Deepfake Detection, Attribution, and Authentication: Insights from the FF4ALL Project

Irene Amerini[1], Mauro Barni[2], Sebastiano Battiato[3], Paolo Bestagini[4], Giulia Boato[5,6], Pietro Bongini[2], Vittoria Bruni[1], Roberto Casula[7], Lorenzo Cirillo[1], Roberto Caldelli[8,9], Giuseppe Daidone[1], Francesco De Natale[5,13], Rocco De Nicola[10], Luca Guarnera[3,*], Simone Maurizio La Cava[7], Sara Mandelli[4], Gian Luca Marcialis[7], Marco Micheletto[7], Andrea Montibeller[5], Viola Negroni[4], Giulia Orrù[7], Pericle Perazzo[11], Giovanni Puglisi[7], Davide Salvi[4], Benedetta Tondi[2], Stefano Tubaro[4], Massimo Villari[12] and Domenico Vitulano[1]

[1]*Sapienza University of Rome, 00185 Roma, Italy*
[2]*University of Siena, 53100 Siena, Italy*
[3]*University of Catania, 95125 Catania, Italy*
[4]*Politecnico di Milano, 20133 Milano, Italy*
[5]*University of Trento, Italy*
[6]*Truebees S.r.l., Italy*
[7]*University of Cagliari, 09123 Cagliari, Italy*
[8]*CNIT, Florence, Italy*
[9]*Universitas Mercatorum, Rome, Italy*
[10]*IMT School for Advanced Studies, 55100 Lucca, Italy*
[11]*University of Pisa, 56122 Pisa, Italy*
[12]*University of Messina, 98166 Messina, Italy*
[13]*CNIT, University of Trento, Italy*

Abstract

The rapid advancement of deep generative models has enabled the large-scale creation of highly realistic deepfakes. While these technologies support innovative applications, they also pose serious threats to trust, security, and digital integrity. As a response, the FF4ALL project investigates deepfake media forensics through a unified framework that integrates source attribution, passive detection, robustness analysis in realistic conditions, and active authentication mechanisms. This paper provides a consolidated overview of the main scientific outcomes achieved within FF4ALL. On the attribution side, novel hierarchical and open-world strategies are presented to identify both the generation technology and the specific model instance responsible for synthetic content. For passive detection, the project advances state-of-the-art methodologies in audio, visual, and multimodal domains, with particular emphasis on generalization to unseen attacks, adversarial robustness, and explainability. Realistic deployment scenarios are addressed through extensive evaluation under social-media compression, continual learning, and out-of-distribution conditions. Beyond passive analysis, FF4ALL develops active authentication solutions, including geometry-aware forensic features, fragile watermarking, cryptographic croppable signatures, and blockchain-based timestamping.

Keywords

Life-Long Media Authentication, Deepfake Attribution, Open-Set Attribution, Generative Model Fingerprinting, Multimodal Deepfake Detection, Active Authentication

1. Introduction

Synthetic media produced by deep neural networks has become pervasive across images, video, audio, and multimodal content [1]. Progress in deep generative modelling, from adversarial architectures to modern diffusion-based systems, has progressively reduced the perceptual gap between generated and natural signals [2, 3]. State-of-the-art (SOTA) models now deliver high-fidelity imagery and photorealistic human faces, as well as natural-sounding synthetic speech and expressive talking avatars [4, 5]. Advances in deepfake generation support creative content production, personalised media, and accessibility, yet also enable large-scale deception, reputational damage, and privacy violations, as documented in political communication and non-consensual synthetic intimate imagery [6, 7]. In parallel, research on deepfake detection and multimedia forensics has expanded at a comparable pace. Initial efforts focused on specific visual artefacts or constrained manipulation scenarios, while more recent work tackles cross-dataset evaluation, multimodal analysis, and increasingly sophisticated generation techniques [8, 9]. Recent systematic reviews highlight both the diversity of detection approaches and the rapidly growing volume of publications in this area [10]. Despite this intense activity, deepfake detection remains characterised by an arms-race dynamic. Improvements in generative modelling tend to erode previously exploitable artefacts, while detectors often rely on training data that only partially cover the variability of real-world conditions. Many methods exhibit limited robustness when confronted with unseen generators, new compression settings or acquisition conditions [11, 12]. Generalisation weaknesses, coupled with heterogeneous protocols and metrics, complicate cross-paper comparison and hinder a precise assessment of how individual systems would behave in realistic operational environments.

Operational scenarios typically require more than a binary authenticity decision. Depending on the application, relevant tasks include attribution of synthetic media to generative technologies or model instances, passive authentication of biometric and multimedia signals, evaluation of robustness under realistic or adversarial channel conditions, and active protection through watermarking or cryptographic signatures. Taxonomies proposed in recent literature group these tasks into complementary functional areas and usually discuss them separately for different modalities and datasets [13]. A comprehensive view that links attribution, passive detection, robustness analysis, and active protection within a single architectural and experimental framework remains relatively uncommon.

The FF4ALL project (*Detection of Deep Fake Media and Life-Long Media Authentication*) was conceived in this context [14]. The project aims to develop theoretical and practical tools to detect fake or counterfeited content, trace media artefacts back to their origin, and support life-long authentication of multimedia items through a combination of passive analysis and active protection mechanisms. The overall research programme is structured into four work packages (WPs): deepfake attribution and recognition (WP1), passive deepfake authentication methods (WP2), deepfake detection in realistic scenarios (WP3), and active authentication (WP4). The present work offers a consolidated view of the main scientific results achieved within FF4ALL, organised along these four research lines, and focuses on problem formulations, methodological choices, datasets, and evaluation protocols.

The remainder of the paper is structured as follows. Section 2 provides a brief overview of deep-

Joint National Conference on Cybersecurity (ITASEC SERICS 2026), February 09-13, 2026, Cagliari, IT
*Corresponding author.
irene.amerini@uniroma1.it (I. Amerini); mauro.barni@unisi.it (M. Barni); sebastiano.battiato@unict.it (S. Battiato); paolo.bestagini@polimi.it (P. Bestagini); giulia.boato@unitn.it (G. Boato); pietro.bongini@unisi.it (P. Bongini); vittoria.bruni@uniroma1.it (V. Bruni); roberto.casula@unica.it (R. Casula); lorenzo.cirillo@uniroma1.it (L. Cirillo); roberto.caldelli@cnit.it (R. Caldelli); giuseppe.daidone@uniroma1.it (G. Daidone); francesco.denatale@unitn.it (F. D. Natale); rocco.denicola@imtlucca.it (R. D. Nicola); luca.guarnera@unict.it (L. Guarnera); simonem.lac@unica.it (S. M. L. Cava); sara.mandelli@polimi.it (S. Mandelli); marcialis@unica.it (G. L. Marcialis); marco.micheletto@unica.it (M. Micheletto); andrea.montibeller@unitn.it (A. Montibeller); viola.negroni@polimi.it (V. Negroni); giulia.orru@unica.it (G. Orrù); pericle.perazzo@unipi.it (P. Perazzo); puglisi@unica.it (G. Puglisi); davide.salvi@polimi.it (D. Salvi); benedetta.tondi@unisi.it (B. Tondi); stefano.tubaro@polimi.it (S. Tubaro); massimo.villari@unime.it (M. Villari); domenico.vitulano@uniroma1.it (D. Vitulano)

fake generation and detection processes, introducing the terminology and task definitions adopted throughout the manuscript. Section 3 summarises the overall research framework, outlining the main objectives and the role of the four work packages. Section 4 discusses deepfake attribution and model fingerprinting techniques, which aim to identify the sources of synthetic media and the generative models responsible for their creation. To provide a broader perspective, Section 5 examines passive deepfake and biometric authentication methods, Section 6 focuses on the practical challenges of detecting manipulated content in realistic and adversarial scenarios, and Section 7 addresses active authentication and infrastructural mechanisms for long-term integrity. A general discussion is provided in Section 8, synthesising insights across these components and highlighting open challenges. Finally, Section 8 summarises the main lessons and outlines directions for future work in deepfake media forensics.

2. Background and Related Work

2.1. Deepfake Generation

Generative models are a class of machine learning algorithms designed to learn the underlying probability distribution of a given dataset and to generate new samples that resemble real data. Unlike discriminative models, which focus on predicting class labels or regression targets, generative models aim to approximate the data distribution p_{data} and synthesize samples that share its statistical and semantic properties. They have become central in computer vision, speech processing and multimedia applications, enabling tasks such as image and video synthesis, text-to-speech, style transfer and data augmentation [15].

Within this broad family, deepfake media are typically produced by deep generative architectures that operate on faces, voices or full-body representations. Early systems were mainly based on autoencoders and their variants, whereas recent approaches increasingly rely on Generative Adversarial Networks (GANs) and Diffusion Models (DMs), which currently dominate high-quality synthesis [16]. GANs implement an adversarial game between a generator and a discriminator [17], while diffusion models learn to reverse a gradual noising process [18]. Both families can be combined with additional modules, such as facial landmark extractors, 3D geometric models or vocoders, to target specific modalities and manipulation types (e.g., face swapping, facial reenactment, lipsync, voice cloning).

