Deepfake Media Forensics: State of the Art and Challenges Ahead

Irene Amerini[5], Mauro Barni[8], Sebastiano Battiato[1], Paolo Bestagini[6], Giulia Boato[9], Tania Sari Bonaventura[5], Vittoria Bruni[5], Roberto Caldelli[2], Francesco De Natale[9], Rocco De Nicola[10], Luca Guarnera[1], Sara Mandelli[6], Gian Luca Marcialis[4], Marco Micheletto[4], Andrea Montibeller[9], Giulia Orrù[4], Alessandro Ortis[1], Pericle Perazzo[3], Giovanni Puglisi[4], Davide Salvi[6], Stefano Tubaro[6], Claudia Melis Tonti[5], Massimo Villari[7], Domenico Vitulano[5]

[1] University of Catania, [2] CNIT, Florence, and Universitas Mercatorum, [3] University of Pisa, [4] University of Cagliari, [5] Sapienza University of Rome, [6] Politecnico di Milano, [7] University of Messina, [8] Università di Siena, [9] University of Trento, [10] Scuola IMT Alti Studi Lucca

Abstract. AI-generated synthetic media, also called Deepfakes, have significantly influenced so many domains, from entertainment to cybersecurity. Generative Adversarial Networks (GANs) and Diffusion Models (DMs) are the main frameworks used to create Deepfakes, producing highly realistic yet fabricated content. While these technologies open up new creative possibilities, they also bring substantial ethical and security risks due to their potential misuse. The rise of such advanced media has led to the development of a cognitive bias known as Impostor Bias, where individuals doubt the authenticity of multimedia due to the awareness of AI's capabilities. As a result, Deepfake detection has become a vital area of research, focusing on identifying subtle inconsistencies and artifacts with machine learning techniques, especially Convolutional Neural Networks (CNNs). Research in forensic Deepfake technology encompasses five main areas: detection, attribution and recognition, passive authentication, detection in realistic scenarios, and active authentication. This paper reviews the primary algorithms that address these challenges, examining their advantages, limitations, and future prospects.

Keywords: Multimedia Forensics · Deepfakes.

1 Introduction

The advent of Deepfakes, synthetic media generated by Artificial Intelligence (AI) that mimics real images, audio, and video, has significantly impacted various domains including entertainment, politics, and cybersecurity. Deepfakes leverage deep learning techniques, particularly GANs [30] and DMs [39], to create highly convincing but falsified representations of individuals. While these technologies offer creative opportunities, they also pose serious ethical and security challenges due to their potential for misuse. The emergence of such advanced AI-generated

media has led to the development of a cognitive bias known as the *Impostor Bias* [15], which refers to the tendency to doubt the veracity of multimedia elements due to the knowledge that they can be realistically generated by AI models. Deepfake detection has become an essential field of research, aiming to develop methods to distinguish between real and artificially generated media. Techniques for Deepfake detection often involve analyzing inconsistencies and artifacts that are not easily perceptible to the human eye [36, 22] but can be detected using proper detectors based on machine learning algorithms. These detection methods typically focus on both spatial and temporal anomalies in the data, utilizing Convolutional Neural Networks (CNNs) [84, 34] for enhanced accuracy. Starting with the Deepfake detection task, the scientific community has over the years taken on several other new challenges to study the nature of synthetic data in detail. We can therefore distinguish 5 main areas of research in the Forensic Deepfake domain, namely *Deepfake Detection* (Section 2) *Deepfake Attribution And Recognition* (Section 3), *Passive Deepfake Authentication Methods* (Section 4), *Deepfakes Detection Method On Realistic Scenarios* (Section 5.3), and *Active Authentication* (Section 6.3).

In this context, authors of the proposed papers are involved in the FF4ALL initiative (FF4ALL - Detection of Deep Fake Media and Life-Long Media Authentication), which aims to develop theoretical and practical tools for detecting and combating media counterfeits or Deepfakes, tracing their origin and limiting their dissemination. In the following sections, a brief overview of the main algorithms that aim to address the above-mentioned challenges will be presented.

2 Deepfake Detection

Deepfake technology poses significant challenges due to its potential for misuse, which can severely impact public well-being and trust. While current detection methods, primarily based on convolutional neural networks and deep learning paradigms, have shown promising results, they often struggle to generalize across the varied techniques employed in digital content manipulation. This issue primarily arises from the intricate interplay between textures and artifacts in Deepfake data, which traditional detection methods frequently overlook. In this context, artifacts are unintentional distortions or irregularities that occur during the Deepfake generation process, including unusual pixel formations or edge anomalies. Conversely, textures refer to the inherent patterns and fine details present in authentic images, such as the natural appearance of skin and hair. In fact, it is well established that synthetic manipulations typically disrupt the texture consistency of original images [77] and often leave detectable traces in the form of artifacts in both spatial [18] and frequency domains [23], particularly in specific facial regions [81]. Consequently, numerous studies focus their analysis on specific portions of face images to identify these inconsistencies. One promising approach involves using both No-Reference (NR) and Full-Reference (FR) quality measures to detect subtle manipulations in video frames [21]. This method has significantly improved cross-manipulation generalization by focusing on areas susceptible to artifacts, such as the mouth and eyes, and analyzing

