Adversarial and Generative AI-Based Anti-Forensics in Audio-Visual Deepfake Detection: A Comprehensive Review and Analysis

Qurat Ul Ain , Fatima Khalid , Hafsa Ilyas , Ali Javed ,
Khalid Mahmood Malik , Khan Muhammad , Aun Irtaza

PII: S1566-2535(25)01182-0
DOI: https://doi.org/10.1016/j.inffus.2025.104120
Reference: INFFUS 104120

To appear in: *Information Fusion*

Received date: 21 July 2025
Revised date: 28 December 2025
Accepted date: 31 December 2025

Please cite this article as: Qurat Ul Ain , Fatima Khalid , Hafsa Ilyas , Ali Javed , Khalid Mahmood Malik , Khan Muhammad , Aun Irtaza , Adversarial and Generative AI-Based Anti-Forensics in Audio-Visual Deepfake Detection: A Comprehensive Review and Analysis, *Information Fusion* (2025), doi: https://doi.org/10.1016/j.inffus.2025.104120

Highlights

- We review adversarial and anti-forensic attacks on deepfake detection systems.
- We cover fusion-based and decoy-based strategies for defense against such attacks.
- We analyze trust factors: fairness, transparency, privacy, and security in detection.
- We identify key challenges in building robust audio-visual deepfake detectors.
- We provide a roadmap to enhance trust in deepfake detection under adversarial threats.

Adversarial and Generative AI-Based Anti-Forensics in Audio-Visual Deepfake Detection: A Comprehensive Review and Analysis

Qurat Ul Ain[1], Fatima Khalid[2], Hafsa Ilyas[3], *Ali Javed[3], Khalid Mahmood Malik[4], *Khan Muhammad[5], Aun Irtaza[6]

[1]Department of Artificial Intelligence and Data Science (AI & DS), FAST School of Computing, National University of Computer & Emerging Sciences (NUCES), A. K. Brohi Road, H-11/4, 47080, Islamabad, Pakistan

[2]Faculty of Computer Science and Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, 23460, Khyber Pakhtunkhwa, Pakistan

[3]Department of Software Engineering, University of Engineering and Technology-Taxila, 47080, Punjab, Pakistan

[4]Director of Cyber Security and Professor of Computer Science, College of Innovation and Technology, University of Michigan-Flint, MI, 48502, USA

[5]VIS2KNOW Lab, Department of Applied Artificial Intelligence, School of Convergence, College of Computing and Informatics, Sungkyunkwan University, Seoul, 03063, South Korea

[6]Department of Computer Science, University of Engineering and Technology-Taxila, 47080, Punjab, Pakistan

Corresponding Authors: Ali Javed (ali.javed@uettaxila.edu.pk) and Khan Muhammad (khan.muhammad@ieee.org)

***Abstract*---**As the technology behind deepfakes advances, detecting audio-visual deepfakes becomes more and more crucial, and the rise of traditional and generative AI-based adversarial/anti-forensics attacks and generative AI-based anti-forensics attacks on deepfake detection technologies is a growing concern. Securing applications against adversarial and generative AI-based attacks is critical for accurate and robust deepfake detection tools. Therefore, this paper provides a comprehensive overview of various adversarial and generative AI-based anti-forensic attacks, which represent one of the core elements of trustworthiness alongside transparency, explainability, and fairness, as well as defensive countermeasures for audio-visual deepfake generation and detection. It covers topics such as adversarial attacks on deepfake detection algorithms and defensive methods, including model fusion and decoy-based approaches, to mitigate these threats. Although extensive research has been conducted in recent years on adversarial attacks and defense on deepfake detection, there have been few attempts to compare existing work qualitatively and quantitatively. This paper aims to help identify and address key issues that need to be considered to bring transferable adversarial attacks and their countermeasures particularly through techniques such as generative defense, knowledge distillation, and beyond.

Keywords: *Adversarial attack, Audio-visual Deepfakes, Deepfake detector, Deepfake Generator, Passive defense, Proactive defense.*

1 Introduction

Alarming advancements in generative adversarial networks (GANs), autoencoders (AEs), and diffusion models pose an existential threat to the authenticity and credibility of audio-visual content due to their ability to create remarkably convincing deepfakes. These algorithms, trained on large, real multimedia corpora, can synthesize multimedia content that is indistinguishable from authentic multimedia. The synthetic media generated through these methods has been used for a variety of purposes. For instance, deepfakes have been employed to enhance educational content with personalized, engaging, and accessible materials [1]. On the other hand, the technology has also been used for impersonation and the creation of malicious content [2], and voice-based attacks on speakers and speech verification systems [3]. The continuing advancement of deepfake generation technologies poses an existing and evolving threat to existing countermeasures. A vicious cycle exists, where countermeasure creators develop new methods for detection and malicious actors respond, using new generative algorithms or confusing the detectors by combining the power of deepfake generation models and adversarial techniques simultaneously. This convergence of generation technologies and adversarial attacks highlights the critical need for continuous innovation in detection and defensive strategies to safeguard against these increasingly sophisticated threats [4].

Because deepfake generation has an associated risk of misuse, research must also focus on the development of techniques for the detection of synthetic media [5, 6]. Therefore, researchers must continue to explore both the creation and detection of deepfakes to better understand their capabilities and thwart malicious actors. Despite aiming for robustness, deepfake detection methods remain vulnerable to adversarial attacks and anti-forensic attacks using generative attacks. This results in inaccurate output from existing detectors. Even techniques that are not purely

designed for malintent may be repurposed for adversarial attacks. In the near future, the power of multi-order deepfake generation will pose new threats to detectors. For instance, swapping a face multiple times, a process that may be used to enhance deepfakes for legitimate purposes [7], also generates imperceptible perturbations in audio and visual media, resulting in deepfakes that are deceptive. Similarly, the combined effect of deepfakes and replay attacks can make the job of an automated speaker verification system difficult [8]. Such methods have the potential to make malicious deepfakes appear real and, therefore, able to evade detection. When used maliciously, they decrease the ability to trust or rely on deepfake detection systems.

As attackers become more sophisticated in their attempts to evade detection, the number of potential attacks increases. Exploiting the weaknesses of AI-based models, attackers can create new, undetectable threats. These attacks can take many forms, for example, adversarial example generation and reconstruction-based methods [9], which can be difficult to detect and mitigate. Even subtle changes can render existing deep learning-based detectors useless. GANs are vulnerable to adversarial attacks on the discriminator network by changing the loss function, input dataset, or architecture through watermarks and tags [10, 11] The attacks themselves may be imperceptible, but the resulting model errors are catastrophic. They cause the algorithms to misclassify, rendering the solutions powerless against such threats. Therefore, the creation of trustworthy detection methods demands a more integrated approach that considers a variety of variables, such as fairness, privacy, security, explainability, and transparency [12-15].

Several strategies have been proposed in the research to protect deepfake (synthetic media) detectors against adversarial attacks. Examples include training the model on adversarial instances, incorporating an adversarial loss term into the training objective [16, 17], and employing defensive distillation [18]. However, adversarial attack resistance for generators and detectors depends on several factors, e.g., the nature of the attack, the model's architecture, and the implementation of safety measures against potential threats. In addition, a deep understanding of the adversarial attack surface is a prerequisite for the development of robust synthetic media generation and detection approaches.

Taking this into consideration, the primary aim of this survey paper is to provide a clear and organized classification of adversarial attacks on synthetic media generation and detection approaches and summarize the defensive mechanisms and evaluation criteria that are provided in the research. Going further, because an extensive classification system related to this problem does not exist, we introduce a taxonomy for a better understanding of the field and to help ongoing research to establish relationships between various overlapping components. The scope of this systematic literature review is not confined to visual content; it also provides a detailed review of adversarial attacks on synthetic audio generation and detection methods. This survey will serve as a valuable resource for researchers and practitioners interested in the field of adversarial machine learning and security. The following are the key contributions of this work:

i. This survey paper facilitates researchers by creating a comprehensive taxonomy of the diverse forms of adversarial attack and defensive techniques that can be employed on generators and detectors of audio-visual deepfakes.
ii. This survey provides an overarching view of the latest improvements, trends, and challenges in the field of adversarial attack and the defense of deepfake generators and detectors.
iii. This survey paper identifies promising directions for future research.

The rest of the paper is organized as follows: Section 2 describes the literature collection and selection criteria. Section 3 overviews the existing surveys and highlights how this study is important and timely. Section 4 presents adversarial terminologies. Section 5 gives an overview of adversarial attacks. Sections 6 and 7 discuss adversarial attacks on visual deepfake generators and detectors, respectively. Section 8 presents adversarial attacks on voice recognition systems. In Section 9, countermeasures against adversarial attacks on visual deepfake generation and detection are discussed. In Section 10, countermeasures against adversarial attacks on audio signals are discussed. Section 11 presents evaluation and perceptual similarity measures. In Section 12, we summarize the key findings and discuss future directions for both attack and defense of deepfake technologies. To enhance clarity and guide the readers through the structure of this survey, Fig. 1 presents a logical overview of the relationships among the main sections of the paper.

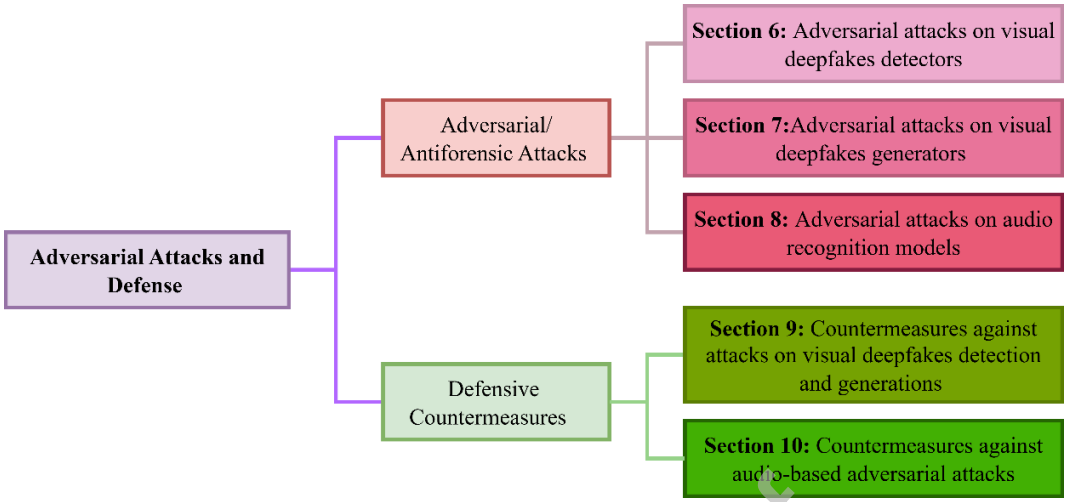

Fig. 1. Attack-to-defense taxonomy of the survey, illustrating adversarial attacks on visual deepfake detection, visual generation, and audio-based recognition systems, along with their associated countermeasures.

2 Literature collection and selection criteria

This survey provides a comprehensive review of the current research on adversarial attacks on deepfake generators and detectors, specifically focusing on studies published from 2018 till January 2025. The publishing counts have experienced substantial growth in recent years along with a rise in the popularity of the deepfake term in Google searches (Fig. 2). Specifically, the graphic exhibits Google trends data indicating the increased frequency of internet searches for “deepfake” and associated phrases over previous years. The reviewed papers focus on adversarial attacks and approaches to counter adversarial attacks on visual and audio deepfake generation and detection methods. A detailed description of the methodology and protocols used to conduct the survey is provided in Table 1.

3 Existing surveys

Here we collect and compile literature from various recent surveys in the field of deepfake technology. Some surveys concentrate solely on deepfake generation, while others examine detection, and a few address both aspects [19]. However, a major issue remains, i.e., most existing surveys do not thoroughly examine the adversarial strategies used in deepfake attacks or the corresponding countermeasures, leaving this critical dimension underexplored. For example, Rana et al. [20] concentrated exclusively on deepfake detection, examining a limited range of literature from 2018 to 2020, encompassing 112 papers over 19 pages, and omitting the broader dimensions of generation and adversarial dynamics. Pei et al. [21] and Croitoru et al. [22] addressed deepfake generation and detection, yet they neglected to discuss adversarial attacks or defensive strategies. Li et al. [23] conducted a comprehensive review of 200 papers, yet limited their analysis to deepfake detection, neglecting the aspects of generation and adversarial issues. Pham et al. [24] examine defensive strategies akin to our survey; however, their focus is confined to deepfake speech detection. Heidari et al. [25] and Gambin et al. [26], presented studies that concentrate specifically on detection and trends in deepfake technology, respectively.

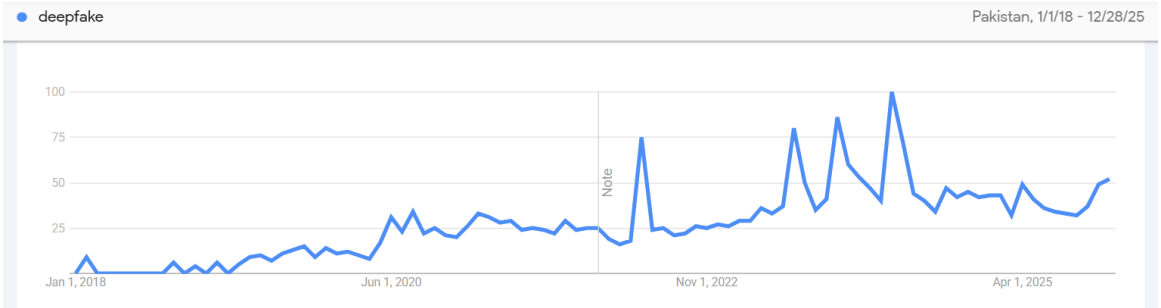

Fig. 2. Google trends graph illustrates the projected search interest for the term "deepfake" on Google from 2018 to 2025 [27].