Generic pipeline for multimodal deepfake creation (image, video, and audio contents). In general, deepfake creation involves a multi-stage pipeline that maps raw inputs to realistic synthetic media [15]. As illustrated in Figure 2.1, the process typically starts from data collection, where source and target images, videos or audio recordings are gathered. A preprocessing stage follows, including operations such as face detection, landmark localization, alignment, normalization and, for audio, feature extraction (e.g., spectrograms or mel-frequency representations). During the model training phase, architectures such as autoencoders, GANs or diffusion models are optimized to capture identity, expression, pose and other relevant attributes. Once trained, the generation stage performs face swapping, facial reenactment, attribute manipulation or voice conversion, often followed by blending and post-processing to obtain seamless composites. Finally, an evaluation stage assesses the visual and perceptual quality of the output and may optionally refine the model through adversarial training or human-in-the-loop feedback.

2.2. Deepfake Detection

The rapid progress of synthetic media generation has stimulated extensive research on deepfake detection, with a growing body of literature proposing taxonomies, methodologies, and benchmarks for assessing manipulated content [10]. Detection techniques aim to distinguish real from synthetic media by analysing visual, temporal, physiological, or multimodal cues that current generative systems struggle to reproduce consistently. Recent surveys highlight a progressive shift from handcrafted, artifact-centric strategies to data-driven approaches capable of capturing subtle and high-dimensional irregularities.

Despite the diversity of methods, deepfake detectors can broadly be categorized into four principal families [8]: undirected learning-based approaches, artifact-oriented methods, biological signal analysis, and texture or spatio-temporal consistency-based techniques. Each category targets conceptually distinct aspects of manipulated content and exhibits different strengths and limitations, particularly in real-world conditions characterized by compression, occlusion, motion blur, and domain shift [19].

Undirected approaches rely on deep networks trained to discriminate between real and fake content without explicitly modeling specific manipulation artifacts or characteristics. Convolutional neural networks, transformers, and hybrid architectures learn latent representations directly from data, capturing discrepancies that may be imperceptible to humans [20]. These methods historically achieve strong performance on in-distribution datasets; however, their generalization remains limited when confronted with unseen manipulations, new generative models, or platform-specific degradations. Recent work explores domain adaptation strategies, latent space regularization, and data augmentation to mitigate overfitting and improve robustness [21]. Although effective, these approaches often lack interpretability, an important consideration in forensic and legal contexts, and can be sensitive to subtle variations in content distribution [22]. Artifact-based methods exploit structural inconsistencies introduced during generation or blending. Surveys consistently report several recurring artifact categories [15] for both audio and video samples: i) spatial artifacts, such as boundary irregularities, inconsistent shadows, unnatural reflections, or texture discontinuities around manipulated regions [23]; ii) frequency-domain artifacts, which arise from limitations in generative models and can be observed through spectral analysis, Fourier transforms, or wavelet decompositions [24] and iii) compression-induced cues, where lossy encoding both obscures and accentuates patterns exploitable by detectors, iv) specificic noise, such as channel pattern noise [25]. Biological signal-based detectors exploit physiological or behavioral patterns that are difficult to replicate with synthetic generation systems [26], such as micro-fluctuations in skin tone or subtle blood-flow indicator, eye-blink patterns, gaze trajectories, facial muscle dynamics, or prosody–lip synchronization. Texture- and spatio-temporal-based approaches leverage fine-grained spatial patterns and their evolution across video frames [27]. In particular, these detectors analyse intra-face texture coherence, detecting mismatches between different facial regions or between subject and background, temporal stability, since authentic videos and audio exhibit consistent texture and time evolution, motion and semantic patterns, whereas manipulated contents often display flickering, abrupt transitions, or unstable high-frequency details [28].

Beyond these general principles, the literature also distinguishes between audio-only, video-only, and audio–video multimodal detection pipelines, depending on the modality under scrutiny and the nature of the manipulation.

3. FF4All Project Overview and Research Objectives

The challenges outlined in the previous sections motivated a coordinated research effort aimed at addressing deepfake media forensics in a systematic and multi-layered way. The FF4ALL initiative (*Detection of Deep Fake Media and Life-Long Media Authentication*)[1] was launched with this goal. The project pursues theoretical and practical advances for detecting and combating manipulated or fully synthetic media, tracing the origin of such content, and supporting long-term authentication of multimedia items in realistic operating conditions. Passive analysis techniques that operate when content is consumed or redistributed are combined with active protection mechanisms applied at creation time, so that subsequent verification becomes more reliable and more transparent for end users.

FF4ALL focuses on deepfake media in a broad sense, including synthetic or manipulated images, videos, and audio, as well as hybrid and multimodal scenarios. Particular attention is devoted to applications where deepfakes have a direct impact on trust and security, such as biometric verification, identity impersonation, social-network misinformation, and evidential use of digital media. The research

[1]Research project line in the context of Spoke 2 “Misinformation and Fakes”, funded by the Italian National Recovery and Resilience Plan (PNRR), Mission 4 “Education and Research”, and by the European Union under the NextGenerationEU programme.

agenda is guided by three overarching questions:

- How to characterise and detect deepfake content in a way that remains robust under new generators, acquisition conditions, and post-processing pipelines?
- How to attribute synthetic media to underlying technologies, model instances, or training data, and how to link such attribution to operational needs in forensic and security contexts?
- How to support life-long media authentication by combining passive forensic analysis with active protection and suitable computational infrastructures?

More information on the consortium, partners, and public deliverables is available on the project website.[2].

3.1. Work packages and research lines

The project is organised into four Work Packages (WPs), each addressing a specific aspect of deepfake detection and media authentication while remaining tightly connected to the others. The structure reflects the layered view adopted in the remainder of the paper.

WP1 – Deepfake Attribution and Recognition. WP1 investigates how synthetic media can be linked to generative processes. Research activities focus on the definition and extraction of fingerprints of generative models and pipelines, on the attribution of content to model families or specific instances, and on the recognition of technological or training-data characteristics that leave stable traces in images, videos, or audio. Task 1.1 concentrates on deepfake fingerprints, including the study of statistical or learned signatures left by different generators. Task 1.2 targets deepfake attribution in closed-set and open-set conditions, considering both technology-level and instance-level attribution of synthetic content.

WP2 – Passive Deepfake Authentication Methods. WP2 develops passive methods that assess authenticity based solely on observed signals, without relying on embedded marks or external information. Research topics include deepfake detection in biometric scenarios, where synthetic or manipulated data target face or voice recognition systems, as well as more general multimedia-authentication settings. Task 2.1 focuses on the interaction between deepfakes and biometric recognition, with emphasis on face and voice. Task 2.2 addresses audio–video deepfake scenarios and multimodal consistency. Task 2.3 explores advanced detection strategies that aim to improve generalisation and robustness across datasets and manipulation pipelines.

WP3 – Deepfake Detection in Realistic Scenarios. WP3 examines how deepfake detectors and attribution tools behave under realistic operating conditions. Social networks, messaging platforms, and content-delivery infrastructures often apply compression, resizing, filtering, and format conversions that can obscure forensic cues. Task 3.1 considers image and video deepfakes "in the wild", where content originates from heterogeneous sources and undergoes uncontrolled processing. Task 3.2 focuses on social-media environments, including platform-specific recompression and re-encoding workflows. Task 3.3 studies adversarial settings, where manipulations are intentionally crafted to evade detectors or attribution tools, and evaluates the robustness of proposed methods under a variety of attack models.

WP4 – Active Authentication and Infrastructure. WP4 develops active mechanisms and supporting infrastructures that facilitate long-term verification of media authenticity. Active fingerprinting strategies for generative models are designed in Task 4.1, with the goal of embedding or inducing identifiable patterns that support later detection or attribution. Task 4.2 addresses authentication of devices and processing chains used for content acquisition and transformation, including signature schemes and related protocols. Task 4.3 focuses on trusted remote media processing, with particular

[2] https://sites.unica.it/ff4all/

Table 1
Results of the deepfake attribution task in the image and audio domain

	Domain	Datset	Performance
Guarnera et al. [34]	Image	Real images: CelebA, FFHQ, and ImageNet Synthetic Images: 9 GAN (AttGAN, CycleGAN, GDWCT, IMLE, ProGAN, StarGAN, StarGAN-v2, StyleGAN, StyleGAN2) 4 text-to-image diffusion models (DALL-E 2, GLIDE, Latent Diffusion, Stable Diffusion).	Level 1 (Deepfake Detection: Real Vs Deepfake) Prec: 0,94 Rec: 0,99 F1: 0,96 Acc: 0,98 Level 2 (Deepfake Technology Recognition: GAN Vs DM) Prec: 0,98 Rec: 0,99 F1: 0,98 Acc: 0,98 Level 3-DM (Deepfake Attribution) Prec: 0,99 Rec: 0,99 F1: 0,99 Acc: 0,99 Level 3-GAN (Deepfake Attribution) Prec: 0,97 Rec: 0,97 F1: 0,97 Acc: 0,97
Di Pierno et al. [37]	Audio	CodecFake, ASVspoof2021, FakeOrReal (FoR)	CodecFake: Prec: 0.9749, Rec: 0.9568, F1: 0.9658 ASVspoof2021: Prec: 0.9402, Rec: 0.9720, F1: 0.9558 FakeOrReal: Prec: 0.9724, Rec: 0.9576, F1: 0.9649

emphasis on cloud and edge computing systems and on distributed learning frameworks that enable privacy-preserving training and deployment of forensic models.