the image quality degradation caused by Deepfake algorithms. In addition to artifacts, texture analysis provides another robust basis for distinguishing between real and fake images. In some Deepfake technologies dedicated to face-swapping operations, the inner and outer faces have different identities, making texture inconsistencies particularly evident [48]. However, focusing exclusively on either artifacts or textures in Deepfake detection can be limiting. While these approaches yield high accuracy in specific contexts, they often fail to adapt to new and evolving Deepfake techniques. To address this limitation, a novel framework called the Texture and Artifact Detector (TAD) has been proposed [26]. The TAD framework enhances Deepfake detection by leveraging both texture and artifact inconsistencies, thereby improving model generalization across various forgery scenarios through ensemble learning. Unfortunately, the performance of these methods is often hindered by the challenges posed by highly compressed data. High compression ratios can obscure subtle manipulations, leading to a significant degradation in detection accuracy [41]. A promising solution involves leveraging a learnable adaptive high-frequency enhancement framework to enrich weak high-frequency details in compressed content, thereby enhancing the robustness of Deepfake detection under compression [27]. Further details on compression impacts and related detection strategies will be discussed in subsequent sections.

3 Deepfake Attribution and Recognition

3.1 Deepfake Fingerprint and Attribution

Deepfake attribution, often referred to as Deepfake Model Recognition [37, 69], encompasses methodologies capable of identifying the specific model used to generate synthetic data. This process includes attempts to estimate the unique model weights [5] of the architecture instance responsible for creating the Deepfake. SOTA techniques are highly effective in detecting Deepfake content generated by widely-used GANs [32, 33, 28] and DMs [35, 68]. These techniques can even specialize to recognize the specific architectures, and, in more details, the specific model used in the creation procedure. Then, a more advanced challenge in this domain is identifying the exact model instance, characterized by a unique set of weights and parameters, within a given architecture: Guarnera et al. [37] demonstrated that using a simple ResNET-18 [38] engine combined with a metric learning approach [53], excellent results can be achieved in identifying the specific model used for creating synthetic data from 100 different instances of StyleGAN2-ADA [44]. A robust model recognition solution would enable the attribution of an image to a specific model owner, which is crucial for intellectual property rights [49]. To establish the ownership or authenticity of an image generated by a particular model within a specific architecture, new strategies and appropriate metrics are required [43]. In the context of forensic investigations involving Deepfake images, videos, or audio, state-of-the-art Deepfake detectors and architecture classifiers can be likened to the task of identifying camera models in traditional forensic analysis. Deepfake model recognition aims to trace the

origin of a Deepfake to a specific model instance within an architecture. This parallel underscores the necessity of developing advanced techniques for Deepfake model attribution to ensure authenticity in digital media.

4 Passive Deepfake Authentication Methods

In the modern era, where video calls have become a cornerstone of global communication, the importance of authenticating audio and video streams cannot be overstated. The advent of Deepfake technology poses a significant challenge to the integrity of digital communication. Traditional Deepfake detection methods may fall short as they often focus on either audio or video data in isolation. However, Deepfakes may involve sophisticated manipulations of both audio and video streams, making them harder to detect with monomodal methods.

This highlights the need for a multimodal approach that simultaneously analyzes both audio and visual data [73]. By correlating information from these two channels, we can significantly improve the accuracy of Deepfake detection. This approach takes advantage of the fact that inconsistencies are often more noticeable when multiple modalities of data are considered together. For instance, [40] leverages the incongruity between emotional cues portrayed by audio and visual modalities. In [4], the authenticity of a speaker is verified by detecting anomalous correspondences between his/her facial movements and what he/she says. Moreover, the results of [46] show that an ensemble of audio and visual baselines outperforms monomodal counterparts.

Given that audio-visual authentication methods may exploit monomodal detectors in a synergistic fashion, another step towards better performance is to enhance monomodal audio or visual detectors separately through the use of advanced techniques. Concerning the visual component, it is possible to leverage fusion of multiple detectors trained on purpose to capture orthogonal traces [60]. Concerning audio, it is possible to exploit modern solutions such as transformers [76], as well as investigating the use of semantic traces [7]. Despite the great effort of the multimedia forensics community, a series of challenges remains. Concerning multimodal solutions, the need for audio-video Deepfake datasets is becoming more than an urgent necessity. Indeed, most of the effort has been put towards monomodal datasets creation. Moreover, given the trend of large language models, it could be interesting to try using the same logic for audio visual reasoning. Concerning monomodal solutions, explainability has definitely not been reach yet, which still proves a problem in case of court of laws.

5 Deepfakes Detection Method on Realistic Scenarios

5.1 Deepfake Detection of image-videos in the wild

In recent years, there has been a growing interest in the study of techniques for the detection of Deepfake and AI-generated media [65, 57]. Consequently, numerous solutions have been proposed to address the problems posed by the increasing spread of fake multimedia content. However, most of these solutions