Table 1. Detailed description of protocols representing data collection and analysis.

Preparation Protocol	**Description**
Purpose	• Offer a brief overview of the existing state-of-the-art adversarial attack generation methods and identify potential gaps. • Present a systematic review and structure to the adversarial attacks on the deepfakes detection and generation methods. • Analyse the defensive techniques to overcome adversarial attacks using deepfakes detection and generation methods. • Investigate the open challenges that exist in the domain of adversarial attacks, deepfake detection, and generation.
Data source	Google Scholar, Springer Link, ACM digital library, IEEE Explorer.
Query	The following queries were used on the above-mentioned data sources for the collection of research papers: Adversarial attacks/ Anti-forensics attacks/ Adversarial attacks and deepfakes/ Adversarial attacks and deepfakes detection/ Adversarial attacks and audio spoofing detection/ Adversarial attacks and deepfakes generation/ Adversarial attacks and GANs and deepfakes/ Anti-forensics attacks and deepfakes/ Anti-forensics attacks and deepfakes detection/ Anti-forensics attacks and deepfakes generation/ Anti-forensics attacks and GANs and deepfakes/ Anti-spoofing techniques and adversarial attacks/ Black box attacks and deepfakes/ White box attacks and deepfakes.
Method	The categorization of the literature was as follows: • Existing adversarial attack generation methods (including white box and black box attacks) and their taxonomy. • Adversarial attacks on deepfake detectors and generation methods. • Adversarial attacks on audio detectors and systems. • Countermeasures against adversarial attacks on visual and audio deepfakes detection and generation methods. • Discussion of the limitations, knowledge gap, and future directions in the domain of adversarial attacks on audio and visual deepfakes detection and generation methods.
Size	A total of 130 relevant papers were retrieved using the queries mentioned above. We further refined and selected the literature relevant to the subject of our survey.
Study type/inclusion and exclusion	The peer-reviewed journal papers, and articles of conference proceedings, were given more importance. Additionally, a few articles from archive literature were also considered.

Our survey addresses this gap, specifically, the lack of comprehensive analysis of adversarial attack techniques and their defensive strategies in both generation and detection, by examining the precise strategies for attack and defence against these attacks and offering a thorough analysis of the adversarial dimensions in audio, image, and video deepfake detectors and generators. By concentrating on this pivotal domain, our research provides a comprehensive understanding of the challenges and solutions associated with combating deepfake technology, thereby constituting a significant contribution to the field. Our research encompasses a significant timeframe from 2018 to 2025. This allows us to track the evolution of deepfake technology and the corresponding advancements in detection and prevention methods over the years. By analyzing trends and patterns over this period, we can offer valuable insights into the future trajectory of deepfake technology and potential strategies for mitigating its negative consequences. A comprehensive summary of previously reviewed surveys in deepfakes are presented in Table 2.

Table 2. A comprehensive summary of previously reviewed surveys in deepfakes. Here, "✓" denotes that the topic was covered in the survey, whereas "–" indicates that it was not addressed.

Title	**Year**	**Deepfake Generation**	**Deepfake detection**	**Adversarial Attacks on Visual Deepfake Detectors**	**Adversarial Attacks on Visual Deepfake Generators**	**Adversarial Attacks on audio deepfake detectors**	**Defensive proactive and passive approaches**	**Venue**	**Year Convergence**	**Number of Papers Reviewed**	**Number of Pages**
Our	**2025**	✓	✓	✓	✓	✓	✓	**Elsevier (INF)**	**2018-2025**	**130**	**49**
Li et al. [23]	2025	-	✓	-	-	-	-	ACM	2023-2025	200	38
Liz-Lopez [19]	2024	✓	✓	-	-	-	-	Elsevier	2018-2023	66	37
Pham et al. [24]	2024	-	✓	-	-	-	-	arXiv	2015-2025	200	25
Croitoru et al. [22]	2024	✓	✓	-	-	-	-	Springer	2020-2021	100	32
Pei et al.[21]	2024	✓	✓	-	-	-	-	arXiv	2021-2023	50	28
Heidari et al. [25]	2023	-	✓	-	-	-	-	Wieley	2020-2023	60	50
Gambin et al. [26]	2023	✓	-	-	-	-	-	Springer	2020-2023	100	32

Rana et al. [20]	2022	-	✓	-	-	-	-	IEEE	2018-2020	112	19

4 Deepfake Generation, Detection, and Adversarial Attacks

4.1 Deepfake Generation

Deepfake generation refers to the process of synthesizing realistic but manipulated data using deep generative models. This data can include visual, auditory, textual, or multimodal content. Formally, let X_r denote the distribution of real data samples. A generator $G_\theta: Z \to X_g$, parameterized by θ, maps latent variables $z \sim p_z(z)$ from a prior distribution to generated samples x_g. The objective of the generator is to approximate the real data distribution such that:

$$p_g(x) \approx p_r(x) \tag{1}$$

This goal is typically achieved through generative frameworks such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), or diffusion models. In practice, G_θ learns to produce synthetic samples that are perceptually indistinguishable from authentic data across one or more modalities. The fidelity and diversity of these generated outputs are often evaluated using quantitative metrics (e.g., FID, precision-recall) or human perceptual studies.

4.2 Deepfake Detection

Deepfake detection aims to identify or localize synthetic or manipulated data to distinguish it from authentic samples. A detector $D_\phi: X \to \{0,1\}$, parameterized by ϕ, is trained to output 0 for genuine data and 1 for fake data. The detection objective can be expressed as:

$$\max_{\phi} \mathbb{E}_{x_r \sim p_r(x)}[\log(1 - D_\phi(x_r))] + \mathbb{E}_{x_g \sim p_g(x)}[\log(D_\phi(x_g))] \tag{2}$$

Detectors may utilize unimodal or multimodal cues (e.g., visual, auditory, textual, or behavioral signals) to identify inconsistencies introduced during synthesis. These inconsistencies can manifest as spatial artifacts, temporal discontinuities, semantic mismatches, or other modality-specific anomalies. Advanced detectors also incorporate attention mechanisms or feature-level fusion to enhance cross-modal analysis.

4.3 Adversarial Relationship between Generators and Detectors

Deepfake generation and detection inherently exist in a mutually adversarial framework. Generators (G_θ) evolve to produce increasingly realistic synthetic data, while detectors (D_ϕ) adapt to better recognize such manipulations. This adversarial interplay forms a dynamic cycle of co-evolution, similar to a minimax game:

$$\min_{\theta} \max_{\phi} \mathbb{E}_{x_r \sim p_r(x)}[\log(1 - D_\phi(x_r))] + \mathbb{E}_{z \sim p_z(z)}[\log(D_\phi(G_\theta(z)))] \tag{3}$$

Progress in generative modelling directly challenges detection systems, while advances in detection motivate the development of more sophisticated synthesis techniques. Understanding this evolving relationship is critical for studying adversarial robustness, generalization, and counter-forensic strategies in deepfake research.

4.4 Adversarial and Anti forensic Attacks and Key Terminologies

In this paper, we review adversarial and anti-forensic attacks in detail, including traditional adversarial methods such as FGSM [28] and PGD [29], as well as anti-forensic techniques based on GANs [30] and diffusion models [31]. For convenience, we use the terms "adversarial" and "anti-forensic" attacks interchangeably throughout the paper; however, the term *adversarial attack* is more commonly used in existing literature. Adversarial attacks involve the addition of perturbations to the input image that cause the models or classifiers to misclassify the input image. These perturbations are imperceptible for the human eyes but can fail the model to provide accurate results. After the deep learning models have been successfully employed in computer vision tasks, Szeged et al. [32] introduced the concept of adversarial attacks in 2014. The authors demonstrated that deep neural networks misclassify the test input image with high confidence if a little perturbation is added to it. A linear amount of undetectable noise was introduced to the original image as a perturbation, which failed the model to identify the correct class for a perturbated input image. The main terminologies that are widely used in the adversarial attacks research field are provided in Table 3, while the timeline of adversarial attacks is given in Fig. 3. This timeline indicates the most frequently used adversarial attacks in multidisciplinary fields.

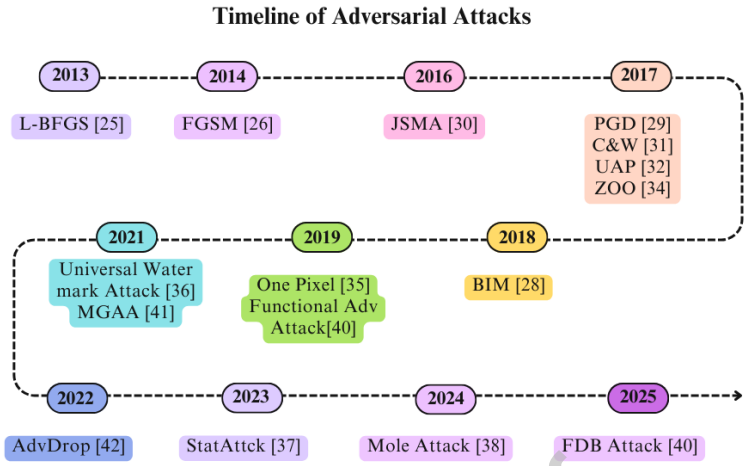

Fig. 3. Timeline of adversarial attacks between 2013 to 2025.

The categorization of adversarial attacks on visual and audio content can be based on many factors, including the intention, knowledge, and authority of the adversary. Adversarial attacks can be divided into targeted and non-targeted attacks depending on the adversary's intention; however, poisoning and evasion categorization of attacks are based on the adversary's authority. Based on the adversary's knowledge, adversarial attacks can be mainly classified as white box and black box. These attacks are briefly described in Table 3, and the taxonomy of adversarial attacks is presented in Fig. 4.

5 Adversarial attack techniques

In this section, we discussed adversarial attacks, focusing on commonly used white box and black box attack techniques. By examining these attack techniques, we gain a comprehensive understanding of the potential vulnerabilities they can exploit. The general framework of white box and black box attacks is given in Fig. 5 and Fig. 6. Whereas Table 4 and Table 5 present the categories of white and black box attack approaches.

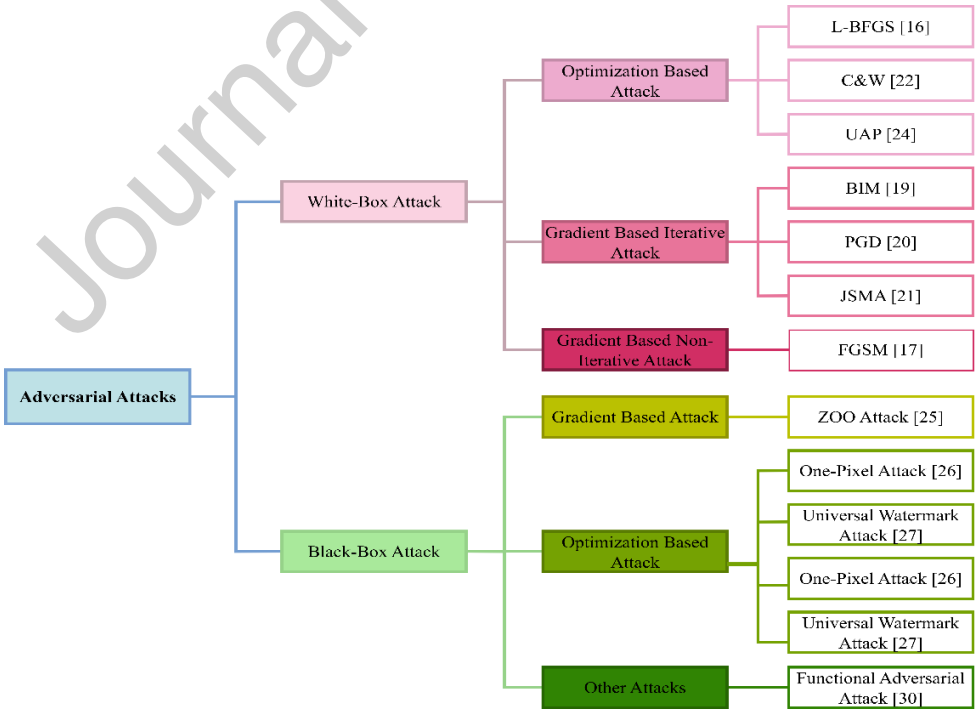

Fig. 4. Detailed categorization of adversarial attacks on deepfake detection systems.

Table 3. Comprehensive terminologies and detailed classification of adversarial attack types.

Terms	Description
Terminologies	
Victim Model	Model on which adversarial attack is performed.
Adversarial Example	Perturbated image used to fail the victim model.
Adversary	One who performs the adversarial attacks on victim models to fail them.
Adversarial Perturbation	Mechanism of modifying the original image to the perturbated image. It can be iterative or non-iterative.
Adversary Knowledge	Information known to the adversary about the victim model.
Transferability	Adversarial examples generated to fail one model can be utilized to invade another model. This characteristic of adversarial examples is referred to as transferability.
Adversarial Attacks	
Targeted Attacks	The adversary specifies a target label for the misclassified samples and tempts the model to assign the targeted label to the adversarial examples.
Non-Targeted Attacks	Adversary induces the model to misclassify perturbated images without specifying a target label with the intention to fail the victim model.
Poisoning Attacks	The adversary has the authority to inject fake/erroneous samples into the training dataset of the victim model, which can cause the model to provide incorrect predictions.
Evasion Attacks	To fail the victim model, the adversarial example is used as input since the adversary has no access to the victim's model and its training data.
Black Box Attacks	Black box attacks are performed in the scenario where the adversary has zero knowledge about the victim's model. Black box attacks are based on the transferable nature of adversarial examples. In such attacks, adversarial examples generated to fail the surrogate model are utilized for the victim model.
White Box Attacks	White box attacks are performed when the adversary has the perfect knowledge of the victim model, including the model's architecture, parameters, gradients, weights, training data, and model output. Keeping in view such information, particularly gradient knowledge, the adversary generates an adversarial example to fail the model.
Gray Box Attacks	While performing gray box attacks, the adversary has limited knowledge (such as knowledge regarding training data, which is known, but the model architecture is unknown). This attack also relies on the transferability of adversarial examples. The adversary trains the surrogate model on the known training data to mimic the victim model. Using the surrogate model information, an adversarial example is generated to fail the victim model.