4. Deepfake Attribution

The rapid spread of deepfake generation tools makes it increasingly urgent to develop advanced techniques to identify the origin of synthetic content and fight its misuse. Deepfake attribution [29] aims to determine which model or generative system was used to produce a given piece of content. This does not only involve detecting the general type of model, such as a GAN or a diffusion model, but also trying to estimate the specific model weights [30], which reveal the exact instance of the used model.

State-of-the-art approaches have shown strong results in recognizing deepfakes produced by GANs [31, 32] and Diffusion Models [33, 34, 35]. These methods are able not only to identify the architecture but also to capture model-specific traces left during the generation process. Early studies have also explored this problem in the audio field, aiming to identify which generator produced synthetic speech signals [36, 37]. In the Deepfake Attribution task, some SOTA methods focus on exploiting the unique traces left by generative models. Ning Yu et al. [38, 39] introduced GAN fingerprints, showing that GANs embed stable and identifiable patterns that enable fine-grained, model-level attribution, even under adversarial perturbations. To handle real-world scenarios with fine-tuned or retrained models, Yang et al. [40] proposed DNA-Det, which captures architecture-level fingerprints, extending applicability to privately trained models. Sun et al. [41] addressed open-world attribution through Contrastive Pseudo Learning, aligning features across both known and unseen forgery types. Finally, Guarnera et al. [29] demonstrated that metric learning combined with a ResNET-18 backbone can discriminate between hundreds of instances of the same architecture (e.g., StyleGAN2-ADA [42]), proving the feasibility of instance-level attribution.

During the FF4LL project, several approaches were proposed. Below are just four of the most recent approaches published for deepfake attribution: the first is based on image domain and the second on audio domain. Guarnera et al. [34] proposed a hierarchical system for detecting and attributing deepfakes able to distinguish real images from AI-generated ones in Level 1, identify the technology used between GAN and diffusion models in Level 2, and finally attribute the specific generative architecture in Level 3. Using a multi-level pipeline based on ResNet-101, it achieves an accuracy of over 97%, surpassing SOTA classifiers and methods, and remains robust under common image manipulations. In the field of audio, Di Pierno et al. [37] proposed LAVA, a framework for audio deepfake attribution that uses a convolutional autoencoder and attention-based classifiers to identify both the generation technology and the specific model underlying the synthetic speech, achieving over 96% accuracy and strong open-set robustness. Table 1 summarizes the obtained results.

The advent of open-world scenarios, where models encounter novel and unseen forgeries, has motivated the development of benchmarks such as OW-DFA++ by Sun et al. [43]. This benchmark combines labeled and unlabeled data to assess attribution methods in diverse and evolving settings, and

their Multi-Perspective Sensory Learning (MPSL) framework leverages multi-scale global–local feature alignment and confidence-adaptive pseudo-labeling to improve attribution performance. A reliable solution for Deepfake model recognition is crucial for intellectual property protection [44], enabling the attribution of synthetic images or videos to their model owner and addressing issues of ownership and accountability. Achieving this level of precision requires new strategies and tailored metrics [45], especially when dealing with models trained with minimally different datasets or hyperparameters. In forensic contexts, Deepfake attribution plays a role analogous to camera source identification, aiming to trace synthetic content back to its specific generative model instance. This parallel highlights the need for advanced techniques capable of ensuring authenticity and accountability in digital media. Key challenges include distinguishing between closely related models, recognizing fine-tuned variants, and maintaining robustness against adversarial attacks designed to conceal model fingerprints [46, 47]. Future research may benefit from integrating self-supervised learning, adversarial training, and ensemble approaches that combine complementary signals for more reliable model attribution.

As part of the FF4ALL project, the WILD collaboration aimed to build a dataset for developing and benchmarking synthetic image source attribution methods that can operate in the wild. The WILD dataset [48] [3] is composed of a grand total of 20,000 images (which increase to 50,000 after post-processing), split in two equal-sized parts: a closed set of 10 text-to-image generators, among which 6 SOTA commercial models; an open set of 10 generators. Each generator was used to produce 1,000 images, half of which are used for training, while the other half is split into test and validation. All the images in this latter half have three post-processed copies, with incremental levels of distortion. Half of the open-set was also post-processed in the same way. Moreover, the closed set image subsets of different generators were all generated using the same 1,000 prompts, to avoid prompt-induced biases. These prompts were created with a dedicated python script, randomizing the image characteristics and minimizing the possible distribution biases. This setup allows to test attribution methods in a variety of real-world scenarios, including closed-set and open-set source attribution, also in presence of post-processing. To validate the benchmarks of [48] and make the project's findings accessible, the collaboration developed the FF4ALL WILD Demonstrator[4] (hosted on Hugging Face Spaces). This web-based tool serves as an interactive interface for the SOTA attribution models trained on the WILD dataset, effectively bridging the gap between the theoretical benchmarks and practical forensic applications. This tool ultimately functions as a proof-of-concept for the FF4ALL project's core mission: establishing a reliable pipeline for Life-Long Media Authentication, where forensic tools can adapt to new generators and withstand the degradations of real-world internet usage.

Moreover, the FF4ALL project contributed another dataset for synthetic image source attribution. This dataset [49] is designed to develop and benchmark resynthesis-based attribution methods. The dataset is composed of 11,000 images from 10 text-to-image generators, among which 7 cutting-edge commercial generators, including Leonardo AI and Midjourney. Along with the dataset, a training-free resynthesis method [49] based on CLIP feature extraction and distance calculation was released. The methodology is computationally equivalent to a one-shot method. This method allows to beat all the SOTA baselines for few-shot source attribution when less than 10 shots are available to train the baselines.

5. Passive Deepfake Authentication Methods

Passive deepfake detection encompasses all techniques that assess the authenticity of multimedia content without relying on embedded auxiliary information such as watermarks, cryptographic signatures, or provenance metadata. Instead, these methods operate exclusively on the intrinsic properties of the signal, its spatial, temporal, spectral, physiological, or semantic characteristics, to determine whether it has been manipulated. As deepfake generation technologies continue to advance and synthetic media increasingly circulate through uncontrolled real-world environments, passive approaches remain

[3]The WILD dataset is available on Kaggle: https://www.kaggle.com/datasets/pietrob92/wild-in-the-wild-image-linkage-dataset
[4]The WILD Demonstrator is available at https://huggingface.co/spaces/AMontiB/Dimostratore_FF4ALL

essential for retrospective analysis, large-scale monitoring, and forensic investigations, where trusted metadata is typically unavailable.

Within the FF4ALL project, passive detection represents a foundational component of the broader goal of safeguarding the integrity of digital media. The consortium has investigated complementary strategies spanning the audio, visual, and multimodal domains, with an emphasis on generalization, robustness to post-processing, and explainability, three aspects repeatedly identified in the literature as critical yet unsolved challenges. These efforts include the study of latent-space and representation-learning methods, artifact-centric analyses, biological and behavioral cue modeling, and cross-modal consistency frameworks.

5.1. Audio-only Deepfake detection

Over the past few years, the rapid evolution of generative models has positioned audio-only deepfake detection as a critical area within multimedia forensics. As synthetic speech becomes increasingly realistic and more accessible to malicious actors, traditional detection methods face growing limitations, particularly in generalization, interpretability, and robustness. These challenges motivate a diverse set of research directions aimed at strengthening our ability to identify, analyze, and attribute manipulated audio signals.

Within the FF4ALL project, substantial progress has been made in audio-only deepfake detection, with research spanning multiple directions, from generalizable and robust binary classification methods to one-class detection approaches.

Notably, efforts were devoted to developing a Mixture of Experts system that combines multiple state-of-the-art detectors through an attention-based gating network [50, 51]. This system achieved top performance in the SAFE challenge by leveraging the complementary strengths of each expert. Complementing this, a one-class anomaly detection framework reframes deepfake detection as an outlier problem, training solely on real speech and producing interpretable anomaly maps, which improves generalization to unseen synthetic methods [52].