perform well only in controlled settings, such as laboratory experiments, but fail to provide reliable results in real-life conditions typical of practical applications. Deep Learning (DL) models can effectively detect Deepfake media and identify their source. However, despite promising results, DL-based methods face several significant challenges, particularly in real-world applications where controlled laboratory conditions are absent. Firstly, DL models require vast amounts of labeled data for training, which is often difficult to obtain in real-life scenarios. Additionally, these models must handle unforeseen situations that were not accounted for during training, a common issue in multimedia forensics applied outside the lab. For DL models to be effective in the ongoing battle between forensic analysts and counterfeiters, it is crucial to address the risk of overfitting to training data, which can lead to failures in new, unexpected situations. While it is possible to train highly accurate detectors, these methods struggle to generalize to new generative techniques due to data drift [64]. Detectors perform well on the techniques they were trained on but often fail with content from new generative models. Another major obstacle is the black-box nature of DL techniques, making it difficult to interpret analysis results and understand decision-making processes. This lack of transparency hampers the practical application of DL in scenarios where accountability is essential. To address these issues, researchers address their attention on several strategies, including the use of one-class classifiers trained only on pristine images [47, 1], developing classifiers with rejection options to opt out when encountering unfamiliar inputs not well-represented in the training set [58, 2] and adopting methods capable of generalization though features fusion [50] together with multimodal approaches that combine audio and video streams [63, 85]. Even considering these efforts, the practical application of automatic detectors has been minimal. Deploying these tools in commercial or mass verification systems presents numerous challenges beyond generalizing from a few known benchmarks [72, 22, 51, 87]. One significant challenge is the need to continuously train these models on new generative/Deepfake techniques in a continual learning fashion [79]. Continual learning, also known as lifelong or incremental learning, is an ongoing approach to maintaining good model performance on evolving tasks without experiencing *catastrophic forgetting* [25]. This approach is well-suited to recognize content generated by new techniques and continuously adjusting models to account for data drift observed during inference versus training. A promising direction is the creation of an end-to-end Deepfake detection system that supports continuous integration and continuous delivery/deployment (CI/CD) with the design of a Machine Learning Model Operations (*MLOps* [64, 74]) pipeline, enabling the end-to-end development of continuously trained and monitored intelligent detectors with a minimal set of components.

5.2 Deepfake and Social Media

An extremely challenging “real-world" case scenario involves the detection of Deepfake multimedia shared on social networks [11, 62, 67]. In fact, to cope with

bandwidth and storage limitations [11], social networks apply severe data compression and resizing. However, such processing, while reducing the overall multimedia size, also reduces the presence of forensic features used for the discrimination of real vs. fake multimedia [11, 83, 59]. A first study on GAN images [62] shared on Twitter demonstrates the adversarial effects of social network compression on Deepfake detectors. Specifically, while the visual quality of the shared images was untouched, the presence of forensic traces and patterns was reduced. The effects of social network sharing were then extensively observed and studied in [11]. In [67], the authors explore the challenges and advancements in the field of media forensics as applied to social network. The study addresses the increasing concerns regarding the authenticity and reliability of digital media shared on social platforms, focusing on challenges affecting source attribution algorithms [55] as well as multimedia verification [83]. For the latter, additional effort was devoted to studying whether the multimedia content is consistent with its descriptive text. The paper [67] discusses the emerging challenges in the field, such as the rise of Deepfakes and the use of bots for spreading disinformation. The authors highlight the need for advanced forensic tools to keep up with these sophisticated methods of media manipulation. In [11], the authors composed a large and diverse dataset counting 80k fake images generated with StyleGAN models and 70k real images collected from several state-of-the-art datasets [45]. In addition to this, while revealing insightful details on the entity and severity of social network compression applied by Twitter, Facebook and Telegram, the authors shown how their dataset can be used to finetuning new detectors, preserving their accuracy on social network compressed images while without experiencing "*catastrophic forgetting loss*" [25]. Interestingly, [8] showed that while social networks degrade the presence of forensic artifacts used by real vs. fake detectors, they introduce other traces that can be exploited to reconstruct the life cycle of multimedia and determine on which social networks it has been shared. While the life cycle of multimedia does not provide specific information on its nature (i.e., real or fake), it can be fundamental in recovering the version of that multimedia closer to the original, unshared one. This, in turn, allows for more accurate real vs. fake detection. Finally, preliminary studies have been conducted also on videos shared on social netowrks [61], showing similar effects to those on images. One of these works [61, 42] studies the effects of social network compression on FaceForensics [61] videos shared on Facebook and Youtube. The study provides a results in line with what observed on images [61] and a new dataset of shared videos to be used to finetune real vs. fake detectors. Nevertheless, while new works on Deepfake detection on social networks are available, the continuous update of social network compression algorithms makes the arms race increasingly challenging. As a consequence, this require additional effort to develop new architectures and updated datasets of social network shared images.

5.3 Detection of Deepfake Images and Videos in Adversarial Setting

An additional problem affecting virtually all the Deepfake forensic techniques developed so far is that such techniques are thought to operate in a benign set-

ting, that is, by neglecting the possible efforts made by an adversary to mislead the forensic analysis. Yet, recent researches [78, 29] have shown how easy is to generate adversarial contents capable of deceiving image and video processing techniques based on DL, when the adversary is informed about the details of the tools employed by the analyst. Some works [66, 9] have also studied the transferability of adversarial examples to networks different than those targeted by the attack, opening the way to the development of powerful attacks even when the attacker is unaware, or only partially aware, of the techniques used by the analyst. Even worse, often it is not even necessary that the adversary applies sophisticated attacks relying on the full or partial knowledge of the to-be-attacked system. By relying on the lack of robustness and the generalization capabilities of the forensic tools already outlined in Section 5.1, the adversary may simply process the deepkfake content in such way to prevent a correct forensic analysis, or at least degrade its performance up to a point to make it unusable. Some examples of this kind of attacks, often referred to as *laundering attacks*, include the application of moderate to strong lossy compression, geometric processing of images and videos, noise addition, histogram stretching and many others.