5.1 White box attacks

5.1.1 L-BFGS

Szegedy et al. [32] was the first to attack the deep learning models using adversarial examples. This attack was named as limited memory broyden fletcher gold farb shanno (L-BFGS) attack. The adversarial examples $\mathbf{I}'$ were generated via introducing a minimum perturbation $\mathbf{p}$ to the input image $\mathbf{I}$ such that a small difference exists between $\mathbf{I}$ and $\mathbf{I}'$. When classified using model M, the adversarial example was misclassified by assigning a target label L'.

$$\begin{aligned} &\min \|\mathbf{I} - \mathbf{I}'\| \quad \text{s.t} \\ &M(\mathbf{I}') = L' \quad \text{where } \mathbf{I}' \in [0,1] \end{aligned} \tag{4}$$

Here, $\mathbf{I}$ and $\mathbf{I}'$ denote the original and adversarial image vectors, respectively, with pixel values normalized in the range [0, 1]. To approximate the minimum perturbation, the L-BFGS algorithm was implemented with the following loss function.

$$\min c\|\mathbf{p}\| + loss\,(\mathbf{I}', L') \quad where\ \mathbf{I}' \in [0,1] \tag{5}$$

In Eq. (5), $c > 0$ is a constant that controls the trade-off between minimizing the perturbation magnitude and the classification loss. The loss term $loss(\mathbf{I}', L')$ is the cross-entropy loss between the model's predicted probabilities for the adversarial image $\mathbf{I}'$ and the one-hot encoded target label L'.

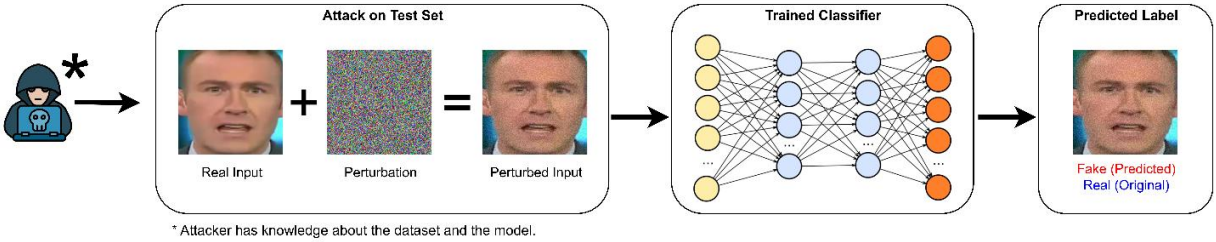

Fig. 5. General framework of white box attack involves an attacker having full knowledge of the target system's architecture, design, and implementation details.

5.1.2 FGSM

In 2015, Goodfellow et al. [28] demonstrated that neural networks were vulnerable to adversarial attacks because of their linear nature. They introduced the one-step fast gradient sign method (FGSM) to create an adversarial example with the gradient computed using backpropagation. Perturbation þ was generated via updating the gradient along the perturbation direction, i.e., a sign of gradient, at each pixel of input image **I**. For instance, the value of the gradient was increased if the gradient was positive, whereas, for a negative gradient, the value was decreased. In a non-targeted FGSM attack, the value of classification loss of the model was maximized, while in the targeted FGMS attack, the probability of target label L' was maximized for the input image **I** via minimizing the loss between the predicted class of **I** and L'. The targeted FGSM attack was introduced in **[33]** to avoid the label-leaking problem of a non-targeted FGSM attack. The formulation to generate the non-targeted and targeted FGSM attacks is shown in Eq. 5 and Eq. 6, respectively.

$$\mathbf{I}' = \mathbf{I} + \varepsilon \cdot sign\left(\nabla_{\mathbf{I}} \text{loss}(\mathbf{I}, L)\right) \tag{6}$$

$$\mathbf{I}' = \mathbf{I} - \varepsilon \cdot sign\left(\nabla_{\mathbf{I}} \text{loss}(\mathbf{I}, L')\right) \tag{7}$$

where ε represents the amount of perturbation added and L indicates the actual label of image **I**. Increasing the ε will increase the misclassification rate at the expense of the perturbated image being significantly different from the original image and vice versa.

5.1.3 Basic iterative attack (BIM)

BIM [34] is the extended version of the FGSM attack, which iteratively introduced the perturbation and clipped the pixel values of the intermediate perturbated image at each iteration. The number of iterations was equal to min (ϵ+4, 1.25ϵ). BIM updates the image slightly at each iteration and thus prevents notable changes in pixel values.

$$\mathbf{I}'_{n+1} = clip_{\mathbf{I},\epsilon}\{\mathbf{I}'_n + \propto \cdot sign\left(\nabla_{\mathbf{I}} loss(\mathbf{I}'_n, L)\right)\} \tag{8}$$

Here, $\propto$ represents the change in the value of each pixel on each iteration (in a BIM attack $\propto = 1$, meaning the pixel value is changed by 1 each iteration), the operator "." denotes element-wise multiplication between the scalar step size α and the sign of the gradient vector, indicating that each pixel's update is scaled by α times the sign of its gradient component.

5.1.4 Projected gradient descent (PGD) attack

Madry et al. [29] introduced the variant of BIM named projected gradient descent that initialized the iteration through random noise in the L_∞ ball around the original image. Instead of clipping the pixels, PGD projected the perturbation into ϵ-l_∞ neighborhood of the input image. PGD attack generates adversarial examples with the largest local max-loss value, which are most likely to fool the target model.

$$\mathbf{I}'_{n+1} = proj_{\mathbf{I},\epsilon}\{\mathbf{I}'_n + \propto .sign\left(\nabla_{\mathbf{I}} \text{loss}(\mathbf{I}'_n, L)\right)\} \tag{9}$$

5.1.5 Jacobian saliency map attack (JSMA)

In [35], a white box attack was introduced which constructed the saliency maps by computing the forward-feed derivative of the DNN framework. The forward feed derivative computed the gradient and was defined as the Jacobian matrix as given below:

$$J_F(\mathbf{I}) = \frac{\partial F(\mathbf{I})}{\partial \mathbf{I}} = \left[\frac{\partial F_n(\mathbf{I})}{\partial I_m}\right]_{m \times n} \tag{10}$$

Based on the gradient, the saliency maps output the set of important pixels to which the perturbation was introduced to generate the adversarial examples that fulfill the adversary goal, i.e., misclassification. Through this attack, only 4% of the total pixels are perturbated, which might cause the perturbation to be visible to human eyes.

5.1.6 C&W attack

Carlini et al. [36] introduced a targeted attack following the optimization problem similar to LBFGS and put forward a new formulation by replacing a loss function with the objective function. In [36], seven different objected functions were presented to be used for scaling the minimization function. However, unlike LBFGS, it was an unconstrained optimization problem, as the author introduced ω which satisfied $\text{þ} = \frac{1}{2}(\tanh(\omega) + 1) - \mathbf{I}$ and controlled the perturbation to the input image. The optimal formula for a C&W attack is shown in Eq. 10.

Table 4. Detailed overview of white box adversarial attacks.

White Box Attacks	Attacks	Distance	Target / Non-Target	Advantages	Limitations
Optimization-Based Iterative Attacks	L-BFGS	L_2	Target	High stability and effectiveness.	Computationally complex.
	C&W	L_0, L_2, L_∞	Target	The attack is stronger than FGSM, PGD, and BIM. Generate highly transferrable adversarial examples.	Computationally complex.
	UAP	L_2	Non-Target	High fooling rate at smaller norm. Generalize well across different models.	---
Gradient-based Iterative Attacks	BIM	---	Target, Non-Target	The image is not distorted at a high value of ϵ. Computationally efficient than L-BFGS. High success rate than FGSM.	Adversarial examples fall into poor local maxima problems and overfit the model. Takes more time compared to FGSM.
	PGD	---	Target, Non-Target	High success rate than FGSM.	Takes more time compared to FGSM.
	JSMA	L_0	Target	High success rate and transferability.	Computationally complex. Applicable only to feedforward DNNs.
Gradient-based non-Iterative Attacks	FGSM	L_2, L_∞	Target, Non-Target	Computationally efficient.	Lower success rate. A distorted image is produced at a large value of ϵ.

$$\min \|\text{þ}\| + c.f\,(\mathbf{I}') \quad where\ \mathbf{I}' \in [0,1] \tag{11}$$

where $f\,(\mathbf{I}') \geq 0$ if and only if the classifier outputs the targeted label. The optimal value of c is found via binary search. Thus, the perturbated image that has a high score for target label L' can be found by minimizing the f.

5.1.7 Universal adversarial perturbation (UAP)

In [37], a strong universal attack was introduced that generalizes well across different models. A single universal perturbation vector $\boldsymbol{v}$ was identified by UAP that can be added to any input sample to fool the classifier up to the specified fooling rate γ. To calculate $\boldsymbol{v}$, the author utilized the DeepFool attack [38] such that $\boldsymbol{v}$ satisfies the following constraints:

$$\|\boldsymbol{v}\| \leq \in \tag{12}$$

$$P\left(g(\mathbf{I}+\boldsymbol{v}) \neq g(\mathbf{I})\right) \geq 1-\gamma \tag{13}$$

5.2 Black box attacks

5.2.1 Zeroth order optimization (ZOO) attack

Inspired by the white box C&W attack formulation, Chen et al. [39] presented a black box attack named the ZOO attack for which no gradient information of the victim model was required. To attack the black box classifier, the ZOO attack monitored the changes in prediction confidence and utilized the zeroth order oracle along with important sampling hierarchical attack, and dimension reduction. Symmetric differential quotient and the Hessian estimate were used to estimate the gradient as follows:

$$\frac{\partial F\,(\mathbf{I})}{\partial \mathbf{I}_j} = \frac{F(\mathbf{I}+h\boldsymbol{e}_j) - F\,(\mathbf{I}-h\mathbf{e}_j)}{2h} \tag{14}$$

$$\frac{\partial^2 F\,(\mathbf{I})}{\partial \mathbf{I}_j^2} = \frac{F(\mathbf{I}+h\boldsymbol{e}_j) - 2F(\mathbf{I}) + F\,(\mathbf{I}-h\mathbf{e}_j)}{h^2} \tag{15}$$

where e_i is a basic vector with i^{th} component as 1, and h is a small constant step size used for finite-difference gradient estimation. By following [34], h is set to 0.0001.

5.2.2 One-pixel attack

In [40], a one-pixel attack was introduced that perturbed only one pixel in the image to fool the classifiers. A one-pixel attack perturbed the pixel in the direction on the axis of one of the n-dimensions of an input image. A differential evolution algorithm was used to solve the optimization problem and find a higher-quality solution compared to the gradient-based solution. The formulation of pixel attack is as follows:

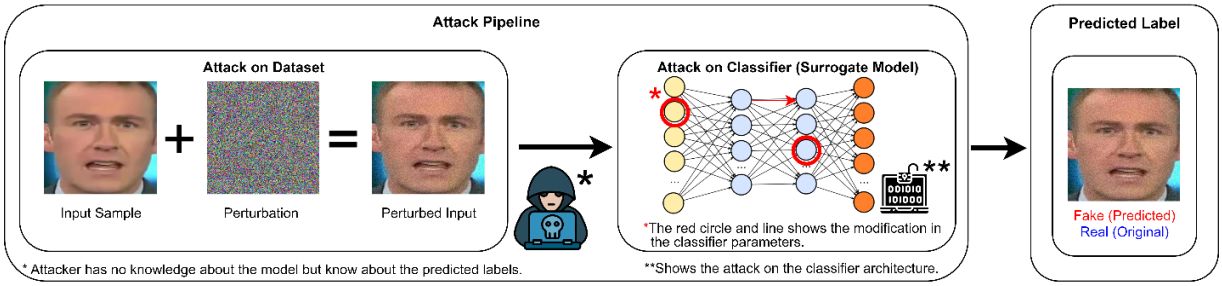

Fig. 6. General framework of black box attack involves manipulating a model without knowledge of its internal workings.

$$\text{maximize } f\ (\mathbf{I} + \text{Þ})\ \ subject\ to\ \ \|\text{Þ}\|_0\ \leq \mathbf{d} \qquad (16)$$

where **d** represents the data point to be modified, and the value of **d** is set to 1 for the one-pixel attack.

5.2.3 Universal watermark perturbation

Wu et al. [41] introduced a universal watermark attack that combined the techniques of optimizing loss function and watermarking to generate adversarial examples. This attack modified the pixel values within the watermark and applied the watermark to any shape at any location of the image. The authors utilized SGD and ADAM optimization algorithms to optimize the loss function. The formulation for universal watermark perturbation is shown in Eq. 17.

$$maximize\ \frac{\sum_j^n f(\mathbf{I}_j) - \sum_j^n f(\mathbf{I}_j + \mathbf{W})}{\sum_j^n f(\mathbf{I}_j)}\ \ where\ \mathbf{W} = \mathbf{W}(R, G, B, S, L) \qquad (17)$$

Here, $W(R, G, B, S, L)$ represents the watermark perturbation function parameterized by its color channels (R, G, B), geometric shape (S), and spatial location (L) on the image. This function generates the universal watermark pattern applied to input images to create adversarial examples.