In the area of source tracing, the project introduced a source verification task inspired by speaker verification, enabling attribution of synthetic speech to its originating generator and addressing open-set challenges [53]. Fine-grained binary detection was also explored through phoneme-level Person-of-Interest methods, allowing detailed and interpretable analysis of impersonations [54]. Additionally, adversarial attacks against deepfake detectors were systematically investigated, assessing perturbations in both time and frequency domains and proposing ensemble-based strategies to evaluate model vulnerabilities [55]. Finally, the project extended its scope to the emerging field of singing voice deepfake detection, analyzing audio representations and feature sets to improve detection performance and understand the differences from standard speech detection [56]. Taken together, these works reflect a broad and coordinated effort to advance the state of audio forensic technology. The FF4ALL project contributes not only improved detection accuracy but also stronger interpretability, enhanced robustness, and new methodologies for source attribution. By tackling fundamental challenges ranging from unseen-generator generalization to adversarial resilience and domain-specific detection, the project helps pave the way for more trustworthy and transparent audio deepfake analysis, addressing both current threats and the evolving landscape of synthetic media.

5.2. Video-only and Image-only Deepfake detection

The visual branch of WP2 follows three complementary directions. A first line of work focuses on representation learning for spatial artefacts and texture inconsistencies in facial regions, with an explicit emphasis on generalisation across datasets and manipulation techniques. The second line targets robustness to heavy compression, where conventional detectors suffer a marked performance drop due to the loss of high-frequency cues. The third line examines how seemingly benign beautification filters affect both deepfake and morphing attack detection, highlighting additional vulnerabilities that arise when aesthetic post-processing is applied to synthetic content or bona fide images.

From the representation-learning perspective, the FF4ALL contributions explored different ways of combining artefact-oriented cues with deep-learning features for frame-based video deepfake detection under scale and compression changes. One approach relies on quality-based artefact modelling, where a set of no-reference and full-reference image-quality measures is computed over multiple facial patches in each frame, both on the original and on a high-pass filtered version, and organised into compact quality–time matrices that are processed by two shallow Convolutional Neural Networks (CNNs) branches whose scores are fused at video level [57]. Experimental results on FaceForensics++ [58] show that this quality-driven representation improves cross-manipulation robustness with respect to plain ImageNet-pretrained backbones, while keeping the architecture relatively simple and interpretable. Furthermore, the proposed approach provided performance that is comparable to more complex state-of-the-art deepfake detection systems often employed as benchmarks, such as RealForensics [59] and LipForensics [60]. A complementary direction is represented by the Texture and Artifact Detector (TAD) framework, which models explicitly the roles of textures and artefacts in a unified architecture [61]. The method represents each face image as the combination of a texture component and an artefact component, and learns two dedicated subnetworks: a texture branch that uses deformable convolutions and a self-supervised reconstruction loss to separate foreground and background facial regions, and an artefact branch that focuses on residual traces introduced by the manipulation or by the acquisition and compression pipeline. Score-level fusion of the two classifiers, combined with ensemble strategies over different manipulation groups, yields improved generalisation on cross-dataset evaluations involving FaceForensics++, CelebDF, DFDC and WildDeepfake, while maintaining competitive intra-dataset performance.

Robustness to real-world compression is addressed in a complementary way by the High-Frequency Enhancement (HiFE) network [62]. Quantitative and qualitative analyses on the raw, c23 and c40 versions of FaceForensics++ [58] show that standard deepfake detectors experience a sharp degradation when moving from high-quality to highly compressed content, mainly due to the loss of discriminative high-frequency details and the presence of compression artefacts. HiFE tackles this problem through an unsupervised enhancement module that can be plugged into existing backbones. The module combines a local branch based on block-wise Discrete Cosine Transform and a global branch based on Discrete Wavelet Transform, together with a two-stage cross-fusion mechanism that reinforces residual high-frequency information in compressed inputs. Experiments on FaceForensics++ at different compression levels, CelebDF V2 and OpenForensics indicate that HiFE consistently narrows the performance gap between uncompressed and highly compressed data, and improves detection accuracy in low-bitrate scenarios without requiring supervision from high-quality reference material.

In [63], Battocchio et al. address the limitations of traditional CNNs in detecting AI-generated videos, particularly their inability to generalize to new, unseen generative models. The authors propose a novel video transformer detector that leverages a 3D Vision Transformer (ViT) backbone to analyze video content not just spatially (frame-by-frame), but temporally, effectively capturing the inconsistent motion artifacts that are often the ”tell” of synthetic media. Testing their method on a newly constructed, diverse dataset comprising videos from five state-of-the-art open-source and proprietary generators, the researchers found that their ViT-based approach significantly outperformed existing baselines. Key results include a True Positive Rate of 95% and a True Negative Rate of 93%, demonstrating superior accuracy and generalization. Furthermore, the study highlights the model’s ”few-shot” learning capabilities, proving it can adapt to detect entirely new types of deepfakes with very little training data.

The interplay between visual manipulation and cosmetic post-processing is investigated in [64], which analyses the impact of beautification filters on both deepfake and morphing attack detection. The work considers a smoothing-based beauty filter with increasing application radius and evaluates two widely used CNN-based detectors, AlexNet [65] and VGG19 [66], on CelebDF [67] for deepfakes and on the AMSL dataset [68] for morphing attacks. Results show a systematic degradation of deepfake detection performance as the smoothing radius increases, with shifts in the score distributions that reduce the separation between real and fake samples. In the morphing scenario, the effect is even more pronounced: error rates for morphing attack detection grow substantially, especially when beautification is applied to genuine images, while filters applied only to morphed faces may even improve separability

Table 2
FF4ALL visual passive detection studies in WP2. Metrics are reported as in the original publications (Acc, AUC, EER, all in %).

Method / study	Protocol	Dataset(s)	Key quantitative results
Quality-based artefact modelling [57]	Intra	FF++ (video, all manip.)	Acc = 99.49, AUC = 99.97
	Cross	FF++ cross-manip. (avg)	Acc = 98.91, AUC = 99.94
Texture and Artifact Detector (E-TAD) [61]	Intra	FF++ (avg over DF, F2F, FS, FSh, NT)	Acc = 95.40
	Cross	WildDF, CelebDF, DFDC (avg)	Acc = 64.18;
HiFE high-frequency enhancement [62]	Intra	FF++ (raw, c23, c40)	AUC = 99.36 (raw), 92.83 (c23), 71.84 (c40).
		CelebDF V2	AUC = 96.64
		OpenForensics	AUC = 99.03
ViT-based synthetic video detector [63]	Intra	VideoDiffusion (raw, c23, c30, c50)	AUC = 88.00 (raw), 79.00 (c23), 70.00 (c30), 67.00 (c50)
	Cross	SORA, LUMA AI, Hunyuan, CogVideo, RunwayML (avg)	AUC = 94.20
"Deceptive Beauty" analysis [64]	Baseline	CelebDF (deepfakes), AMSL (morphs)	CelebDF EER = 22.3 / 30.2 (AlexNet / VGG19); AMSL EER = 27.6 / 19.0.
	Beautified	CelebDF (deepfakes), AMSL (morphs)	CelebDF EER = 28.1 / 35.2; AMSL EER = 41.2 / 37.3 (AlexNet / VGG19).

by emphasising differences with respect to the training distribution.

Table 3 summarises the main characteristics of these works, including their target scenarios, datasets and qualitative performance trends.

5.3. Audio-video Deepfake detection

In recent years, researchers have increasingly turned their attention to *multimodal* deepfake detection, aiming to analyze several types of signals at the same time for stronger and more reliable results. This shift is driven by the shortcomings of conventional methods, which typically target only audio or only video. Since modern deepfakes often manipulate both streams in intricate ways, relying on a single modality can leave important clues unnoticed. Multimodal approaches address this issue by jointly examining audio and visual cues to uncover cross-modal inconsistencies or hidden artifacts. For example, a fabricated video might display highly realistic facial movements while revealing subtle irregularities in the accompanying audio, such as poorly synced speech or background noises that feel out of place.

Within the FF4ALL project, and based on prior findings [69, 27], current research has evolved in two main directions. The first examines emotional consistency as a potential indicator of fake content. While synthetic media can generate highly realistic faces and voices, reproducing natural emotional dynamics remains a challenge. This line of investigation uses LSTM-based emotion prediction from low-level audio and video descriptors, using the temporal progression of emotional cues as a signal to differentiate authentic videos from manipulated ones. The second direction focuses on cross-modal coherence. Deepfake generation methods may convincingly forge audio or visual streams in isolation, yet ensuring semantic and temporal consistency between them is notably more difficult. To explore this aspect, a multimodal detection strategy is being developed that analyzes audio–visual features over time using time-aware neural models. A key characteristic of this approach is its reliance on separate monomodal datasets for training, reducing dependence on scarce multimodal deepfake resources while still aiming for strong generalization to unseen multimodal forgeries. Early investigations into different fusion strategies suggest that combining complementary modalities can yield more robust detection than treating audio and video independently.