Understanding and ensuring the security of Deepfake forensic tools is a crucial problem, if such tools have to be used under the intrinsically adversarial conditions typical of multimedia forensics applications. For this reason, several efforts have been made to defend against adversarial attacks [14, 52], both in the realm of computer vision applications and multimedia forensics. Still, no general effective solutions have been found yet [6]. Among the solutions developed so far, adversarial training [56] has received some consensus and has proven to at least mitigate the effectiveness of adversarial attacks in computer vision applications. As argued in [82], adversarial training forces DL models to focus on robust, possibly semantic, features, which are inherently more difficult to attack. Whether such a beneficial effect of adversarial training also applies to Deepfake forensic applications is still an open problem. It is no clear, in fact, if in multimedia forensics the equivalent of semantic computer vision features exist or not.

With regard to laundering attacks, the solutions proposed so far are similar to those already discussed in Section 5.1, given the, ultimately, the effectiveness of laundering attacks can be drastically reduced by improving the robustness and generalization capabilities of the forensic tools. A common approach to do so, involves the use of data augmentation techniques that enrich the training set with processed samples, thus improving the robustness of the forensic tools against the processing operators included in the data augmentation procedure. Yet, accounting for *all* possible kinds of processing during training is clearly unfeasible. Among the solutions proposed so far, we mention the possibility of identifying a kind of *worst possible laundering attack* and include it in the training procedure. Examples of such an approach are described in [10, 70]. Despite all the efforts made, even for laundering attacks, a definitive solution has not been devised yet, thus adding yet another point to the *to-do* list of multimedia forensic researchers.

6 Active Authentication

6.1 Active deepefake detection

Passive Deepfake detection techniques [20, 19] work a posteriori, after that the forged content has been generated, distributed and possibly processed, on the contrary, active methods work in a preemptive way, pre-processing the media in such a way to ease the subsequent analysis. This is the case, for instance, of Deepfake detection methods based on DNN watermarking, whereby the content generated by DNNs is watermarked in such a way to ease the distinction between genuine and fake media, and the attribution of the fake content to the network which generated it. An alternative possibility is to modify the computational imaging chain characterizing modern acquisition devices, to insert within the generated content a unique fingerprint to be used later on for authentication purposes. This means a change of paradigm that needs to be properly explored. Active authentication techniques represent a valid, and more reliable, alternative (or complement) to passive authentication, whenever the operating conditions allow their use.

Watermarking has recently been proposed as a means to protect the IPR of DNNs [24]. By tying a watermark to a DNN model, in fact, it would be possible to prove the ownership of the model or trace its illegal use. On this basis, DNN watermarking can be also used to link AI-generated contents [86], like Deepfakes, to the model which generated them, thus providing an easy and convincing way to distinguish between synthetic and natural contents. Such a goal is achieved by requiring that all the contents generated by a network contain a predefined watermark (a kind of synthetic fingerprint) that can be used later on to distinguish the synthetic images (or videos) generated by a trained model from real ones. This marks a drastic paradigm change with respect to current solutions based on multimedia forensics, since authentication is now achieved with the active help of the party which trained the media-generation network. Though some solutions have appeared in this direction, putting this idea at work requires that considerable advances are made particularly in terms of watermark robustness against image, video and network manipulations, security against adversarial attacks, payload and also imperceptibility. Succeeding into designing a new class of robust and secure solutions, based on active approaches, for Deepfake detection/attribution surely represents an open challenge and an interesting opportunity for scientific research in the field of multimedia forensics.

6.2 Efficient Media Origin Authentication

Customary deep-fake detection methods, both passive and active, are subject to false positives and false negatives, whose rate highly depends on the employed method and the goodness of the training data. False positives are due to various factors, such as the complexity of the content, the quality of the training data, or the intrinsic limitations of the detection algorithm itself. False negatives could happen if the deep-fake is very well made, or in general if the detection

method fails to recognize certain patterns or features that indicate a deep-fake. On the other hand, cryptographic signatures are "almost perfect" from this point of view, in the sense that false negatives (i.e., authentic signatures which are not recognized so) are zero and false positives (i.e., fake signatures that are taken as authentic) are considered computationally infeasible to forge. This suggests that cryptographic signatures could be used fruitfully to detect deep-fakes (or better, the absence of deep-fakes) with perfect precision. In this direction, the work-in-progress standard JPEG Trust [80] by the Joint Photographic Experts Group (JPEG) aims to establish trust in digital media by addressing aspects of authenticity, provenance, and integrity. JPEG Trust will provide a framework for establishing trust in media through secure annotation of media assets throughout their life cycle, using cryptography as a key component. Cryptographically signing media allows deep-fakes to be repudiated by the interested person, since the signature on them will be absent or invalid. Such an anti-fake signature should allow "good" manipulation of original file (at least cropping), disallow "bad" manipulation, but also be space efficient, to save bandwidth on web servers once the media file is disseminated. Unfortunately, customary signature schemes like ECDSA do not have these properties.