5.2.4 Statistical consistency attack

Hou et al. [42] introduced a statistical consistency attack (StatAttack) that minimized the statistical difference between real and deepfake images. More precisely, statistically sensitive natural degradations such as exposure (P_e), blur(P_b), and noise (P_n) were introduced to the fake images and then the distribution aware loss was utilized to optimize different degradations. StatAttack generated adversarial examples in which feature distribution was closer to the real images and thus evaded the spatial and frequency-based deepfakes detectors. The extended version, namely MStatAttack, was also presented in [42], which involved sequential multi-layer degradations and tuned the combination weights via utilizing loss. These attacks produce high visual quality adversarial images and show a higher success rate than the existing PGD, FGSM, and MIFGSM attacks.

$$P_{\boldsymbol{\theta}}(\boldsymbol{X}^{\boldsymbol{fake}}) = P_n(P_b(P_e(\boldsymbol{X}^{fake}))) \qquad (18)$$

5.2.5 Facial Mole attack

Despite advancements in deepfake detection methods, recent studies have shown that these systems are still susceptible to adversarial black-box attacks. For instance, a study demonstrated [43] a novel facial mole black-box adversarial attack that successfully disrupted shared attention patterns in detection models. To add black moles to each masked frame Mf_{xy}, the first step is to calculate the number of moles n that will be added. The parameter m_i indicates the height and width of the mole, which can be either 1-pixel or 2-pixels in size, represented as $m1$ and $m2$, respectively. This attack was able to significantly reduce detection accuracy and achieve a success rate on state-of-the-art detectors.

5.2.6 Facial Distraction Black-Box Attack

Facial Distraction Black-Box Attack (FDB attack) [44] framework that is visually realistic, resilient, and demonstrates formidable attacking capabilities. The proposed black-box attack is capable of successfully evading deepfake detectors without the need for access to the target detector's parameters or architectural specifications. It exhibits high transferability across a variety of deepfake detectors, including end-to-end deep learning, fused, and unified models.

5.2.7 Functional adversarial attack

To produce natural adversarial examples, Laidlaw et al. [45] presented a functional threat model that applied a single perturbation to each input image pixel. For example, changing all the green color pixels to light green color. The

functional threat model utilized various regularization functions such as difference (F_{dif}) and smooth (F_{smo}) to make the modifications imperceptible. The F_{dif} limits the amount of perturbation in the image, while the F_{smo} perturbed similar features in the same direction.

5.3 Other attacks

Inspired by meta-learning, a plug-and-play meta-gradient adversarial attack (MGAA) was introduced in [46] that can be integrated with any other gradient-based attack. To generate adversarial examples, MGAA iteratively simulated the white box and black box attacks from the model. This attack [46] improves the transferability of adversarial examples by narrowing the gap of gradient directions between the white box and black box setting, thus enhancing the success rate of both black box and white box attacks. Duan et al. [47] introduced an AdvDrop attack that dropped the imperceptible features (i.e., subtle texture information) from the input image to generate an adversarial example. AdvDrop attack first transformed the input image into a frequency domain, and then some frequency components were reduced quantitatively. Cilloni et al. [48] introduced a gradient-based attack, namely a focused adversarial attack that introduced the perturbation in the sensitive region of the image. For this, a focused threshold was used to identify whether the output feature map should be considered or not for computing the perturbation. Intermediate layer attack with attention guidance (IAA) was presented in [49], which enhances the black box attack transferability while maintaining the performance of the attack in the white box setting. IAA introduced the perturbation to a specific intermediate feature space via attention loss and stabilized the optimization direction through projection loss. IAA attack degrades key features of the input image that are not model-specific and thus makes the adversarial example effective for different models.

Table 5. Detailed overview black box adversarial attacks.

Black Box Attacks	Attacks	Distance	Target / Non-Target	Advantages	Limitations
Gradient-based iterative	ZOO Attack	L_2	Target, Non-target	Achieves comparable performance to the C&W attack. Avoid performance loss while transferability.	Computationally complex for large models. Requires expensive computation to query and estimate gradients.
Optimization-based iterative	One-Pixel Attack	L_0	Target, Non-target	Only the label information is needed to perform the attack. It can affect multiple types of frameworks.	----
	Universal Watermark Perturbation	L_∞	Non-target	Powerful attack compared to UAP and FGSM	Perform well for the networks with shallow layers.
---	Functional adversarial Attack	L_2, L_∞	---	Perturbation is imperceptible.	Features cannot be perturbated individually, making the attack more restrictive.

6 Adversarial attacks on visual deepfake detectors

This section provides an in-depth review of existing papers on evading image and video deepfake detectors using diverse adversarial attacks. The overview of such existing papers is provided in Table 6 and the taxonomy of adversarial attacks on visual deepfakes detectors is shown in Fig. 7.

Table 6. Comprehensive overview of existing literature on adversarial attacks on deepfakes detectors.

Year	Attacks	Datasets	Victim Models	Results			Perceptual Similarity Measures	Limitations
				Before Attacks	After Attack			
					White Box Attack	Black Box Attack		
Adversarial Attacks on Image Deepfakes Detectors								
GAN-based Adversarial Attacks								
2021	GAN-based attack [30]	StyleGAN generated fake images	EfficientNet-b3	Acc = 97	Acc = 0	Acc = 9	MSE, PSNR, SSIM, learned perceptual image patch similarity (LPIPS)	The transferability of the adversarial examples towards the other forensics detectors is limited.
			XceptionNet	Acc = 93	Acc = 0	Acc = 5		

2021	GAN-based attack [50]	100K-Faces, TPDN (StyleGAN)	CCNDetector, ResNet18, VGG16, VGG19, XceptionNet	CNNDetector Acc = 98	CNNDetector Acc = 34.8	ResNet50 Acc = 37.75 XceptionNet Acc = 51.25 VGG19 Acc = 65.25 VGG16 Acc = 38.25	---	Low performance on XceptionNet and ResNet18 compared to FGSM.
2021	GAN-based attack [51]	GAN-generated images (StyleGAN, StarGAN)	Xception	Acc = 99.67	ASR = 100	ASR = 91.07	PSNR	---
			ResNet-50	Acc = 78.26	ASR = 81.23	ASR = 24.35		
			DenseNet	Acc = 96.39	ASR = 97.08	ASR = 87.30		
2021	Anti-forensic GAN attack [52]	GAN generated images	MISLNet, SRNet, DenseNEt, VGG-19	Avg. Acc = 98	Avg. ASR = 96		PSNR, SSIM	---
2023	Trace removal attack [53]	All-in-one dataset (CelebA, ProGAN, STGAN, DeepfakeTIMIT)	Xception	Acc = 99.9	Acc = 18.9		PSNR, SSIM	Less effective for spatial-based (Xception, Path-CNN) and frequency-based (DCTA, F3Net)
			Patch-CNN	Acc = 92.8	Acc = 13.06			
			DCTA	Acc = 93	Acc = 30.16			
			F3Net	Acc = 99.9	Acc = 56.10			
			LF	Acc = 93	Acc = 14.10			
			NF	Acc = 74.9	Acc = 21.74			
2025	GAN-based attack [54]	AI-generated images	DenseNet121 Inception-V3 MobileNetV3 ResNet101 Xception	Avg. Acc = 93.8	Avg. Acc = 27.7		PSNR SSIM LPIPS	---
Reconstruction-based Adversarial Attacks								
2020	FakePolisher [55]	Used 16 GAN-based method to generate fake images.	GANFingerprint	Acc = 99.74	PCA Reconstruction Acc = 36.65 KSVD Reconstruction Acc = 54.36		cosine similarity (COSS), PSNR and SSIM	Not removing all the fake artifacts from the fake images
			DCTA	Acc = 99.6	PCA Reconstruction Acc = 53.43 KSVD Reconstruction Acc = 64.77			
			CNNDetector	Acc = 68.4	PCA Reconstruction Acc = 14.77 KSVD Reconstruction Acc = 48.44			
2021	DeepNotch [9]	Used 16 GAN-based method to generate fake images.	GANFingerprint,	Acc = 99.74	Acc = 14.52		COSS, PSNR, SSIM	---
			DCTA	Acc = 99.6	Acc = 30.96			
			CNNDetector	Acc = 68.4	Acc = 11.2			
2021	GAN based attack [56]	DFDC, Celeb-DF, FF++	XceptionNet	Acc = 98.95 (Celeb-DF)	Acc = 7.33		---	---
			FAW	Acc = 84.39 (FF++)	Acc = 23.04			
			DenseNet	Acc = 96.72 (FF++)	Acc = 7.97			
			ResNet	Acc = 96.55 (FF++)	Acc = 5.57			
Other Adversarial Attacks								
2020	Flip the lowermost bit of pixels. [57]	Private dataset Fake images were generated using 11 methods. The dataset not only consists of faces	Existing Deepfakes detection methods [4,5]	AUC reduces from 99% to almost zero in different scenarios			L_0, L_2 Norm	Highly restrictive attack.
2020	Universal adversarial attack, individual	FF++ (Deepfakes, Faceswap, Face2Face)	MesoInception-4	Acc = 94.88	---	UAP Acc = 19.19	RMSE	Perceptual loss in adversarial example
			ForensicTransfer	Acc = 86.39	---	UAP Acc = 24.06		

	adversarial attack [58]		Y-shaped Network	Acc = 91.98	UAP Acc = 3.17			generated using IAA.
2021	Disrupting attack [59]	WIDER, 300-W, UMD faces, Celeb-DF	Faster RCNN, Fv16(VGG16), Fr101(ResNet101), Pr50(ResNet50), Sv16(VGG16)	---	SSIM = 90 (approx.)		Data utility quality (DUQ), SSIM	Time consuming attack.
2021	Poisson noise Deep Fool (PNDF) [60]	11 dataset including FF++	ResNet 50 and other GANS	Acc with cycle GAN = 0.972	Acc = Almost 0		---	Computationally complex due to iterative nature.
2021	Universal adversarial attack [61]	DFDC	EfficientNetB7	AUC = 0.717	ASR = 100		mean distortion $L\infty$	Perturbation is perceptual on higher magnitude.
			EfficientNet Seli B7	AUC = 0.724	ASR = 100			
			EfficientNetB3	AUC = 0.724	ASR = 100			
			XceptionNet	AUC = 0.7	ASR = 100			
2021	Key Region Attack [62]	FF++	Xception	Acc = 99	Acc = 0.006		---	KRA combined with other attacks (PGD) is better instead of individual.
			Resnet-50		Acc = 0.001			
			Resnet-101		Acc = 0.001			
			Inception-v3		Acc = 0.40			
2021	Noise attack [63]	DFGC-21 testing dataset, FaceForensic++, Deepfake Detection Challenge (DFDC), Deeper Forensics Challenge	Own teacher, student models.	---	**AUROC** DFDC test = 0.682, FF++-test = 0.732		---	Less diverse dataset (DFGC-21), limited to generated samples.
2021	Label flipping attack, backdoor attack [64]	FF++	Xception	Combine Acc = 96	Acc = 37.5		---	Detectors were limited to detect in adversary.
2022	Double-masked guided attack [65]	StyleGAN generated fake faces	ResNet	Acc = 100	SR = 99.9	DenseNet121 SR = 99.2	SSIM, PSNR, LPIPS	The transferability of attack needs to be improved.
						MesoIncepion4 SR = 94.09		
			Xception	Acc = 99.99	SR = 100	AlexNet SR = 21.15		
						Discriminator SR = 99.04		
			EfficientNet-b0	Acc = 99.93	SR = 99.9	GramNet SR = 96.84		
						RFM SR = 89.10		
			ResNet-18	Acc = 93.2	FGSM Acc = 7.5 C&W Acc = 0	FGSM Acc = 20.8 C&W Acc = 4.6		
2022	Two-phase attack, SA-GD, SA-EA [66]	---	ResNet-50	Acc = 99.6	Two phase attacks SR = 90 SA-GD white box SR = close to 100	SA-EA Black box Below 20%	---	To increase the success rate, a generation of large perturbation is required which affects the visual quality of the perturbated image
2022	Frequency adversarial attack [67]	DFDC, FF++	EfficientNet-b4 and others	FF++ Acc = 94.3 DFDC	FF++ SR = 83.2 DFDC	ResNet FF++ SR = 22.7 DFDC SR = 20.1	MSE, peak signal-to-noise ratio (PSNR) and	Transferability is not good.