Although the multimedia forensics field has made notable strides, several key challenges continue to stand in the way of progress. One pressing issue for multimodal research is the scarcity of high-quality audio–video Deepfake datasets. Most existing resources still emphasize a single modality, leaving a significant gap for systems that need coordinated audio–visual training data. Explainability also remains a major hurdle. Many current detection models provide little insight into how their decisions are formed, which poses serious problems for forensic and legal settings where transparency and justification are mandatory. Strengthening the interpretability of these systems is therefore essential to establish trust and ensure they can be deployed in critical real-world scenarios.

6. Deepfakes Detection Method on Realistic Scenarios

The capabilities of generative models to produce high-quality fake content pose several challenges for deepfake detectors acting in real-world scenarios [70]. Although deep learning models have shown effectiveness in closed settings, their performance often deteriorates when applied in open environments, where data distributions shift and novel manipulation techniques emerge [71, 72, 73]. In addition, understanding and explaining the decision-making processes of deep learning models becomes increasingly important in such unconstrained environments, both for ensuring trustworthiness and for diagnosing failure cases [74]. Therefore, there is a growing need for detection methods that generalize appropriately across real-world distributions, offering interpretable explanations of their predictions. Building on the mentioned issues, a significant amount of recent research tries to identify the root causes of poor generalization in deepfake detectors when evaluated on out-of-distribution (OOD) samples. Recent studies agree that detectors tend to overfit to low-level acquisition or generation artifacts present in their training sets, limiting their generalizability to unseen forgeries at test time [75, 76, 77]. As a consequence, several studies propose to use semantic features to encourage models to focus on forgery-invariant features rather than dataset-specific artifacts [78, 79].

Within the FF4ALL project, multiple efforts have been conducted to address OOD generalization and the broader challenge of deploying deepfake detectors in realistic scenarios. In particular, Maiano et al. [80] propose to detect OOD samples by leveraging the contextual properties of the attention mechanism, while Wani et al. [81] decompose spectrograms into frequency bands in order to increase robustness of deep models to unseen synthesis methods. Additionally, different approaches have been tested on deepfake detection, as Leporoni et al. [82], who show that generating fake content introduces possible inconsistencies in the depth of the generated images, and Mongelli et al. [83] exploit two CNNs to extract and process spatial and temporal features concurrently. In a more realistic context, Cirillo et al. [84] propose a framework that leverages explainability to assess the adversarial robustness of deepfake detectors with generalization capabilities, evaluating models under both in-distribution (ID) and OOD conditions and even in the presence of adversarial perturbations across the WILD dataset [48]. Quantitative results are reported in Table 3.

Additionally, in [85], significant effort has been devoted to the development of a novel real-world dataset, TrueFake. This dataset identifies a critical vulnerability in current synthetic image detection systems: their inability to generalize to “in-the-wild” scenarios where images are shared via social media. By constructing a massive dataset of 600,000 images, real and synthetic from GANs and Diffusion Models, and passing them through the actual compression pipelines of social networks, the authors demonstrated that social media processing acts as a ”laundering” mechanism for forensic traces. The study found that while the visual quality of the images remained largely intact, the subtle high-frequency artifacts that detectors rely on were aggressively smoothed out by platform-specific compression, causing the performance of state-of-the-art detectors to drop precipitously. However, the paper also provides a solution: the authors discovered that by fine-tuning these detection models on the TrueFake dataset (specifically the subset of images that had undergone social media processing), they could effectively recover detection accuracy. This underscores the necessity of moving away from laboratory-clean benchmarks and instead training forensic models on data that mirrors the destructive transformations of the real digital ecosystem. Similarly, in [86] Montibeller et al. tackle a significant bottleneck in

Table 3
FF4ALL deepfake detection methods on realistic scenarios in WP3. Accuracy (Acc), precision (Pre), recall (Rec), and attack success rate (ASR) are averaged and reported in % as in the original publications.

Method / study	Protocol	Dataset(s)	Key quantitative results
Explainability-driven adversarial assessment [84]	ID	ProGAN images	Acc = 99.70, ASR = 50.05
	OOD	WILD (DALL-E 3)	Acc = 51.20, ASR = 73.70
		WILD (Midjourney)	Acc = 50.00, ASR = 75.61
Continuous fake media detection [88]	ID	IMLE	Acc = 97.30, Pre = 99.42, Rec = 93.66
		FaceForensics++	Acc = 64.65, Pre = 70.14, Rec = 48.03
		CRN	Acc = 94.36, Pre = 99.65, Rec = 89.51
		WildDeepfake	Acc = 51.22, Pre = 52.29, Rec = 21.32
	OOD	IMLE	Acc = 94.74
		FaceForensics++	Acc = 94.12
		CRN	Acc = 95.98
		WildDeepfake	Acc = 53.87

video forensics: the ”domain gap” where video deepfake detectors performing perfectly in the lab fail catastrophically when applied to videos shared on platforms like YouTube or Facebook. The authors identify that proprietary social network algorithms apply aggressive compression (e.g., specific bitrate control and resizing) that obliterates the fine-grained forensic traces used by standard detectors. To solve this without the prohibitive cost and API limitations of uploading millions of training videos to social media, the paper introduces the Social Network Video Sharing Emulator. The key finding is that by probing a social network with a small set of ”probe” videos (fewer than 50), one can reverse-engineer the platform’s specific encoding parameters (such as the constant rate factor and resizing logic). Using these estimated parameters, the authors built a local emulator that mimics the degradation of major platforms. The results demonstrate that detectors fine-tuned on locally emulated data achieve detection accuracy comparable to those trained on actual social media uploads. This demonstrates that researchers can robustly train ”social-media-proof” video detectors offline, effectively bypassing the need for direct, large-scale interaction with social media APIs.

In parallel, continual learning frameworks are increasingly adopted to ensure that detection models retain prior knowledge while adapting to novel manipulation. Wani et al. [87] address the critical challenges of catastrophic forgetting and incremental learning by ensuring discriminative feature extraction and computational efficiency. Tassone et al. [88] propose an analysis of two continuous learning techniques on a short and a long sequence of fake media. The corresponding average results are reported in Table 3. Overall, our experiments using explainability-based techniques and continual learning frameworks demonstrate a consistent deterioration of deepfake detector performance under OOD conditions, reflected in reduced accuracy and higher attack success rates. At the same time, these approaches offer valuable insights, revealing the spurious or non-transferable cues that detectors over-rely on, and the benefits of incremental adaptation to evolving manipulation techniques. Together, they outline concrete pathways for improving deepfake detectors’ robustness, generalizability, and resilience in real-world deployment. Moreover, Figure 1 shows a continual learning pipeline tailored for deepfake detection under realistic, incremental update scenarios, as explained in [88]. Different sources, such as generative models and social media, provide continuous data that are analyzed by forensic experts. This ensures the system’s continual retraining. After that, these data are used for continual learning and monitoring. The data drift distribution module raises an alert whenever OOD data distributions are detected.

7. Active Authentication

Most deepfake detection approaches proposed in the literature are based on passive forensic analysis, where authenticity is assessed after content creation and dissemination by inspecting intrinsic signal artifacts [89, 90, 91]. In contrast, *active* deepfake detection relies on the proactive embedding of

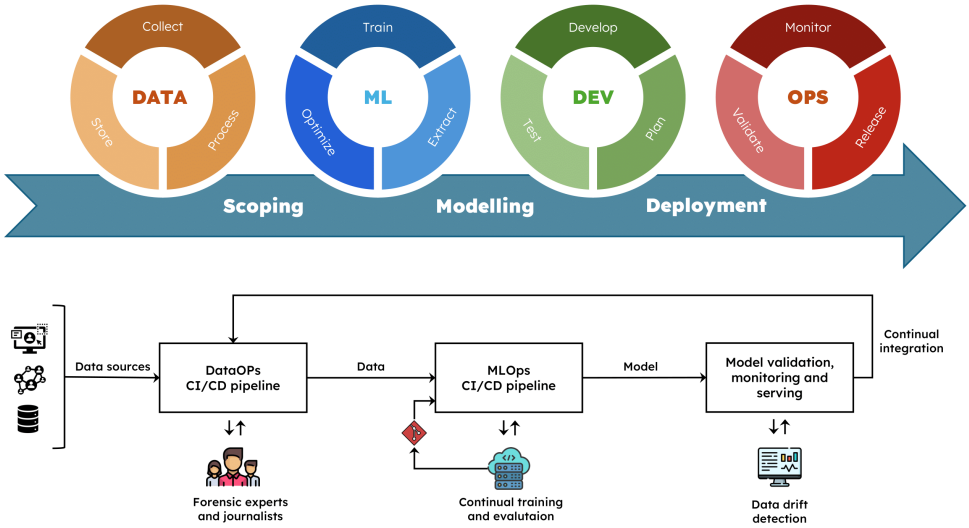

Figure 1: Proposed CI/CD pipeline for deepfake detection reported in [88]. Different sources, such as generative models and social media, provide continuous data that are analyzed by forensic experts. This ensures the system's continual retraining. After that, these data are used for continual learning and monitoring. The data drift distribution module raises an alert whenever OOD data distributions are detected.