To address this challenge, a solution could rely on novel aggregatable signatures, such as the Boneh-Lynn-Shacham (BLS) signature [13, 12], which has been successfully used in blockchain technologies like Ethereum 2.0 to optimize storage[1]. The BLS signature scheme makes use of a novel form of cryptography called pairing-based cryptography, which allows for a plethora of new functionalities like attribute-based encryption[31, 71, 75]. Aggregatable signatures could be employed in JPEG (possibly within the JPEG Trust standard itself) in such a way to permit benign alterations of the image like cropping while preventing malicious tampering, without increasing too much the bandwidth occupation on web servers.

6.3 Trusted Remote Media Processing on Cloud and Edge Computing Systems

In the emerging Smart Cities context, systems based on IoT (Internet of Things) play an important role to allow citizens to interact with the environment and to benefit from advanced services, such as video surveillance, intelligent traffic lightning, and air quality sensing. From a technological point of view, using sensors and actuators to automate services is strategic, but managing, configuring, and optimizing the digital infrastructures to adapt their behavior to the specific needs of the context is a big challenge, both in terms of system design and security. Deepefake media detection in these scenarios represent a challenge due to the nature of possible manipulations of future citizens day life, hence a more holistic approach should be considered where the media production occur, since

[1] https://github.com/ethereum/consensus-specs/blob/dev/specs/phase0/beacon-chain.md#bls-signatures

at the Edge Computing Systems [16]. In just 20 years, with the objective to increase system response and reduce communication latency, computation moved from mainframes and computing rooms towards Cloud Computing, Fog Computing, and lastly, Edge Computing. A Federated Cloud-Edge infrastructure is considered, where different administrative domains are in place and where Machine Learning software artifacts, in the assertion of Federated Learning even at the Edge, help to distribute intelligence in this scenario.

Multimedia acquisition devices based on IoT generate an unprecedented amount of data, with the need of developing Cloud-based video big data analytics frameworks. A distributed approach in video recording and elaboration systems, such as video surveillance systems based on IP cameras, is highly recommended to overcome the maximum storage or throughput limitation of Network Video Recorders installed on single machines. To perform such a variety of tasks, and to be able to modify a device's behavior on-demand, the Function as a Service (FaaS) computational paradigm is generally adopted. FaaS allows to define several minimal applications and to run one or more instances of these on the same device at the same time. FaaS framework relies on two configuration approaches: a local configuration file, generally YAML, or a secure remote server. However, both come with limitations: a local file configuration requires direct access to the device, physically or through a secure connection, to modify it. Alternatively, a remote server can store and send updated configuration files, but it might be vulnerable to well-known cyberattacks suck as Man-in-the-middle (MITM) or Distributed Denial of Service (DDoS), making it unusable and unreliable. To overcome such limitations, it is possible to benefit from three technologies that have been increasingly recognized to be able to address information access problems and system trustiness in different application domains: Federated Learning [3], Blockchain and IPFS (InterPlanetary File System) starting at the Edge:

- Federated Learning is a decentralized approach to training Machine Learning Models. In traditional Machine Learning, data is centralized in the Cloud, where a single model is trained on the entire dataset. Federated Learning, on the other hand, allows for training Machine Learning Models across multiple decentralized devices or servers that hold local data samples without exchanging them. Moreover, Federated Learning at the Edge refers to the application of Federated Learning techniques on Edge devices, such as IoT devices, or Edge Servers. This approach combines the benefits of Federated Learning, which ensures Data Privacy and reduces communication costs, with the advantages of Edge Computing Systems, which enables data processing and model training to occur closer to where the data is generated, hence, fake Media might not exit from the Edge.
- The use of Blockchain, supported by the flexibility and robustness of Smart Contracts, allows the combining of the well-known FaaS paradigm with the intrinsic features of data non-repudiation and immutability, replacing the service configuration with a Smart Contract, guaranteeing protection against distributed cyber-attacks [17].

– IPFS is a distributed system for storing and accessing files. Since the block size of the Blockchain does not allow storing files, these can be uploaded to this special file storage, which produces a unique hash value to be used as a key to access its content [54].

7 Conclusion

In this paper, we have conducted a comprehensive review of the state-of-the-art techniques and challenges in Deepfake media forensics. Our exploration covered the core areas of Deepfake detection, attribution and recognition, passive authentication, detection in realistic scenarios, and active authentication. Each of these areas addresses specific facets of the Deepfake phenomenon, from the identification of synthetic media and tracing their origins to ensuring the robustness of detection systems in real-world environments and embedding verifiable information within media for instant authentication. Future work will focus on conducting a more in-depth analysis of practical countermeasures and gaining deeper insights into real-world applications (e.g. highly compressed data).

Acknowledgments. This study has been partially supported by SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

Disclosure of Interests. The authors have no competing interests.

References

1. Abady, L., Dimitri, G.M., Barni, M.: A one-class classifier for the detection of gan manipulated multi-spectral satellite images. Remote Sensing **16**(5) (2024)
2. Abady, L., Wang, J., Tondi, B., Barni, M.: A siamese-based verification system for open-set architecture attribution of synthetic images. Pattern Recognition Letters **180**, 75–81 (2024)
3. Abdelgaber, Y.E., Ahmed, Y.A., Salem, M.A.M., Salem, M.A.G.: Federated learning for resource management in edge computing. In: 2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS). pp. 102–109 (2023). https://doi.org/10.1109/ICICIS58388.2023.10413933
4. Agarwal, S., Hu, L., Ng, E., Darrell, T., Li, H., Rohrbach, A.: Watch those words: Video falsification detection using word-conditioned facial motion. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2023)
5. Asnani, V., Yin, X., Hassner, T., Liu, X.: Reverse engineering of generative models: Inferring model hyperparameters from generated images. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
6. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: Proceedings of PMLR, International conference on machine learning. pp. 274–283 (2018)
7. Attorresi, L., Salvi, D., Borrelli, C., Bestagini, P., Tubaro, S.: Combining automatic speaker verification and prosody analysis for synthetic speech detection. In: International Conference on Pattern Recognition (ICPR) (2022)