				Acc = 91.1	SR = 97.1	XceptionNet FF++ SR = 1.4 DFDC SR = 2.7	structural similarity (SSIM)	Less effective for spatial-based models.
2022	Apply makeup artifacts to identified landmarks [68]	FF++ F2F subset	MesoInception-4,	Acc = 86	Acc = 52		---	Not effective as other attacks i.e., PGD, FGSM etc.
			TwoStreamNet	Acc = 99.62	Acc = 55			
2023	StatAttack, MStatAttack [42]	StyleGAN, StarGAN, ProGAN, DeepFakes subset of FF++ dataset	ResNet50 EfficientNet DenseNet MobileNet	---	Success rate ranges from 26.5 – 88.3	SR ranges from 96.5 – 100	BRISQUE	Poor performance in case of black box attack on frequency based deepfakes detectors
			DCTA DFTD		Success rate ranges from 85 – 96.4	Success rate ranges from 12 – 38.8		
2024	AdvShadow [69]	DFDC, FF++, Celeb-DF	spatial-based, frequency-based, and physiological-based Deepfake detectors	FF++ Acc = 76.81 DFDC Acc = 65.09 Celeb-DF Acc = 73.74	SSIM = 0.9584 PSNR = 31.5303		SSIM, PSNR	Computationally complex due to iterative nature.
Adversarial Attacks on Video Deepfakes Detectors								
Other Adversarial Attacks								
2021	Robust white box and black box attack [70]	FF++, DFDC	XceptionNet	Acc = 96.04	Acc = 1.77 SR = 98.23	Acc = 16.06 SR = 83.94	L∞ distortion metrics	---
			MesoNet	Acc = 84	Acc = 0.5 SR = 99.5	Acc = 21.67 SR = 78.33		
			3D EfficientNet	Acc = 91.74	Acc = 0 SR = 100	Acc= 48.98 SR = 51.02		
2022	FGSM, C&W [71]	FF++	Conv-LSTM	Acc = 81.3	FGSM Acc = 14.8 SR = 72.31 C&W Acc = 8.3 SR = 99.72	FGSM Acc = 44.7 SR = 21.73 C&W Acc = 38.4 SR = 63.82	L∞ distortion	---
			FacenetLSTM	Acc = 84.5	FGSM Acc = 20.9 SR = 68.23 C&W Acc = 13.5 SR = 98.83	FGSM Acc = 53.5 SR = 33.89 C&W Acc = 28.7 SR = 67.14		
2022	Universal adversarial perturbation [72]	FF++, DFDC	XceptionNet, MesoNet, 3D CNN	---	3D CNN Acc = 91.74	Robust and Transferable attack Acc = 0	mean distortion	---
2023	Adversarial Deepfakes Video Generation Framework [73]	FF++, CelebDF	ResNet, XceptionNet, MesoNet	Acc = 99.4	Acc = 1.3 SR = 98.7	Acc ranges from 28.4 to 47.3. SR ranges from 52.1 to 71.6	PSNR, SSIM	---
2024	eXplainable AI [74]	FF++, CelebDF	XceptionNet, MesoNet	Acc = 80.11	Acc = 21.94	Average time taken MesoNet=4.94 XceptionNet, 2.52	Average time taken	---
2025	GAN-based attack [54]	AI-generated images	DenseNet121 Inception-V3 MobileNetV3 ResNet101 Xception	Avg. Acc = 93.8	Avg. Acc = 27.7	PSNR SSIM LPIPS	---	----

*(Acc=Accuracy, ASR=Attack Success Rate)

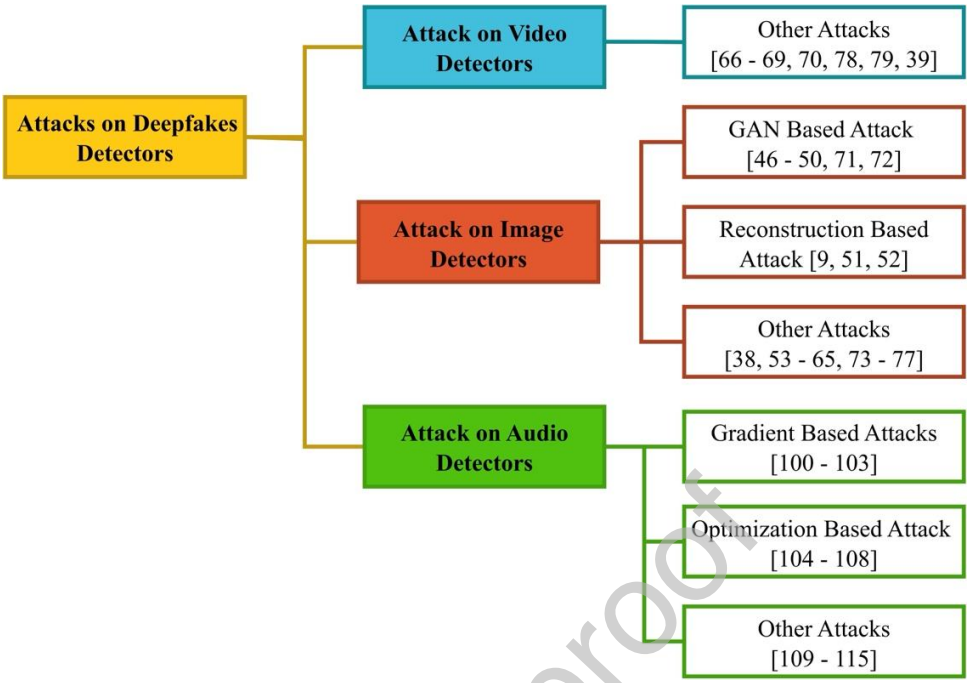

Fig. 7. Taxonomy of attacks on audio and visual deepfakes detectors.

6.1 Adversarial attacks on image deepfakes detectors

Adversarial attacks on image deepfake detectors involved attacks on images and video frames. These attacks are categorized as GAN-based attacks, reconstruction attacks, and other attacks. The details and literature related to these categories are coming in this section.

6.1.1 GAN-based adversarial attacks

GAN-based adversarial attacks mostly employ the GAN frameworks to generate adversarial examples to evade the different deepfakes detectors that detect GAN-generated fake images. In [30], the author introduced a GAN-based approach in which gradient descent was performed to the input latent vector and modified the manifolds of StyleGAN to generate the adversarial examples. The attack was performed on two deepfakes detectors (XceptionNet and EfficientNet-b3) in the black box and white box settings while reducing the detection accuracies to less than 1%. This attack method [30, 50] generates high-quality adversarial examples compared to FGSM and PGD attacks. However, the transferability of the adversarial examples to other forensics detectors is limited. Wang et al. [50] generated the content preserving adversarial examples from the fake images using the two GAN-based frameworks for evading the CNNDetector (ResNet50) under black box and white box scenarios. In the black box scenario, real data was provided to the discriminator of the GAN model along with the introduction of a Real Extractor as an auxiliary network for extracting real data features to increase the adversarial ability of the framework. For the white box scenario, the victim classifier was included in the GAN architecture and generated images without sampling to fool the classifier. For the seen data, a 63.2% and 32.55% drop in accuracy occurred; however, for unseen data, an accuracy drop of 61% and 58.55% were observed for white box and black box attacks, respectively. The method [50] also showed good transferability when an attack was performed on four other classifiers, including ResNet18, XceptionNet, VGG16, and VGG19. Liu et al. [53] introduced the trace removal attack based on an adversarial learning method encompassing multiple discriminators and one generator. By thoroughly investigating the deepfakes creation pipeline, the trace removal net empirically discovered and removed the deepfakes traces, such as spatial anomalies, spectral disparities, and noise fingerprints from the fake images. This attack was performed on spatial-based, frequency-based, and fingerprint-based deepfake detectors and significantly reduced the detection accuracies. Trace removal attacks are less effective than PGD attacks when performed on spatial and frequency-based detectors; however, the attack is more effective for fingerprint-based forensics detectors.

Uddin et al. [54] introduced transferable anti-forensic attacks on deepfake face forgery detectors targeting both real and synthetic samples, and proposed a robust detection framework based on multi-model knowledge distillation. Zaho et al. [51] introduced an adversarial attack that complicates GAN-generated image detection by forensic classifiers. The attack, using an anti-forensic generator, introduces traces resembling actual images, deceiving detectors. The

attack is transferable and highlights vulnerabilities in current synthetic image detection methods. Zhao et al. [52] developed a method to deceive forensic CNNs by creating a generator that removes forensic traces from manipulated images, making them appear unaltered. This attack is effective against multiple CNNs, transferable, and maintains high image quality by minimizing visual distortions. However, its effectiveness depends on its transferability and the precise manipulation operations on the images. Ding et al. [75] developed a novel GAN model called ExS-GAN that employs an extra supervision system to generate high-quality, manipulated images for anti-forensics applications. While maintaining image quality, the model can launch attacks; however, its long-term efficacy may be restricted by the potential for sophisticated forensic tools. Peng et al. [76] developed a novel approach that employs GAN to convert computer-generated and genuine facial images, thereby improving anti-forensic applications. The model produces convincing facial images from computer-generated (CG) inputs, which complicates the differentiation between CG and natural faces. However, the study limits its dependence on the quality of training data.

Among the reviewed GAN-based attacks, [30] is more effective in fooling the spatial-based deepfakes detectors and has greater transferability compared to the other two attacks [30, 53]. Furthermore, a comparison of the SSIM value of the attacks indicates that the adversarial examples crafted via the GAN-based attack [30], which are better in quality and look more similar to the original fake images compared to the trace removal attack [53]. So, it can be inferred that GAN-based attacks generate high visual quality adversarial examples which performed quite well in white box settings.

6.1.2 Reconstruction-based adversarial attacks

Reconstruction-based adversarial attacks involve the post-processing methods used to further enhance the visual quality of the fake images to recreate the high visual quality forged images that are then used to fool the forensics classifiers. FakePolisher introduced in [55], was a post-processing method that reduced the fake artifacts introduced in the GAN-generated fake images. FakePolisher reconstructed fake images using dictionary learning methods such as K-SVD and principal component analysis (PCA). It was used as a black box attack to evade deepfake detectors, including DCTA, GAN Fingerprint, and CNNDetector, and significantly reduced the detection accuracy. K-SVD reformed images are more like the actual fake images; however, PCA reconstruction is more effective than K-SVD reconstruction while fooling the deepfakes detector. Similarly, in [9], a DeepNotch method was introduced that performed notch filtering in the spatial domain by intelligently adding the noise using the semantic information of the image. Later, image filtering was performed to reproduce noise-free fake images. The fake images were generated using 16 Deepfakes methods, and the attack was performed using three existing detection methods: GAN Fingerprint, CNNDetector, and DCTA. The detection accuracy of the victim models was improved when re-trained on the reconstructed images, indicating that the attack was not too effective in evading the deepfakes classifier. In [55], the PCA reconstruction attack appears to be more effective for all three victim models as compared to the K-SVD attack. However, the DeepNotch attack [9] significantly reduced the detection accuracies of the classifiers when compared with the PCA reconstruction attack. Ding et al. [56] proposed a GAN-based adversarial attack and introduced a novel loss function to make the GANs an adversarial tool. Instead of adding perturbation or visual distortion, this method generated more natural and realistic deepfakes, which could easily cause the detector to fail. This attack degrades the accuracy of models up to 4.30% from 97.24%. This work can be extended to add perturbation to real images or add visual artifacts on deepfake videos to bypass the deepfake detector.

6.1.3 Other adversarial attacks

To fail the forensic classifiers that detect the GAN-generated fake faces, Zhang et al. [65] presented a double-masked guided attack that introduced the perturbation to the important face regions on which most deepfakes detectors concentrated when detecting fake images. The attack was performed on nine forensics classifiers in a white box and black box setting and achieved a good success rate. However, the transferability and robustness of the adversarial examples need to be improved. Similarly, Carlini et al. [57] introduced the perturbation to the GAN-generated images (StyleGAN, ProGAN) by flipping the lowermost bit of pixels to fail the existing deepfake detection methods [55, 77]. The rate of misclassification increases as the number of perturbated pixels increases. For method [55], the misclassification rate was 89.7% by perturbing 4% of the pixels, while for method [77], the addition of perturbation to 50% of the pixels caused 100% misclassification. Being a restrictive attack, it is difficult to execute such an attack in the real world. In [66], semantic perturbations were added that alter the target semantic attributes (pose, expression, age) while retaining the original identity to fail the forensics system (based on ResNet50). For the white box attack, a two-phase attack and the semantic aligned gradient descent (SA-GD) approach were introduced, while for a black box attack, a semantic aligned evolutionary algorithm (SA-EA) was utilized to generate adversarial examples. The two-phase attack achieved the highest success rate, reaching 90%, whereas the SA-EA attack attained the lowest success rate (below 20%). However, increasing perturbation affects the visual quality; for example, a two-phase attack

introduces noise in the image while the SA-GD and SA-EA attacks alter the non-targeted attributes. To fail several face detectors and facial landmark extractors, Li et al. [59] introduced imperceptible adversarial perturbation for facial images. In the white box attack scenario, training instances were disrupted. However, zero-mean Gaussian noise was introduced for the gray box and black box. On the landmark extractor and face detector, the highest SSIM was achieved at 90% and 98%. This technique [59] is more time-consuming for a large number of images. Fan et al. [60] introduced a Poison Noise DeepFool (PNDF) adversarial attack to degrade the performance of deepfake detectors, including ResNet50 and two other customized classifiers. Images from ten different deepfake generators were collected to be perturbed iteratively in the white box manner. The AUC and accuracy of detectors reduce to 0.3331 and 0.071 with the PNDF attack on ProGAN-generated images. PNDF attack is quite effective in degrading the performance but is computationally complex. Liu et al. [69] developed AdvShadow, a transferable adversarial attack designed to exploit natural shadows in real-world scenarios to target Deepfake detectors. AdvShadow uses a randomized shadow generator, shadow overlay network, and adversarial shadow-generating method to minimize differences in brightness between actual and generated images. However, its evaluation focuses on DeepFake detectors, and its application to other detection architectures has not been explored.

In [58], DNN-based forensics classifiers were evaluated against adversarial attacks and demonstrated that individual and universal adversarial perturbations could cause the deepfakes classifiers to misclassify the input image. To craft the individual adversarial attack (IAA), a gradient-based iterative method was used, which considered the length of the gradient along with its direction to add perturbation to the image. However, for universal adversarial attack (UAA), a new objective function was introduced that utilized the resource-conserving over-firing approach to craft the universal perturbation in a data-free manner. Both attacks (IAA, UAA) drastically reduced the classification accuracy of the Y-shaped network [78]. The transferability of UAA was also demonstrated on MesoNet [79] and Forensic Transfer [80]. The UAAs are more imperceptible than IAAs because condensed local distortion causes serious perceptual loss in case of individual adversarial perturbations. In [67], frequency adversarial attack (FAA) was proposed to evade both spatial-based and frequency-based deepfakes detectors. The average success rate of the attack for spatial models using the FaceForensic++ (FF++) dataset was 73%, whereas for frequency-based models, it was 90.9%. Similarly, for spatial and frequency-based models, the average success rates of 87.5% and 99.3% were achieved, respectively, using the DFDC dataset. Although perturbation is imperceptible to humans, FAA is less effective for spatial-based models compared to frequency-based victim models. Moreover, the transferability of the attack is not good enough, as the success rate lies between 1.5% and 49% in black-box settings. However, FAA attacks outperform the PGD and FGSM attacks. Hou et al. [42] introduced a statistical consistency attack (StatAttack) that minimized the statistical difference between real and deepfake images. More precisely, statistically sensitive natural degradations such as exposure, blur, and noise were introduced to the fake images and then the distribution aware loss was utilized to optimize different degradations. StatAttack generated adversarial examples in which feature distribution was closer to the real images and thus evaded the spatial and frequency-based deepfakes detectors. The extended version, namely MStatAttack, was also presented in [42], which involved sequential multi-layer degradations and tuned the combination weights via utilizing loss. These attacks produce high visual quality adversarial images and show a higher success rate than the existing PGD, FGSM, and MIFGSM attacks. Lim et al. [68] introduced a black box attack in which makeup artifacts (eyeliner, blush, lipstick) were applied to the regions identified through facial landmarks to produce the perturbated images. The accuracies of the victim models (MesoInception-4 and TwoStreamNet) decreased to 50% on the perturbated images, which indicates that the attack is not effective compared to other attacks (i.e., PGD, FGSM, and C&W). However, the result indicates that such subtle changes in the input image can easily fool the deepfakes classifier.