fingerprints or watermarks during the content generation process, enabling direct verification of provenance and authenticity at test time. This paradigm shift moves part of the forensic responsibility upstream in the media production pipeline and requires the cooperation of the entity that trains and deploys the generative model. FF4ALL has explicitly investigated *active* and *trusted* strategies that shift part of the verification process upstream to the content generation, acquisition, and distributed processing stages. This integrated perspective addresses both the intrinsic fragility of passive detectors under adversarial conditions and the need for deterministic guarantees on content origin and integrity. The limitations of purely passive detection have been clearly demonstrated through the analysis of adversarial post-processing attacks. In particular, Coccomini et al. [92] showed that super-resolution can be effectively exploited as a black-box attack to suppress forensic artifacts in face-swapped images, leading to severe performance degradation of state-of-the-art deepfake detectors while preserving high visual quality. These results highlight that detection systems based solely on visual artifacts are inherently vulnerable to content enhancement operations that are increasingly accessible to end users. To improve robustness and generalization, FF4ALL research also investigated detection features rooted in physical image formation rather than appearance alone. Ciamarra et al. [93] proposed local surface frames as geometric descriptors capable of revealing inconsistencies introduced by generative models, particularly diffusion-based generators that lack a true optical acquisition process. This approach was further extended by Affatato et al. [94] through the integration of surface frames with texture information extracted from 3D Morphable Models, achieving significantly improved cross-generator generalization and robustness to post-processing operations in cross-dataset evaluations. These results indicate that geometry-aware forensic features represent a reliable direction to mitigate the rapid evolution of generative models. Beyond passive analysis, FF4ALL has explored active protection mechanisms based on data hiding and cryptographic authentication. Ghiani et al. [95] introduced a fragile watermarking framework based on deep steganography for biometric and document images. The watermark is embedded at acquisition time and enables both integrity verification and manipulation classification, including morphing and replacement attacks, while preserving compliance with ICAO imaging standards. This approach provides a practical realization of active deepfake detection, where content authenticity can be verified deterministically rather than inferred statistically. Complementary to signal-level watermarking, cryptographic authentication of visual content has been addressed through the introduction of croppable signatures for JPEG images. Perazzo et al. [96] proposed a block-wise

signing scheme based on aggregatable BLS signatures, enabling robust origin authentication even after legitimate spatial editing operations such as cropping. Any content-altering manipulation, including deepfake generation, invalidates the signature, thus providing cryptographic repudiation of forged media. Blockchain-based timestamping was further employed to extend the long-term validity of such signatures beyond certificate expiration, ensuring non-repudiation and temporal authenticity over extended time horizons. Finally, the problem of trust has also been addressed at the level of distributed learning infrastructures for media verification. Garofalo et al. [97] introduced Federated Objective (FedObj), a truth-aware aggregation strategy for federated learning that weights client contributions based on their generalization performance on a trusted reference dataset. This mechanism significantly improves robustness against poisoned or deceptive participants. In parallel, browser-based federated learning architectures integrated within a cloud–edge continuum [98] have enabled large-scale, privacy-preserving participation in collaborative training for multimedia forensics.

8. Conclusions

This work summarizes the main achievements of the FF4ALL project in deepfake detection, attribution, and media authentication, highlighting a multilayer approach that integrates passive forensics, source attribution, robustness evaluation, and active authentication. By addressing complementary aspects of the deepfake life cycle, FF4ALL moves beyond isolated detection strategies and provides a unified view of how reliability, provenance, and trust can be jointly enforced in modern multimedia ecosystems.

The project introduces realistic benchmarks and open-world evaluation settings for source attribution, enabling the systematic analysis of synthetic content generated by both known and previously unseen models. On the detection side, FF4ALL advances more generalizable techniques that are resilient to diffusion-based generators, social-media compression, and adversarial manipulations, while also promoting explainability-aware analysis for forensic interpretation. In parallel, active approaches based on fragile watermarking, cryptographic croppable signatures, and blockchain-based timestamping enable proactive media authentication and long-term provenance verification, complementing passive forensic methods. Finally, trustworthy and privacy-preserving federated learning strategies demonstrate how large-scale collaborative training can be made robust against malicious participants while safeguarding sensitive data.

Despite these advances, significant challenges remain. The rapid evolution of generative models continues to stress the generalization capabilities of forensic tools, standardization of datasets and protocols is still limited across tasks, and explainability and legal admissibility remain open issues for operational deployment. Overall, FF4ALL lays a solid foundation for future research on trustworthy deepfake analysis and represents a step toward bridging the gap between academic forensics and real-world media authentication.

Acknowledgments

This work was supported by Project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] S. Yazdani, A. Singh, N. Saxena, Z. Wang, A. Palikhe, D. Pan, U. Pal, J. Yang, W. Zhang, Generative ai in depth: A survey of recent advances, model variants, and real-world applications, Journal of Big Data 12 (2025) 1–43. doi:10.1186/s40537-025-01247-x.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Nets, Advances in Neural Information Processing Systems 27 (2014).

[3] J. Ho, A. Jain, P. Abbeel, Denoising Diffusion Probabilistic Models, Advances in Neural Information Processing Systems 33 (2020) 6840–6851.

[4] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

[5] T. M. Wani, S. A. A. Qadri, F. A. Wani, I. Amerini, Navigating the soundscape of deception: A comprehensive survey on audio deepfake generation, detection, and future horizons, Foundations and Trends in Privacy and Security 6 (2024) 153–345. doi:10.1561/3300000048.

[6] M. Pawelec, Deepfakes and democracy (theory): How synthetic audio-visual media for disinformation and hate speech threaten core democratic functions, Digital Society 1 (2022) 1–27. doi:10.1007/s44206-022-00010-6.

[7] R. Umbach, N. Henry, G. F. Beard, C. M. Berryessa, Non-consensual synthetic intimate imagery: Prevalence, attitudes, and knowledge in 10 countries, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, ACM, 2024, pp. 779:1–779:20. doi:10.1145/3613904.3642382.

[8] R. Tolosana, L. Verdoliva, J. Fierrez, A. Morales, J. Ortega-García, Deepfakes and beyond: A survey of face manipulation and fake detection, Information Fusion 64 (2020) 131–148. doi:10.1016/j.inffus.2020.06.014.

[9] L. Verdoliva, Media forensics and deepfakes: An overview, IEEE Journal of Selected Topics in Signal Processing 14 (2020) 910–932. doi:10.1109/JSTSP.2020.3002101.

[10] V. K. Sharma, R. Garg, Q. Caudron, A systematic literature review on deepfake detection techniques, Multimedia Tools and Applications 84 (2025) 22187–22229. doi:10.1007/s11042-024-19906-1.

[11] A. Heidari, N. Jafari Navimipour, H. Dag, M. K. Ünal, Deepfake detection using deep learning methods: A systematic and comprehensive review, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 14 (2024) e1520. doi:10.1002/widm.1520.

[12] T. Wang, X. Liao, K.-P. Chow, X. Lin, Y. Wang, Deepfake detection: A comprehensive survey from the reliability perspective, ACM Computing Surveys 57 (2025) 58:1–58:35. doi:10.1145/3699710.

[13] G. Pei, J. Zhang, M. Hu, Z. Zhang, C. Wang, Y. Wu, G. Zhai, J. Yang, C. Shen, D. Tao, Deepfake generation and detection: A benchmark and survey, arXiv preprint arXiv:2403.17881 (2024).

[14] I. Amerini, M. Barni, S. Battiato, P. Bestagini, G. Boato, V. Bruni, R. Caldelli, F. De Natale, R. De Nicola, L. Guarnera, et al., Deepfake media forensics: Status and future challenges, Journal of Imaging 11 (2025) 73.

[15] Y. Patel, S. Tanwar, R. Gupta, P. Bhattacharya, I. E. Davidson, R. Nyameko, S. Aluvala, V. Vimal, Deepfake generation and detection: Case study and challenges, IEEE Access 11 (2023) 143296–143323.

[16] A. AV, S. Das, A. Das, et al., Latent flow diffusion for deepfake video generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3781–3790.

[17] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A. A. Bharath, Generative adversarial networks: An overview, IEEE signal processing magazine 35 (2018) 53–65.

[18] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, Deep Unsupervised Learning Using Nonequilibrium Thermodynamics, in: International Conference on Machine Learning, PMLR, 2015, pp. 2256–2265.