8. Baracchi, D., Boato, G., De Natale, F., Iuliani, M., Montibeller, A., Pasquini, C., Piva, A., Shullani, D.: Towards open-world multimedia forensics through media signature encoding. IEEE Access (2024)
9. Barni, M., andf E. Nowroozi, K.K., Tondi, B.: On the transferability of adversarial examples against CNN-based image forensics. In: Proceedings ICASSP - IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 8286–8290 (2019)
10. Barni, M., Nowroozi, E., Tondi, B.: Higher-order, adversary-aware, double JPEG-detection via selected training on attacked samples. In: Proceedings of 25th European signal processing conference (EUSIPCO). pp. 281–285 (2017)
11. Boato, G., Pasquini, C., Stefani, A.L., Verde, S., Miorandi, D.: Trueface: a dataset for the detection of synthetic face images from social networks. In: 2022 IEEE International Joint Conference on Biometrics (IJCB). pp. 1–7 (2022). https://doi.org/10.1109/IJCB54206.2022.10007988
12. Boneh, D., Gorbunov, S., Wahby, R.S., Wee, H., Wood, C.A., Zhang, Z.: BLS Signatures. Internet-Draft draft-irtf-cfrg-bls-signature-05, Internet Engineering Task Force (Jun 2022), work in Progress
13. Boneh, D., Lynn, B., Shacham, H.: Short signatures from the weil pairing. In: Boyd, C. (ed.) Advances in Cryptology — ASIACRYPT 2001. pp. 514–532. Springer Berlin Heidelberg, Berlin, Heidelberg (2001)
14. Bountakas, P., Zarras, A., Lekidis, A., Xenakis, C.: Defense strategies for adversarial machine learning: A survey. Computer Science Review (2023)
15. Casu, M., Guarnera, L., Caponnetto, P., Battiato, S.: Genai mirage: The impostor bias and the deepfake detection challenge in the era of artificial illusions. Forensic Science International: Digital Investigation **50**, 301795 (2024). https://doi.org/https://doi.org/10.1016/j.fsidi.2024.301795
16. Catalfamo, A., Celesti, A., Fazio, M., Randazzo, G., Villari, M.: A platform for federated learning on the edge: a video analysis use case. In: 2022 IEEE Symposium on Computers and Communications (ISCC). pp. 1–7 (2022). https://doi.org/10.1109/ISCC55528.2022.9912968
17. Catalfamo, A., Ruggeri, A., Celesti, A., Fazio, M., Villari, M.: A microservices and blockchain based one time password (mbb-otp) protocol for security-enhanced authentication. In: 2021 IEEE Symposium on Computers and Communications (ISCC). pp. 1–6 (2021). https://doi.org/10.1109/ISCC53001.2021.9631479
18. Chai, L., Bau, D., Lim, S.N., Isola, P.: 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 16. pp. 103–120. Springer (2020)
19. Ciamarra, A., Caldelli, R., Becattini, F., Seidenari, L., Del Bimbo, A.: Deepfake detection by exploiting surface anomalies: The surfake approach. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). pp. 1024–1033 (2024). https://doi.org/10.1109/WACVW60836.2024.00112
20. Coccomini, D.A., Zilos, G.K., Amato, G., Caldelli, R., Falchi, F., Papadopoulos, S., Gennaro, C.: Mintime: Multi-identity size-invariant video deepfake detection. IEEE Transactions on Information Forensics and Security pp. 1–1 (2024). https://doi.org/10.1109/TIFS.2024.3409054
21. Concas, S., La Cava, S.M., Casula, R., Orrù, G., Puglisi, G., Marcialis, G.L.: Quality-based artifact modeling for facial deepfake detection in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3845–3854 (2024)