Neekhara et al. [61] introduced an imperceptible universal adversarial attack, which was performed on EfficientNet (EN) B7, B3, and XceptionNet (XN) victim models. A perturbation was introduced to each frame of 100 videos collected from the DFDC dataset. An attack was performed on the models in white and black box settings using gradient-based saliency maps. In comparison with the white box and transfer attacks, universal attacks in the black box setting attained the highest attack success rate on all victim models. The attack success rate of the universal attack was 100.0% on EN-B7 Selim, 77.5% on XN, and 66.5% on EN-B3, with an $L\infty$ threshold. This attack perturbed video frames imperceptibly; however, perturbation is visible at a higher magnitude. Liao et al. [62] introduced a Key Region Attack (KRA) that added imperceptible perturbation in key regions of the image to disrupt deepfakes detectors. A KRA attack was also performed with a PGD attack, and the highest ASR achieved 0.99 with a KAR-PGD attack on ResNet-50 and Xception in the white box setting. This technique performed well as a combined attack, so it should be improved to perform well individually. Peng et al. [63] designed a two-party game between deepfake generators and detectors. Three different creation tracks were introduced based on face shifters and faceswap generators. In the first track, the FGSM attack was used to add adversarial noise to fake images generated through the face shifter

method. In the second track, "teacher and student", face shifter models were used and L_2 loss was introduced in the student model to resemble the outcomes with its teacher model. In the third track, more realistic face-swapped images were created to fail deepfake detectors. However, these techniques were employed only on generated images, not on real ones. Ivanovska et al. [31] proposed Denoising Diffusion Models (DDMs) to generate realistic images for black-box attacks on deepfake detection systems being explored. A guided conditional DDM is used to reconstruct FF++ deepfakes with a predetermined number of diffusion steps. However, this approach lacks generalizability to attacks produced with different denoising steps. Cao et al. [64] highlighted the vulnerability of deepfake detectors in backdoor adversarial attack settings. In the label flipping attack, the training data was poisoned by flipping the labels of some training data from real to fake, while in the backdoor attack, the small triggers were embedded into images to perturb it. On label flipping, the accuracy of the classifier only dropped to 0.07, whereas triggers were added to 5% of training data that failed testing in the deepfake detector.

6.2 Adversarial attacks on video deepfakes detectors

Adversarial attacks against the video deepfake detectors entail the creation of an adversarial video that is used to fool the video deepfake classifiers. For this purpose, perturbation is introduced to the frames of the fake videos to generate an adversarial video. In [70], adversarial videos generated via adding the perturbation to the frames of the video (FF++ dataset) were used to fail two CNN models including MesoNet and XceptionNet, and one sequence-based model named 3D EfficientNet. The authors conducted robust white box and black box attacks on the models and reported success rates on compressed and uncompressed videos. The success rate of these attacks was good on uncompressed videos, while performance degrades on compressed videos. Likewise, Shahriyar et al. [71] showed the effectiveness of FGSM and C&W attacks on the sequenced-based deepfakes detector in a white box and black box setting. For this purpose, the victim models were Conv-LSTM [81] and FacenetLSTM [82] which attained an accuracy of 81.3% and 84.5% on unperturbed images from the FF++ dataset. Similar to [70], the white box attacks crafted for one model were used as black box attacks for the other model and vice versa. Adversarial attacks reduced the accuracy of models to 8%-20.9% in the white box setting, while in the scenario of a black box attack, accuracy decreased to 28.7%-53%.

Moreover, C&W attacks outperform the FGSM attack while evading sequence-based forensic classifiers. Gowrisankar et al.[74] reveals that traditional explainable Artificial Intelligence (XAI) evaluation methods are insufficient for deepfake detection models due to their unique functionality. The evaluation focuses on deepfake detection models without considering their application to other classifiers. The proposed XAI approach lacks consideration for computational efficiency and resource requirements. In [73], a novel framework was introduced that generated coherent adversarial videos to evade the video deepfake detectors including ResNet, XceptionNet, and MesoNet. To maintain consistency among adjacent frames, this framework [73] utilized optical flow to restrict perturbation generation and then introduced adaptive distortion cost for visual quality improvement and imperceptibility via constraining total perturbations. The attack was performed in the white box and black box scenarios utilizing F++ and CelebDF datasets. For the white box setting, the success rate achieved ranged from 97.5% to 99.4%, whereas for the black box setting, the success rate was in the range of 52.7% to 71.6%. Hussain et al. [72] employed different white box and black box attacks on deepfake videos to bypass deepfake detectors. Simple white box, robust and transferable, query-based black box, and query-based robust black box attacks were performed based on existing attacks like universal and transformation-based attacks. From all attacks, robust and transferable attacks attained a higher misclassification rate of up to 99.9%, with the lowest classification accuracy of 0.02 on the neural texture subset of the FF++ dataset. In [43] facial mole-based black-box attack was proposed that reduces detector accuracy by up to 40.3%. The findings underscore the need for more robust, attack-resistant deepfake detectors.

6.3 Analysis and discussion

In this analysis, the potential strengths and limitations of adversarial attacks performed on visual deepfakes detectors, are discussed. For the GAN-based adversarial attacks, it is observed that these attacks are evaluated on GAN-generated facial images. Attacks [30, 50] are evaluated on the perturbated images generated using the StyleGAN-generated synthetic fake images. While the trace removal attack [53] is evaluated on the adversarial examples generated from a private dataset consisting of different deepfakes types including synthetic faces, attribute manipulation, and face replacement. However, GAN-based adversarial attacks are not evaluated for the perturbated images generated utilizing the challenging and standard datasets (i.e., FF++, DFDC, CelebDF) in the domain of deepfakes detection. In terms of victim models, these attacks are evaluated only for spatial-based deepfakes detectors except for trace removal attacks. The effectiveness of trace removal attack [53] is evaluated for spatial-based, frequency-based, and fingerprint-based deepfakes detectors; however, it only significantly affects the misclassification rate of fingerprint-based detectors. For reconstruction-based attacks, it is inferred that these attacks are not assessed specifically for fake facial images; however, they are evaluated for different fake images generated via 16 GAN-based methods. So, a need exists to

access reconstruction-based attacks specifically for the deepfakes facial images covering the diverse types of deepfakes. Moreover, such attacks cannot be considered too powerful compared to state-of-the-art adversarial attacks (i.e., C&W, FSGM, PGD), as the adversarial training on reconstructed images can significantly improve the detection results. In the reviewed papers, different attacks are performed on different models using adversarial examples that are crafted using either GAN-generated images or other deepfakes dataset images such as FF++ and DFDC. Most of the evaluated victim models are spatial-based classifiers. Only the frequency adversarial attack and StatAttack are evaluated on frequency-based models. Furthermore, none of the reviewed papers have utilized the GANFingerprint-based classifiers as a victim model. From the reviewed papers, it can also be concluded that, for the GAN-generated images, the success rate of the attacks is quite good in both white box and black box settings, compared to the other deepfakes dataset images. However, the SA-EA black box attack was not effective as it achieved a success rate below 20%. Additionally, the UAP attack is quite effective in both white box and black box scenarios and shows notable transferability. Overall, the adversarial attacks performed very well in the white box setting but in the case of black box setting, the misclassification rate decreases which indicates the limited transferability aptitude. However, pure black box adversarial attacks that are assessed on deepfakes detectors, are limited in numbers. Also, the success rate of the black box attacks is comparatively lower compared to the white box adversarial attacks, while fooling the deepfakes detection methods.

7 Adversarial attack on visual deepfakes generators

This section provides a detailed review of existing works on disrupting deepfakes creation through adversarial attacks. Adversarial attacks on deepfakes generations can be classified as watermark-based, and other adversarial attacks on real images. An overview of such existing literature is provided in Table 7 and the taxonomy of adversarial attacks on visual deepfake generators is given in Fig. 8.

7.1 Watermark-based adversarial attacks

Researchers have also introduced image tags and watermark-based approaches to stop the disinformation spreading on social media and the creation of fake personas. For instance, Wang et al. [10] introduced an image tagging approach based on an encoder, GAN simulator, and decoder to defend against deepfakes generation proactively. The DeepTag approach embedded a tag on the image using the encoder and then effectively recovered the hidden message through the decoder after the GAN-based manipulation was applied using the GAN simulator. The GAN simulator employed the existing frameworks, i.e., STGAN, StarGAN, and StyleGAN, to generate the manipulated facial images. This approach produces high-quality visuals but is computationally complex and resource intensive. Moreover, if the image is highly compressed, the embedded message might be lost. Likewise, Wang et al. [11] introduced a fake tagger, an imperceptible tag with images to bypass the deepfake generators to generate a falsified image. For implementation, U-Net was used as an image encoder to embed an imperceptible message into the image, and the decoder decodes the message from the tagged image. The CelebA-HQ dataset was used with several GAN-based architectures. The experiment was performed in a white and black box set. The quality of the encoded image was measured using PSNR and SSIM; the highest measure was achieved on entire synthesis deepfakes. This technique performed well in comparison with other watermark-based techniques. Zhao et al. [83] proposed a proactive detector utilizing an encoder-decoder architecture to encode facial identity features with watermarks that serve as anti-deep fake labels.

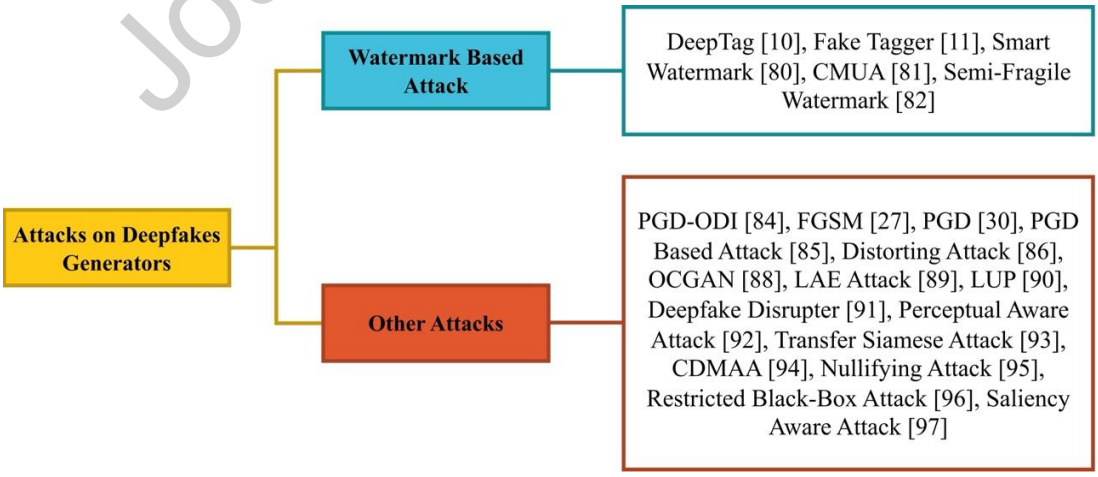

Fig. 8. Comprehensive taxonomy of adversarial attacks targeting deepfake generators.

The injected label exhibits sensitivity to translations involving face swaps. Experimental results demonstrate an average detection accuracy of over 80%. This approach offers users a dependable method to authenticate the legitimacy of their images, thereby mitigating the adverse consequences associated with deepfakes. In [84], the smart watermark framework was proposed, which consists of a watermark and attention module to generate unperceptive water-marked images. Such adversary images caused the deepfake generation models to produce the blur deepfake images. The watermark module extracted semantic information to create a personalized watermark while the attention module enabled the model to add the perturbation to the facial area utilizing the attention mask. Likewise, [85] introduced a cross-model universal adversarial (CMUA) watermark that can be applied to multiple facial images to generate adversarial examples to combat the various deepfakes generation models. Neekhara et al. [86] introduced a watermark-based technique named FaceSigns that is robust for image transformations (such as compression and filters) and fragile to GAN-based manipulation. FaceSigns framework embedded an imperceptible message in the images and was designed to detect the image as fake if the secret message's recovery rate is lower.

7.2 Other adversarial attacks

To protect the real images from being converted into deepfakes, many researchers have introduced the approach of performing adversarial attacks on the real images and thus hinder the creation of deepfakes. The major objective of such frameworks is to prevent the deepfakes GAN-based generators from manipulating the individuals' images that can be easily accessible from social media. To attack the style transfer model named CycleGAN, Liu et al. [87] introduced the PGD-based attack in which the starting point of the iteration was found using the output diversification initialization (ODI) method. ODI utilized the output space distance to locate the initial point for the iteration, and the adversarial image was generated using the iterative PGD attack, which was distinct from the input image. Ruiz et al. [88] applied existing adversarial attacks such as FGSM, I-FGSM, and PGD on the input real images to prevent fake images generated using the conditional image translation models, including GANimation, StarGAN, CycleGAN, and pix2pixHD. The authors also presented the spread-spectrum attack to evade the blur defense while generating deepfake images. In [89], a multi-objective algorithm based on a PGD attack was introduced for generating adversarial examples that caused the multiple image translation models to generate deteriorated fake images. When this non-transferable attack was performed on models (StarGAN-v2, STGAN) separately, the achieved success rate was 94.8% and 89.9%. However, the success rate reduced to 73.3% and 66.6%, respectively, when the models were simultaneously attacked. Yeh et al. [90] disrupted the image-to-image translation frameworks at inference time by utilizing the GAN discriminator as an adversarial loss function, where the utilized attacking technique was PGD. More specifically, two adversarial attacks, named nullifying attack and distortion attack, were performed on CycleGAN, pix2pix, and pix2pixHD frameworks. The nullifying attack enabled the model to output the image identical to the input image, while the distortion attack caused the model to generate distorted images that were clearly recognized as fake.