[19] L. Cuccovillo, C. Papastergiopoulos, A. Vafeiadis, A. Yaroshchuk, P. Aichroth, K. Votis, D. Tzovaras, Open Challenges in Synthetic Speech Detection, in: IEEE International Workshop on Information Forensics and Security (WIFS), 2022.

[20] V. L. Thing, Deepfake detection with deep learning: Convolutional neural networks versus transformers, in: 2023 IEEE International Conference on Cyber Security and Resilience (CSR), IEEE, 2023, pp. 246–253.
[21] Z. Yan, Y. Luo, S. Lyu, Q. Liu, B. Wu, Transcending forgery specificity with latent space augmentation for generalizable deepfake detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 8984–8994.
[22] R. Lanzino, F. Fontana, A. Diko, M. R. Marini, L. Cinque, Faster than lies: Real-time deepfake detection using binary neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3771–3780.
[23] L. Chai, D. Bau, S.-N. Lim, P. Isola, What makes fake images detectable? understanding properties that generalize, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI, Springer, 2020, pp. 103–120.
[24] R. Durall, M. Keuper, J. Keuper, Watch your up-convolution: Cnn-based generative deep neural networks are failing to reproduce spectral distributions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7890–7899.
[25] Z. Wang, G. Wei, Q. He, Channel pattern noise-based playback attack detection algorithm for speaker recognition, in: IEEE International Conference on Machine Learning and Cybernetics (ICMLC), 2011.
[26] U. A. Ciftci, I. Demir, L. Yin, Fakecatcher: Detection of synthetic portrait videos using biological signals, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2020).
[27] B. Hosler, D. Salvi, A. Murray, F. Antonacci, P. Bestagini, S. Tubaro, M. C. Stamm, Do deepfakes feel emotions? a semantic approach to detecting deepfakes via emotional inconsistencies, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
[28] M. Lomnitz, Z. Hampel-Arias, V. Sandesara, S. Hu, Multimodal approach for deepfake detection, in: IEEE Applied Imagery Pattern Recognition Workshop (AIPR), 2020.
[29] L. Guarnera, O. Giudice, M. Nießner, S. Battiato, On the exploitation of deepfake model recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop, 2022, pp. 61–70.
[30] V. Asnani, X. Yin, T. Hassner, X. Liu, Reverse engineering of generative models: Inferring model hyperparameters from generated images, IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).
[31] L. Guarnera, O. Giudice, S. Battiato, Fighting deepfake by exposing the convolutional traces on images, IEEE Access 8 (2020) 165085–165098.
[32] O. Giudice, L. Guarnera, S. Battiato, Fighting deepfakes by detecting gan dct anomalies, Journal of Imaging 7 (2021) 128.
[33] L. Guarnera, O. Giudice, S. Battiato, Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models, in: Intelligent Systems Conference, Springer, 2024, pp. 615–625.
[34] L. Guarnera, O. Giudice, S. Battiato, Mastering deepfake detection: A cutting-edge approach to distinguish gan and diffusion-model images, ACM Transactions on Multimedia Computing, Communications and Applications (2024).
[35] O. Pontorno, L. Guarnera, S. Battiato, Deepfeaturex net: Deep features extractors based network for discriminating synthetic from real images, in: International Conference on Pattern Recognition, Springer, 2025, pp. 177–193.
[36] D. Salvi, P. Bestagini, S. Tubaro, Exploring the synthetic speech attribution problem through data-driven detectors, in: IEEE International Workshop on Information Forensics and Security (WIFS), 2022.
[37] A. Di Pierno, L. Guarnera, D. Allegra, S. Battiato, Towards reliable audio deepfake attribution and model recognition: A multi-level autoencoder-based framework, in: Proceedings of the 1st on Deepfake Forensics Workshop: Detection, Attribution, Recognition, and Adversarial Challenges in the Era of AI-Generated Media, 2025, pp. 101–109.
[38] N. Yu, L. S. Davis, M. Fritz, Attributing fake images to gans: Learning and analyzing gan finger-

prints, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7556–7566.
[39] N. Yu, V. Skripniuk, S. Abdelnabi, M. Fritz, Artificial fingerprinting for generative models: Rooting deepfake attribution in training data, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 14448–14457.
[40] T. Yang, Z. Huang, J. Cao, L. Li, X. Li, Deepfake network architecture attribution, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 4662–4670.
[41] Z. Sun, S. Chen, T. Yao, B. Yin, R. Yi, S. Ding, L. Ma, Contrastive pseudo learning for open-world deepfake attribution, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 20882–20892.
[42] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, T. Aila, Training generative adversarial networks with limited data, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 12104–12114.
[43] Z. Sun, S. Chen, T. Yao, R. Yi, S. Ding, L. Ma, Rethinking open-world deepfake attribution with multi-perspective sensory learning, International Journal of Computer Vision (2024) 1–24.
[44] R. Leotta, O. Giudice, L. Guarnera, S. Battiato, Not with my name! inferring artists' names of input strings employed by diffusion models, in: International Conference on Image Analysis and Processing, Springer, 2023, pp. 364–375.
[45] Z. Huang, B. Li, Y. Cai, R. Wang, S. Guo, L. Fang, J. Chen, L. Wang, What can discriminator do? towards box-free ownership verification of generative adversarial networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 5009–5019.
[46] L. Guarnera, F. Guarnera, A. Ortis, S. Battiato, G. Puglisi, Evasion Attack on Deepfake Detection via DCT Trace Manipulation, in: International Conference on Pattern Recognition, Springer, 2024.
[47] F. Guarnera, L. Guarnera, A. Ortis, S. Battiato, G. Puglisi, A novel adversarial gray-box attack on dct-based face deepfake detectors, IEEE Access (2025).
[48] P. Bongini, S. Mandelli, A. Montibeller, M. Casu, O. Pontorno, C. V. Ragaglia, L. Zanchetta, M. Aquilina, T. M. Wani, L. Guarnera, B. Tondi, G. Boato, P. Bestagini, I. Amerini, F. De Natale, S. Battiato, M. Barni, Wild: a new in-the-wild image linkage dataset for synthetic image attribution, in: 2025 International Joint Conference on Neural Networks (IJCNN), 2025, pp. 1–8. doi:10.1109/IJCNN64981.2025.11227289.
[49] P. Bongini, V. Molinari, A. Costanzo, B. Tondi, M. Barni, Training-free source attribution of ai-generated images via resynthesis, in: 2025 IEEE International Workshop on Information Forensics and Security (WIFS), 2025.
[50] V. Negroni, D. Salvi, A. I. Mezza, P. Bestagini, S. Tubaro, Leveraging Mixture of Experts for Improved Speech Deepfake Detection, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025.
[51] V. Negroni, D. Salvi, A. I. Mezza, P. Bestagini, S. Tubaro, Attention-based mixture of experts for robust speech deepfake detection, in: IEEE WIFS, 2025.
[52] E. Coletta, D. Salvi, V. Negroni, D. U. Leonzio, P. Bestagini, Anomaly detection and localization for speech deepfakes via feature pyramid matching, in: Eurasip EUSIPCO, 2025.
[53] V. Negroni, D. Salvi, P. Bestagini, S. Tubaro, Source verification for speech deepfakes, in: INTERSPEECH, 2025.
[54] D. Salvi, V. Negroni, S. Mandelli, P. Bestagini, S. Tubaro, Phoneme-level analysis for person-of-interest speech deepfake detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 1586–1595.
[55] W. E. Wang, D. Salvi, V. Negroni, D. U. Leonzio, P. Bestagini, S. Tubaro, Bim-based adversarial attacks against speech deepfake detectors, Electronics 14 (2025) 2967.
[56] M. Gohari, D. Salvi, P. Bestagini, N. Adami, Audio features investigation for singing voice deepfake detection, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025.
[57] S. Concas, S. M. La Cava, R. Casula, G. Orrù, G. Puglisi, G. L. Marcialis, Quality-based artifact