22. Corvi, R., Cozzolino, D., Poggi, G., Nagano, K., Verdoliva, L.: Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 973–982 (2023)
23. Durall, R., Keuper, M., Keuper, J.: 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. pp. 7890–7899 (2020)
24. Fei, J., Xia, Z., Tondi, B., Barni, M.: Supervised gan watermarking for intellectual property protection. In: 2022 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–6 (2022)
25. French, R.M.: Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences **3**(4), 128 – 135 (1999)
26. Gao, J., Micheletto, M., Orrù, G., Concas, S., Feng, X., Marcialis, G.L., Roli, F.: Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection. Engineering Applications of Artificial Intelligence **133**, 108450 (2024)
27. Gao, J., Xia, Z., Marcialis, G.L., Dang, C., Dai, J., Feng, X.: Deepfake detection based on high-frequency enhancement network for highly compressed content. Expert Systems with Applications **249**, 123732 (2024)
28. Giudice, O., Guarnera, L., Battiato, S.: Fighting deepfakes by detecting gan dct anomalies. Journal of Imaging **7**(8), 128 (2021)
29. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv:1412.6572 (2014)
30. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. Advances in Neural Information Processing Systems **27** (2014)
31. Goyal, V., Pandey, O., Sahai, A., Waters, B.: Attribute-based encryption for fine-grained access control of encrypted data. In: Proceedings of the 13th ACM Conference on Computer and Communications Security. p. 89–98. CCS '06, Association for Computing Machinery, New York, NY, USA (2006). https://doi.org/10.1145/1180405.1180418
32. Guarnera, L., Giudice, O., Battiato, S.: Deepfake detection by analyzing convolutional traces. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 666–667 (2020)
33. Guarnera, L., Giudice, O., Battiato, S.: Fighting deepfake by exposing the convolutional traces on images. IEEE Access **8**, 165085–165098 (2020)
34. Guarnera, L., Giudice, O., Battiato, S.: Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models. arXiv preprint arXiv:2303.00608 (2023)
35. Guarnera, L., Giudice, O., Battiato, S.: Mastering deepfake detection: A cutting-edge approach to distinguish gan and diffusion-model images. ACM Transactions on Multimedia Computing, Communications and Applications (2024)
36. Guarnera, L., Giudice, O., Nastasi, C., Battiato, S.: Preliminary forensics analysis of deepfake images. In: 2020 AEIT international annual conference (AEIT). pp. 1–6. IEEE (2020)
37. Guarnera, L., Giudice, O., Nießner, M., Battiato, S.: On the exploitation of deepfake model recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 61–70 (2022)

38. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
39. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
40. Hosler, B., Salvi, D., Murray, A., Antonacci, F., Bestagini, P., Tubaro, S., Stamm, M.C.: Do deepfakes feel emotions? a semantic approach to detecting deepfakes via emotional inconsistencies. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
41. Hu, J., Liao, X., Wang, W., Qin, Z.: Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network. IEEE Transactions on Circuits and Systems for Video Technology **32**(3), 1089–1102 (2021)
42. Hu, J., Liao, X., Wang, W., Qin, Z.: Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network. IEEE Transactions on Circuits and Systems for Video Technology **32**(3), 1089–1102 (2022). https://doi.org/10.1109/TCSVT.2021.3074259
43. Huang, Z., Li, B., Cai, Y., Wang, R., Guo, S., Fang, L., Chen, J., Wang, L.: What can discriminator do? towards box-free ownership verification of generative adversarial networks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5009–5019 (2023)
44. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 12104–12114. Curran Associates, Inc. (2020)
45. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
46. Khalid, H., Kim, M., Tariq, S., Woo, S.S.: Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In: Workshop on synthetic multimedia-audiovisual deepfake generation and detection (2021)
47. Khalid, H., Woo, S.S.: Oc-fakedect: Classifying deepfakes using one-class variational autoencoder. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 2794–2803 (2020)
48. La Cava, S.M., Orrù, G., Drahansky, M., Marcialis, G.L., Roli, F.: 3d face reconstruction: The road to forensics. ACM Computing Surveys **56**(3), 1–38 (2023)
49. Leotta, R., Giudice, O., Guarnera, L., Battiato, S.: Not with my name! inferring artists' names of input strings employed by diffusion models. In: International Conference on Image Analysis and Processing. pp. 364–375. Springer (2023)
50. Leporoni, G., Maiano, L., Papa, L., Amerini, I.: A guided-based approach for deepfake detection: Rgb-depth integration via features fusion. Pattern Recognition Letters **181**, 99–105 (2024). https://doi.org/https://doi.org/10.1016/j.patrec.2024.03.025
51. Li, C., Huang, Z., Paudel, D.P., Wang, Y., Shahbazi, M., Hong, X., Van Gool, L.: A continual deepfake detection benchmark: Dataset, methods, and essentials. p. 1339 – 1349 (2023)
52. Liang, H., He, E., Zhao, Y., Jia, Z., Li, H.: Adversarial attack and defense: A survey. Electronics (2022)
53. Liu, E.Y., Guo, Z., Zhang, X., Jojic, V., Wang, W.: Metric learning from relative comparisons by minimizing squared residual. In: 2012 IEEE 12th International Conference on Data Mining. pp. 978–983. IEEE (2012)