To improve the distortion attack [90], a training-resistant oscillating GAN (OCGAN) attack was introduced in [91] and failed the face-swapping frameworks by introducing imperceptible perturbation to the video frames. In the OCGAN framework, the adversarial generator and faceswap autoencoder were trained against one another such that the input image, along with the targeted distortion was given to the generator to produce an adversarial image. When the faceswap model was applied, it resulted in a distorted fake image. OCGAN attack introduced spatiotemporal perturbation and improved the transferability of different face-swapping models compared to the distortion attack. In [92], the latent adversarial exploration (LAE) method was introduced that searched the latent representations of the image to embed the perturbation, which resulted in high-quality perturbated real images that can defeat the different types of deepfakes generation. There is a lesser chance of the reconstructed image being identified as anomalous; however, the appearance of the image is sometimes slightly different from the original image. This demonstrates that the LAE method is not able to produce a perturbated image like the original one. Ruiz et al. [93] introduced an effective and simple black box attack named leaking universal perturbation (LUP) for disrupting the fake image generation through StarGAN and GANimation models. LUP attack achieved a success rate of 98% and 100% for GANimation and StarGAN frameworks, respectively, and significantly reduced the number of queries to perform the attack. Wang et al.[94] argued that making the fake image visually perceptible from the human viewpoint is insufficient as sometimes such images can fool the deepfake detectors. The deepfake disrupter, a perturbation generator was presented that added imperceptible perturbation to the real image and protected it to be manipulated by the deepfakes generation models such as StarGAN and GANimation. The deepfake images generated through perturbated real images can be identified as fake by both human and deepfake detection models.

Without degrading the visual quality, a perceptual aware perturbation was introduced in [95], which generated the adversarial examples natural to human eyes via introducing the noise on the color space in an incessant manner. The perturbation was added to the whole image and disrupted the generation of different deepfakes types (i.e., attribute

manipulation, face swap, facial reenactment). The adversary can crop the background to evade the perturbation if it is not in the facial region. Dong et al. [96] performed adversarial attacks against faceswap autoencoders to disrupt the fake images generated through GANs. Three different attacks were performed simultaneously on the Face-Scrub dataset. The first attack performed on images was a transfer adversarial attack, which was image agnostic. Later, Siamese attacks and Latent Siamese attacks were performed specifically to source images. All three experiments were performed on both reference and non-reference images of the dataset. The best SSIM and FSIM achieved on reference images were 0.49 and 0.79 respectively, based on the Siamese attack, while on a non-reference image, the quality latent Siamese adversarial attack achieved the best BRISQUE of 46.19.

Qiu et al. [97] proposed a gradient-based Cross-domain and model adversarial attack (CDMAA) to disrupt the generation samples of GANs. CDMAA attack was based on an I-FGSM attack, and generalization multi-gradient descent (MGDA) was used to provide generalization. The perturbation was added to the original images of the CelebA dataset, which fails GAN-based algorithms by generating unrealistic and distorted images. On StarGAN and U-GAT-IT, the highest achieved attack rate was 100%, whereas on AttaGAN and STGAN, the attack rate was 62.9 and 69.6, which is comparatively less. This attack was limited to images generated through GANs; it will be extended to audio and video as well. Yeh et al. [98] introduced a novel Nullifying adversarial attack for GANs to defend against a generation of malicious content. This attack was designed in a black box setting corresponding to the Limit-Aware Self-Guiding Gradient sliding attack (LaS-GSA) and nullified the effect of translation in GAN-generated images. The attack was performed on Black2Blond, None2Glasses, and Blue2Red GAN architectures and achieved the highest success rate of 95% on None2Glasses GAN, indicating the venerability of GANs towards adversarial attacks. In future work, the vulnerability of GANs against adversarial attacks will be reduced. Dong et al. [99] introduced a restricted black box attack that utilized a transferable cycle adversary generative adversarial network (TCA-GAN) to disrupt the face-swapped image generation. This technique protects deepfake generation based on translation methods. In comparison with existing attacks like FGSM, PGD, etc., this attack achieved the highest FISM and SSIM of 0.73 and 0.87. This technique was limited to the disruption of faceswap deepfake generation, and it will be extended to the generation of other deepfake types as well. Li et al. [100] proposed a saliency-aware attack framework to defend a well-trained deepfake generation model by manipulating the raw image with unperceived perturbation. It is achieved by selectively perturbing only the foreground person region and maintaining the irrelevant background to fool the model while minimizing alterations to the original image. This method [100] is limited to reliance on a pre-trained saliency detection model, which may not accurately detect the foreground region and reduce the effectiveness of the perturbation.

Table 7. Comprehensive overview of existing literature on adversarial attacks on deepfakes generations.

Year	Attack	Dataset	Victim Model	Results	Perceptual Similarity Measure	Limitations
Watermark-based Adversarial Attacks						
2020	DeepTag [10]	CelebA-HQ	StarGAN	PSNR = 26.32 SSIM = 0.862	PSNR, SSIM	Computationally complex and resource intensive.
			STGAN	PSNR = 27.48 SSIM = 0.901		
			StyleGAN	PSNR = 29.89 SSIM = 0.927		
2021	Fake tagger [11]	---	DeepFaceLab, Face2Face, STGAN, Style GAN	---	SSIM, PSNR	---
2021	Smart watermark [84]	CelebA	StarGAN	SSIM = 0.8231	SSIM	---
2022	CMUA-watermark [85]	CelebA, Film100	StarGAN	FID = 2.3032	Frechet Inception Distance (FID)	---
			AGGAN	FID = 1.7072		
			AttGAN	FID = 1.8133		
			HiSD	FID = 1.9672		
2022	Semi-Fragile Watermarks [86]	CelebA	---	PSNR = 36.08 SSIM = 0.975	PSNR, SSIM	Embedded messages will be lost for unseen deepfakes manipulation. Does not perform experiments on victim models to show attacks' effectiveness.

2023	Proactive mechanism [83]	FFHQ, CelebA-HQ, CelebA	AttGAN STGAN	Acc = 80	---	---
Other Adversarial Attacks						
2020	FGSM, I-FGSM, PGD [88]	CelebA	GANimation	FGSM $L_1 = 0.090$, $L_2 = 0.017$ I-FGSM $L_1 = 0.142$, $L_2 = 0.046$ PGD $L_1 = 0.139$, $L_2 = 0.044$	L_1 norm, L_2 norm	---
			StarGAN	FGSM $L_1 = 0.462$, $L_2 = 0.332$ I-FGSM $L_1 = 1.134$, $L_2 = 1.525$ PGD $L_1 = 1.119$, $L_2 = 1.479$		
2020	PGD-based attack [89]	CelebA	StarGAN	PSNR = 2.1561 SSIM = 0.1852 $L_2 = 1.6697$	PSNR, SSIM	---
			STGAN	PSNR = 26.5652 SSIM = 0.9153 $L_2 = 0.0082$		
			StarGAN-v2	PSNR = 11.711 SSIM = 0.4304 $L_2 = 0.0884$		
2020	Nullifying attack, Distorting attack [90]	CelebA-HQ	CycleGAN	**Nullifying attack** Similarity score = 0.21, Distortion score = 0.0725 **Distorting attack** Similarity score = 0.03, Distortion score = 0.165	Similarity score, distortion score	Transferability of distortion attacks needs to be improved.
			Pix2PixHD	**Nullifying attack** Similarity score = 0.34, Distortion score = 0.12 **Distorting attack** Similarity score = 0.02, Distortion score = 0.15		
			Pix2Pix	**Nullifying attack** Similar score = 0.27, Distortion score = 0.09 **Distorting attack** Similarity score = 0, Distortion score = 0.17		
2020	OCGAN [91]	Own dataset	realface,	ST score = 0.018	Spatial-temporal score (ST score)	---
			dfl-h128	ST score = 0.061		
			dfl-sae	ST score = 0.075		
2021	PGD-ODI [87]	CelebA	CycleGAN	$L_2 = 0.283$ Distortion score = 0.47	L_2 norm, distortion score	Limited to the gradient-based attacks and target only one GAN model.
2021	Transfer, Siamese, Latent Siamese [96]	Face Scrub dataset	Deepfake autoencoders	BRISUE=46.19	SSIM, FISM, BRISUE	Results were concluded on 1500 samples only.
2021	Nullifying Attack [98]	HQ-CelebA	Img2Img GANs	LAS-GAS ASR= 85	Query Count	The vulnerability of Img2Img GANs against attacks should be reduced
2022	LAE attack [92]	CelebA-HQ	SimSwap	Similarity = 0.21	L norms, normalized mean error (NME), mean confidence difference (MCD)	Not able to produce an adversarial example identical to the original fake image.
			GANimation,	$L_2 = 0.001$, NME = 0.025		
			StarGAN	$L_2 = 0.014$, MCD = 0.10		
2022	LUP [93]	CelebA	GANimation	Avg. Norm = 3.07, Success Rate = 98.6	---	---
			StarGAN	Avg. Norm = 6.36, Success Rate = 100		
2022	Deepfake disrupter [94]	CelebA, VoxCeleb1	StarGAN,	$L_2 = 0.326$	L_2 norm	---
			GANimation	$L_2 = 0.073$		

2022	Perceptual aware adversarial attack [95]	CelebA	StarGAN	PSNR = 5.292, SSIM = 0.251	MSE, PSNR, SSIM	Adversary can evade the attack by cropping the background, if the perturbation is not introduced in the facial area
			AttGAN	PSNR = 15.975, SSIM = 0.634		
			Fader Network	PSNR = 11.047, SSIM = 0.320		
2022	CDMAA (based on I-FGSM) [97]	CelebA	StarGAN	ASR = 100	---	Experiments were performed on image dataset only.
			AttGAN	ASR = 62.9		
			SRGAN	ASR = 62.4		
			U-GAT-IT	ASR = 100		
2023	Restricte Black Box Attack [99]	TCA-GAN	StarGAN,	StarGAN = 0.591	BRISQUE, SSIM, FISM	Limited to distort face swap generation.
			GANimation	GANimation = 0.796		
			SaGAN,	SaGAN = 0.864		
			AttGAN	AttGAN = 0.841		
2023	Saliency-aware Attack [100]	CelebA	StarGANs	L_1 = 0.016, L_2 = 0.068	L_1 and L_2 error	Limited to fail starGAN only.

*(Acc=Accuracy, ASR=Attack Success Rate, PSNR = Peak Signal-to-Noise Ratio, SSIM = Structural Similarity Index)

7.3 Analysis and discussion

This section includes an analytical discussion of adversarial attacks on visual deepfake generation methods. Adversarial attacks on deepfakes generators mainly involve two types of attacks, one that causes the deepfakes generation model to output the disrupted images. The other causes the generation model to be non-functional and output an image like the input image. LAE attack [92] disables the deepfakes generation by not putting the desired manipulation in the input adversarial example (LAE-generated image). When the LAE-generated image is passed to the manipulation algorithm, it outputs an image like the target image but still a deepfake image that is not visibly disrupted. So, this attack [92] cannot be considered too effective in preventing the deepfakes generation. Mostly, existing works introduce attacks that lead image translation models (such as StarGAN, Pix2Pix, CycleGAN, etc.) to deliver distorted images. For instance, the paper [90] performed both types of attacks but evaluated only for CycleGAN, Pix2Pix, and Pix2PixHD models (which generate attribute-manipulated deepfakes). The methods [84] and [85] utilized the watermark techniques, which caused the deepfakes generation models to output the blur or disrupted fake images. The smart watermark framework [84] only defends against the StarGAN manipulation algorithm, while the CUMA watermark [85] fails the StarGAN, AttGAN, HiSD, and AGGAN manipulation techniques. Most of the attacks are performed using GAN-based methods that generate either attribute-manipulated images or entire synthetic faces. Thus, the effectiveness of such attacks for other types of deepfakes generated through different deepfakes generation models is unknown. Most of the adversarial attacks generate adversarial examples against specific deepfake generation models, indicating the limited transferability aptitude of such attacks. However, the OCGAN attack [91] showed good transferability but only for the two different face swap generation models. Moreover, in existing works, significantly less attention is given to the black box attacks for disrupting the deepfakes.

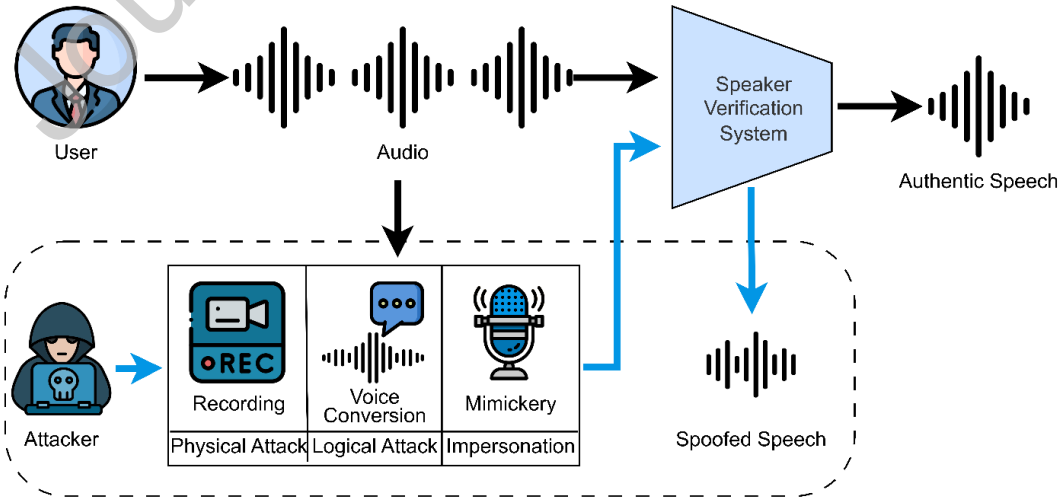

Fig. 9. General framework of attacks on speaker verification system.