modeling for facial deepfake detection in videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3845–3854.
[58] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Niessner, Faceforensics++: Learning to detect manipulated facial images, volume 2019-October, 2019, p. 1 – 11.
[59] A. Haliassos, R. Mira, S. Petridis, M. Pantic, Leveraging real talking faces via self-supervision for robust forgery detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 14950–14962.
[60] A. Haliassos, K. Vougioukas, S. Petridis, M. Pantic, Lips don't lie: A generalisable and robust approach to face forgery detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 5039–5049.
[61] J. Gao, M. Micheletto, G. Orrù, S. Concas, X. Feng, G. L. Marcialis, F. Roli, Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection, Engineering Applications of Artificial Intelligence 133 (2024) 108450.
[62] J. Gao, Z. Xia, G. L. Marcialis, C. Dang, J. Dai, X. Feng, Deepfake detection based on high-frequency enhancement network for highly compressed content, Expert Systems with Applications 249 (2024) 123732.
[63] J. Battocchio, S. Dell'Anna, A. Montibeller, G. Boato, Advance fake video detection via vision transformers, in: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, 2025, pp. 1–11.
[64] S. Concas, S. M. La Cava, A. Panzino, E. Masala, G. Orrù, G. L. Marcialis, Deceptive beauty: Evaluating the impact of beauty filters on deepfake and morphing attack detection, arXiv preprint arXiv:2509.14120 (2025).
[65] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems 25 (2012).
[66] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
[67] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df: A large-scale challenging dataset for deepfake forensics, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 3207–3216.
[68] T. Neubert, A. Makrushin, M. Hildebrandt, C. Kraetzer, J. Dittmann, Extended stirtrace benchmarking of biometric and forensic qualities of morphed face images, Iet Biometrics 7 (2018) 325–332.
[69] D. Salvi, H. Liu, S. Mandelli, P. Bestagini, W. Zhou, W. Zhang, S. Tubaro, A robust approach to multimodal deepfake detection, Journal of Imaging 9 (2023). doi:10.3390/jimaging9060122.
[70] T. M. Wani, S. A. A. Qadri, F. A. Wani, I. Amerini, Navigating the soundscape of deception: A comprehensive survey on audio deepfake generation, detection, and future horizons, Foundations and Trends® in Privacy and Security 6 (2024) 153–345. URL: http://dx.doi.org/10.1561/3300000048. doi:10.1561/3300000048.
[71] L. Maiano, A. Benova, L. Papa, M. Stockner, M. Marchetti, G. Convertino, G. Mazzoni, I. Amerini, Human versus machine: A comparative analysis in detecting artificial intelligence-generated images, IEEE Security &; Privacy 22 (2024) 77–86. doi:10.1109/MSEC.2024.3390555.
[72] F. Pro, N. Dionelis, L. Maiano, B. L. Saux, I. Amerini, A semantic segmentation-guided approach for ground-to-aerial image matching, in: IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium, 2024, pp. 2630–2635. doi:10.1109/IGARSS53475.2024.10642526.
[73] E. Mule, M. Pannacci, A. G. Goudarzi, F. Pro, L. Papa, L. Maiano, I. Amerini, Enhancing ground-to-aerial image matching for visual misinformation detection using semantic segmentation, in: Proceedings of the Winter Conference on Applications of Computer Vision (WACV) Workshops, 2025, pp. 795–803.
[74] L. Cirillo, C. Schiavella, L. Papa, P. Russo, I. Amerini, Shedding light on depth: Explainability assessment in monocular depth estimation, in: 2025 International Joint Conference on Neural Networks (IJCNN), 2025, pp. 1–8. doi:10.1109/IJCNN64981.2025.11228948.
[75] K. Yao, J. Wang, B. Diao, C. Li, Towards understanding the generalization of deepfake detectors

from a game-theoretical view, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2031–2041. doi:10.1109/ICCV51070.2023.00194.

[76] Z. Yan, Y. Zhang, Y. Fan, B. Wu, Ucf: Uncovering common features for generalizable deepfake detection, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22355–22366. doi:10.1109/ICCV51070.2023.02048.

[77] U. Ojha, Y. Li, Y. J. Lee, Towards universal fake image detectors that generalize across generative models, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24480–24489. doi:10.1109/CVPR52729.2023.02345.

[78] D. Zhang, D. Li, A. K. Sangaiah, F. Li, Z. Deng, C. Wu, Generalizing face forgery detection by suppressed texture network with two-branch convolution, IEEE Transactions on Computational Social Systems 12 (2025) 1330–1338. doi:10.1109/TCSS.2024.3441251.

[79] K. Yao, J. Wang, B. Diao, C. Li, Towards understanding the generalization of deepfake detectors from a game-theoretical view, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2031–2041. doi:10.1109/ICCV51070.2023.00194.

[80] L. Maiano, F. Casadei, I. Amerini, Enhancing abnormality identification: Robust out-of-distribution strategies for deepfake detection, 2025. URL: https://arxiv.org/abs/2506.02857. arXiv:2506.02857.

[81] T. M. Wani, I. Amerini, Multi-scale self-supervised learning for efficient audio deepfake detection, IEEE Signal Processing Letters (2025) 1–5. doi:10.1109/LSP.2025.3634032.

[82] G. Leporoni, L. Maiano, L. Papa, I. Amerini, A guided-based approach for deepfake detection: Rgb-depth integration via features fusion, Pattern Recognition Letters 181 (2024) 99–105. doi:https://doi.org/10.1016/j.patrec.2024.03.025.

[83] L. Mongelli, L. Maiano, I. Amerini, CMDD: A novel multimodal two-stream CNN deepfakes detector, in: M. Petrocchi, M. Viviani (Eds.), Proceedings of the 4th Workshop on Reducing Online Misinformation through Credible Information Retrieval co-located with the 46th European Conference on Information Retrieval, ROMCIR@ECIR 2024, Glasgow, UK, March 24, 2024, volume 3677 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 17–30.

[84] L. Cirillo, A. Gervasio, I. Amerini, Explainability-driven adversarial robustness assessment for generalized deepfake detectors, EURASIP Journal on Information Security 23 (2025). URL: https://doi.org/10.1186/s13635-025-00211-9. doi:10.1186/s13635-025-00211-9.

[85] S. Dell'Anna, A. Montibeller, G. Boato, Truefake: A real world case dataset of last generation fake images also shared on social networks, arXiv preprint arXiv:2504.20658 (2025).

[86] A. Montibeller, D. Shullani, D. Baracchi, A. Piva, G. Boato, Bridging the gap: A framework for real-world video deepfake detection via social network compression emulation, in: Proceedings of the 1st on Deepfake Forensics Workshop: Detection, Attribution, Recognition, and Adversarial Challenges in the Era of AI-Generated Media, 2025, pp. 29–36.

[87] T. M. Wani, I. Amerini, Audio deepfake detection: A continual approach with feature distillation and dynamic class rebalancing, in: A. Antonacopoulos, S. Chaudhuri, R. Chellappa, C.-L. Liu, S. Bhattacharya, U. Pal (Eds.), Pattern Recognition, Springer Nature Switzerland, Cham, 2025, pp. 211–227.

[88] F. Tassone, L. Maiano, I. Amerini, Continuous fake media detection: Adapting deepfake detectors to new generative techniques, 2024. arXiv:2406.08171.

[89] D. A. Coccomini, G. K. Zilos, G. Amato, R. Caldelli, F. Falchi, S. Papadopoulos, C. Gennaro, Mintime: Multi-identity size-invariant video deepfake detection, IEEE Transactions on Information Forensics and Security (2024) 1–1. doi:10.1109/TIFS.2024.3409054.

[90] A. Ciamarra, R. Caldelli, F. Becattini, L. Seidenari, A. Del Bimbo, Deepfake detection by exploiting surface anomalies: The surfake approach, in: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), 2024, pp. 1024–1033. doi:10.1109/WACVW60836.2024.00112.

[91] F. Marra, C. Saltori, G. Boato, L. Verdoliva, Incremental learning for the detection and classification of gan-generated images, in: 2019 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2019, pp. 1–6.

[92] D. A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, G. Amato, Exploring strengths and weaknesses

of super-resolution attack in deepfake detection, in: European Conference on Computer Vision, Springer, 2024, pp. 351–362.

[93] A. Ciamarra, R. Caldelli, A. Del Bimbo, On the generalisation capability of local surface frames in detecting diffusion-based facial images, in: Proceedings of the Winter Conference on Applications of Computer Vision, 2025, pp. 1402–1411.

[94] G. Affatato, A. Ciamarra, E. D. Cannas, S. Mandelli, B. Tondi, R. Caldelli, P. Bestagini, et al., 3d morphable models meet surface frames for generalizable and robust deepfake detection, in: European Signal Processing Conference (EUSIPCO), 2025, pp. 1223–1227.

[95] D. Ghiani, J. D. R. Chivata, S. Lilliu, S. M. La Cava, M. Micheletto, G. Orrù, F. Lama, G. L. Marcialis, Fragile watermarking for image certification using deep steganographic embedding, arXiv preprint arXiv:2504.13759 (2025).

[96] P. Perazzo, M. Mattei, G. Anastasi, M. Avvenuti, G. Dini, G. Lettieri, C. Vallati, Jpegs just got snipped: Croppable signatures against deepfake images, arXiv preprint arXiv:2512.01845 (2025).

[97] M. Garofalo, A. Catalfamo, M. Colosi, M. Villari, Federated objective: Assessing client truthfulness in federated learning, in: 2024 IEEE International Conference on Big Data (BigData), IEEE, 2024, pp. 7755–7763.

[98] M. Colosi, A. Catalfamo, M. Garofalo, M. Villari, Enabling flower for federated learning in web browsers in the cloud-edge-client continuum, in: 2024 IEEE/ACM 17th International Conference on Utility and Cloud Computing (UCC), IEEE, 2024, pp. 290–299.