54. Lukaj, V., Martella, F., Fazio, M., Galletta, A., Celesti, A., Villari, M.: Gateway-based certification approach to include iot nodes in a trusted edge/cloud environment. In: 2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW). pp. 237–241 (2023)
55. Lukáš, J., Fridrich, J., Goljan, M.: Detecting digital image forgeries using sensor pattern noise. In: Security, steganography, and watermarking of multimedia contents VIII. vol. 6072, pp. 362–372. SPIE (2006)
56. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083 (2017)
57. Maiano, L., Benova, A., Papa, L., Stockner, M., Marchetti, M., Convertino, G., Mazzoni, G., Amerini, I.: Human versus machine: A comparative analysis in detecting artificial intelligence-generated images. IEEE Security &; Privacy **22**(03), 77–86 (may 2024). https://doi.org/10.1109/MSEC.2024.3390555
58. Maier, A., Riess, C.: Reliable out-of-distribution recognition of synthetic images. Journal of Imaging **10**(5), 110 (2024)
59. Maier, A., Riess, C.: Reliable out-of-distribution recognition of synthetic images. Journal of Imaging **10**(5), 110 (2024)
60. Mandelli, S., Bonettini, N., Bestagini, P., Tubaro, S.: Detecting gan-generated images by orthogonal training of multiple cnns. In: IEEE International Conference on Image Processing (ICIP). pp. 3091–3095 (2022)
61. Marcon, F., Pasquini, C., Boato, G.: Detection of manipulated face videos over social networks: A large-scale study. Journal of Imaging **7**(10), 193 (2021)
62. Marra, F., Gragnaniello, D., Cozzolino, D., Verdoliva, L.: Detection of gan-generated fake images over social networks. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 384–389 (2018)
63. Mongelli, L., Maiano, L., Amerini, I.: CMDD: A novel multimodal two-stream CNN deepfakes detector. In: Petrocchi, M., Viviani, M. (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. CEUR Workshop Proceedings, vol. 3677, pp. 17–30. CEUR-WS.org (2024)
64. Paleyes, A., Urma, R.G., Lawrence, N.D.: Challenges in deploying machine learning: A survey of case studies. ACM Computing Surveys **55**(6) (2022)
65. Papa, L., Faiella, L., Corvitto, L., Maiano, L., Amerini, I.: On the use of stable diffusion for creating realistic faces: from generation to detection. In: 11th International Workshop on Biometrics and Forensics, IWBF 2023, Barcelona, Spain, April 19-20, 2023. pp. 1–6. IEEE (2023). https://doi.org/10.1109/IWBF57495.2023.10156981
66. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia conference on computer and communications security. pp. 506–519 (2017)
67. Pasquini, C., Amerini, I., Boato, G.: Media forensics on social media platforms: a survey. EURASIP Journal on Information Security **2021**(1), 4 (2021)
68. Pontorno, O., Guarnera, L., Battiato, S.: Deepfeaturex net: Deep features extractors based network for discriminating synthetic from real images. arXiv preprint arXiv:2404.15697 (2024)
69. Pontorno, O., Guarnera, L., Battiato, S.: On the exploitation of dct-traces in the generative-ai domain. arXiv preprint arXiv:2402.02209 (2024)
70. Purnekar, N., Abady, L., Tondi, B., Barni, M.: Improving the robustness of synthetic images detection by means of print and scan augmentation. In: Proceedings of Information Hiding & Multimedia Security Conference (IH&MMSEC) (2024)

71. Rasori, M., Perazzo, P., Dini, G.: ABE-Cities: An attribute-based encryption system for smart cities. In: 2018 IEEE International Conference on Smart Computing (SMARTCOMP). pp. 65–72 (2018)
72. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Niessner, M.: Faceforensics++: Learning to detect manipulated facial images. vol. 2019-October, p. 1 – 11 (2019)
73. Salvi, D., Liu, H., Mandelli, S., Bestagini, P., Zhou, W., Zhang, W., Tubaro, S.: A robust approach to multimodal deepfake detection. Journal of Imaging **9**(6) (2023). https://doi.org/10.3390/jimaging9060122
74. Semola, R., Lomonaco, V., Bacciu, D.: Continual-learning-as-a-service (claas): On-demand efficient adaptation of predictive models. arXiv preprint arXiv:2206.06957 (2022)
75. Sicari, S., Rizzardi, A., Dini, G., Perazzo, P., La Manna, M., Coen-Porisini, A.: Attribute-based encryption and sticky policies for data access control in a smart home scenario: a comparison on networked smart object middleware. International Journal of Information Security **20**, 695–713 (2021)
76. Singh Yadav, A.K., Bhagtani, K., Baireddy, S., Bestagini, P., Tubaro, S., Delp, E.J.: Mdrt: Multi-domain synthetic speech localization. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2024)
77. Sun, X., Wu, B., Chen, W.: Identifying invariant texture violation for robust deepfake detection. arXiv preprint arXiv:2012.10580 (2020)
78. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv:1503.02531 (2013)
79. Tassone, F., Maiano, L., Amerini, I.: Continuous fake media detection: adapting deepfake detectors to new generative techniques (2024)
80. Temmermans, F., Caldwell, S., Papadopoulos, S., Pereira, F., Rixhon, P.: Towards an international standard to establish trust in media production, distribution and consumption. In: 2023 24th International Conference on Digital Signal Processing (DSP). pp. 1–5 (2023)
81. Tolosana, R., Romero-Tapiador, S., Vera-Rodriguez, R., Gonzalez-Sosa, E., Fierrez, J.: Deepfakes detection across generations: Analysis of facial regions, fusion, and performance evaluation. Engineering Applications of Artificial Intelligence **110**, 104673 (2022)
82. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A.: Robustness may be at odds with accuracy. arXiv:1805.12152 (2018)
83. Verdoliva, L.: Media forensics and deepfakes: An overview. IEEE Journal on Selected Topics in Signal Processing **14**(5), 910 – 932 (2020)
84. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8695–8704 (2020)
85. Wani, T.M., Amerini, I.: Deepfakes audio detection leveraging audio spectrogram and convolutional neural networks. In: Image Analysis and Processing - ICIAP 2023 - 22nd International Conference, ICIAP 2023, Udine, Italy, September 11-15, 2023, Proceedings, Part II. Lecture Notes in Computer Science, vol. 14234, pp. 156–167. Springer (2023)
86. Yu, N., Skripniuk, V., Abdelnabi, S., Fritz, M.: Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14428–14437 (2021). https://doi.org/10.1109/ICCV48922.2021.01418
87. Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.G.: Wilddeepfake: A challenging real-world dataset for deepfake detection. p. 2382 – 2390 (2020)