Additionally, the existing works cause the generation models to generate disrupted fake images, which can be easily identified as fake by humans, but the performance of deepfake detection methods on such images is not demonstrated. Only in [94] it shown that the disrupted images generated through their method were also correctly identified as fake ones by deepfakes detectors.

Overall, the limitation exists in the transferability of adversarial examples among different deepfakes generation methods. So, it is required to show the effectiveness of adversarial attacks for unknown deepfake generation methods, because, in a real-world scenario, it is unknown which deepfakes generation method is going to be used. Additionally, adversarial attacks on deepfakes generation models can be ineffective in the case when adversarial noise detectors are utilized to identify the anomalies before the deepfakes generation. Also, the blurring and compression techniques, if applied to the adversarial examples generated, can minimize their effectiveness while disrupting the deepfakes generation.

8 Adversarial attacks on audio/voice recognition models and system

Adversarial attacks on the audio/voice domain include the addition of perturbations to the audio signal to trick deep learning models or voice detectors. Existing surveys [3, 23, 101, 102] include multiple attacks with categorization in this domain; however, we reviewed selective literature of recent years, including targeted or non-targeted attacks. In addition, depending on the knowledge of attacks, white box and black box attacks are also reviewed. We categorize these attacks into gradient sign methods and optimization attacks. The general framework of attacks on audio detection and speaker verification systems shown in Fig. 9 and Table 8 presents the overview of audio-based attacks.

8.1 Gradient sign-based adversarial attacks

In this categorization, gradient-based methods were used to generate adversarial examples. Liu et al. [103] proposed targeted gradient sign-based FGSM and PGD attacks. These attacks were performed on mel spectrograms of the ASVSpoof 2019 dataset in a white and black box setting. Several victim models were used, including LCNN-Big, LCNN-Small, and SeNet. The best EER achieved on the LCNN model is 93.11% by PGD attack. Li et al. [104] proposed room impulse response (RIR) modeling to investigate the realism and practicability of adversarial training with a real-time, over-the-air attack. A practical and systematic white box adversarial attack named X-vector was conducted to challenge the state-of-the-art deep neural network (DNN) based speaker recognition system. It caused the speaker recognition system to make inaccurate predictions or even force the audio to be recognized as any adversary-desired speaker by adding a carefully constructed, unobtrusive noise to the original audio. It is possible that incorporating an assault on the estimated room impulse response (RIR) into the adversarial example training process will result in useful, over-the-air audio adversarial examples. This attack [104] gives a 98% success rate in digital attacks and a 50% success rate with ASR 50 on over-the-air attacks to the x-vector using the VCTK dataset. Zhang et al. [105] proposed an ensemble method based on the MI-FGSM adversarial attack in a white box setting. In addition, adversarial examples created through this technique failed ResNet 34 with an ASR of 100%. The black box attacks vary significantly in ASR across models, from 24% to 84%, and this method causes perceptible perturbation Wang et al. [106] proposed an imperceptible, inaudible white box adversarial attack. This method [106] achieved an ASR of 98.5% by limiting the perturbation to the actual audio's mask threshold and producing targeted, inaudible perturbed samples of the original sound waveform. They also applied their method to other waveforms, such as music.

8.2 Optimization based attacks

Nakamura et al. [107] proposed a white box verification-to-synthesis attack on the voice conversion (VC) system. With the possibility of the trained network distorting the input voice's phonetic features, an automatic speech recognition model is incorporated into the optimization process to control the amount of phonemic information that is lost. The resulting output voice is not only convincing to the ASV system but also retains its perceived quality. The experiments were performed on a d-vector using a Japanese dataset. Xie et al. [108] proposed an attack, which adds distortions in the sound due to the physical over-the-air propagation of the room, utilizes that information to estimate the impulse response of the room (RIR), and achieves an attack success rate of over 90% and an ASR of 90.19 on the x-vector using the VCTK dataset. Li et al. [109] examined the susceptibility of Gaussian Mixture Model (GMM) i-vector-based speaker verification systems to adversarial attacks. Adversarial samples created for GMM i-vector-based systems may be used successfully in x-vector-based systems, and FGSM enhances testing phrases for producing adversarial samples in a white box setting. Both the i-vector and x-vector versions of the GMM are vulnerable to attacks as the accuracy of the system falls from 87.50% to 25.75% on the VoxCeleb1 dataset. Chen et al. [110] introduced optimization-based FGSM for adversarial training, where white box and black box attacks were performed against the same ASV in a cross-corpus and cross-feature scenario. This method [110] employed an optimization-based approach to estimate the score threshold and perform targeted attacks against the i-vector and GMM-UBM

using the LibriSpeech and VoxCeleb1 datasets. For speech emotion recognition (SER), a generator-based STAA-Net attack was proposed [111] that efficiently generates transferable and sparse adversarial perturbations in an end-to-end manner. The generator was trained using Emo18 and WavLM as threat models and produced adversarial examples in a single forward pass. The generated adversarial examples were used to attack the Zhao19 and wav2vec 2.0 models and show good transferability.

8.3 Other attacks

These attacks are comprised of techniques that are either an ensemble or a combination of multiple methods, not included in the categories. In [112], dictionary attacks on ASV were proposed, in which attackers can target a large speaker population with this method, even if they don't have detailed information about any particular speakers or models. It also suggested an adversarial optimization method for artificially creating master voices. Dictionary attacks are a realistic security concern for mission-critical applications of speaker verification. The attack was performed in a white box setting on the VGGVox model using the VoxCeleb2 dataset. Tian et al. [113] investigated black box attacks on ASV using a feedback-controlled VC architecture. The feedback ASV score is used together with the objective function to maximize the training of the feedback-controlled VC. The findings suggested that ASV functionality can be harmed by black box attacks. The adversarial examples generated without ASV input are indistinguishable, and experiments were performed on an i-vector using the ASV spoof 2019 dataset with a 30.73% EER rate. Chen et al. [114] proposed multiple targeted and non-targeted attacks in white and black box settings against the speaker recognition system that are combinations of gradient sign, evolutionary, and optimization-based methods. These perturbations based on PGD, FGSM, FAKEBOB, and C&W attacks, etc., were injected into Mel-spectrograms of voice signals to fool the system. Experiments were performed on the LibriSpeech dataset in different experimentation settings, but overall, the C&W attack with the L2 white box attack attained the highest attack success rate of 97.1%.

Abdullah et al. [115] proposed a discrete Fourier transform (DFT) and singular spectrum analysis (SSA)-based targeted black box attack that introduces perturbations to every single phoneme after a few words. This feature-based method targets the MFCC of the TIMIT and word audio datasets. Chen et al. [116] proposed an arbitrary source-to-target attack by adding perturbation to waveform signals of voice on 14 different speaker verification systems; these attacks include multiple gradient signs and optimization methods. This attack included target and non-targeted scenarios in white and black box settings on the LibriSpeech dataset. This attack achieved the best ASR rate of 100% on the ECAPA model, but it is limited to the source-to-target attack only. Rabhi et al. [117] show audio deepfake classifiers are vulnerable to adversarial assaults. Two new methods target the Deep4SNet classification algorithm, which first detects counterfeit audio samples with 98.5% accuracy. Generative adversarial network (GAN) assaults reduce detector accuracy to almost 0%. The detector's accuracy reduces to 0.08%, even with gray box attacks. The paper emphasizes the need to research the reliability of alternative audio-deep fake detection structures. He et al. [118] explored the unique characteristics of adversarial audio, focusing on phonetics. Researchers analyzed 2,400 audio samples, revealing 612,000 acoustic-statistical features, including energy gaps, speech-like morphology, disordered signals, and anomalous linguistic patterns. They developed a naturalness score and proposed an adversarial example detector with an average precision of 91.1%. Experimental investigations showed that adversarial examples significantly affect model accuracy, increasing false positives and false negatives and decreasing overall accuracy. Further research is needed to refine detection techniques, explore innovative defense strategies, and understand adversarial attacks. Umar et al. [119] introduced an ensemble of surrogate model–based losses combined with a transcription loss to enhance the transferability of GAN-based anti-forensic attacks across white-box, gray-box, and black-box audio deepfake detectors.

8.4 Analysis and discussion

The susceptibility of audio-based systems to adversarial attacks and their potential vulnerabilities are discussed in this section. A gradient-based adversarial attack [103-106] includes modifying the loss function's gradient concerning the input instances and then introducing a small, carefully prepared perturbation in the direction that maximizes the loss. This technique is quick and capable of producing adversarial samples quickly, but these attacks may degrade the performance during transferability from the white to black box setting [103]. Gradient-based adversarial attacks may be less effective than optimization-based attacks, as they may not identify the optimal perturbation that maximizes the loss.

An optimization-based adversarial method [107-110] involves the resolution of an optimization process to determine the ideal perturbation that maximizes the loss. This method is computationally costly, but it can produce more effective adversarial examples since it determines the ideal modification that maximizes the loss. Optimization-based adversarial attacks may employ gradient descent, optimization, genetic algorithms, or other optimization approaches. However, optimization-based adversarial attacks may be more complex to implement and demand additional computer

resources than gradient-based attacks. There are several techniques that are either combinations of other methods or attacks [112-116]. In general, the decision between several categories of adversarial attacks is determined by the requirements and resources of the system. Gradient-based adversarial attacks are quicker and less complicated, but they may not be as effective as optimization-based attacks. Optimization-based adversarial attacks are more successful, but they demand additional computer resources and may be more challenging to implement.

Table 8. An overview of adversarial attacks on audio.

Year	Attack	Victim Model	Dataset	Perceptual Similarity Measure	Results		Limitations
					Before Attack	After Attack	
Adversarial Attacks on Audio							
Gradient Sign Based Adversarial Attack							
2019	FGSM, PGD (White and Black box) [103]	LCNN-Big LCNN-Small SeNet	ASVspoof2019	EER=93.11	-	Acc= 48.4	Less effective in a black box setting
2020	Adding perturbation with RIR, black box [104]	x-vector	VCTK corpus, Kaldi toolkit	-	ACC=98	ASR=96.9	Some perturbations are less effective
2020	MI-FGSM ensemble-based attack [105]	LCNN/AFNet, SENet50, ResNet34	ASVspoof2019	ASR = 84(black) 100(white)	-	-	ASR degraded during white to black box. Transferability.
2021	Inaudible [106]	x-vector	Aishell-1	Score=91.5, ASR=98.5	-	-	Limited to 1 dataset testing
Optimization-Based Adversarial Attack							
2019	White box V2S attack [107]	d-vector	Japanese dataset	Score= 0.713	-	-	Results were not measured by standard measures.
2020	Adding perturbation with RIR, White box [108]	x-vector	VCTK corpus	ASR= 90.19	-	-	computationally complex, evaluated on one dataset
2020	FGSM (white box) [109]	GMM (i-vector and x-vector)	Voxceleb	EER= 99.95, FAR=99.99	-	-	Computationally complex, limited to one model.
2021	FGSM (Targeted White and black) [110]	i -vector, GMM-UBM	LibirSpeech Voxceleb	FRR=4.2, FAR=11.2, ASR=99, UTR=99	-	-	Tested with two datasets only.
2024	STAA-NET [111]	Emo18, Zhao19, Wav2vec, WavLM	DEMoS, IEMOCAP	ASR=17.5 SNR=92.8	-	-	Tested with two datasets only.
Other Adversarial Attack							
2019	Dictionary attack, white box [112]	VGGVox	VoxCeleb	EER=8, FAR=1	-	-	Not robust
2019	Feedback-controlled VC attack [113]	i-vector	(WSJ) corpus and CMU-ARCTIC database	area of shift (female=7.40, male= 4.62)	-	-	Results were not measured by standard measures.
2021	FGSM [114]	GMM i-vector, x-vector, AudioNet	LibirSpeech	ASR=100, L_2=0.082, SNR=46.93, PESQ=3.77	ACC=99.8	C&W=20, ACC=5.9	Limited to few attacks.
2021	4 targeted black box attacks [115]	Google (Normal), Google (Phone), Wit, DeepSpeech1, and Sphinx	Word audio dataset and TIMIT	PEQS=1.7, ASR=100	AUC=0.93	AUC=0.52	
2022	AS2T attack(white and black) [116]	GMM, CNN, LSTM, and 11 other	LibirSpeech	ASR= 97.4, EER= 2.2	-	-	Limited to specific systems
2024	White, grey, and black box attacks [117]	Deep4SNet	LJSpeech dataset	ASR= 87.6	Acc=98.5	Acc=0.8	Limited to the specific dataset

2024	White and black box attack [118]	Kaldi DNN-HMM model, DeepSpeech model, and LAS model	VCTK corpus	Acc=91.1	-	-	-
2024	GAN-based attack [119]	AI-generated audio	RawNet2, TSSDNet, ResNet, MS-ResNet2	PSNR, SSIM	Acc=93.0,	After_Acc = 61.4	Less effective and less performance drops.

*(Acc=Accuracy, ASR=Attack Success Rate, PESQ = Perceptual Evaluation of Speech Quality, EER= Equal Error Rate, FRR = False Rejection Rate, FAR = False Acceptance Rate)

9 Countermeasures against adversarial attacks on visual deepfakes detection and generation

Adversarial defense techniques are designed to prevent malicious attacks from deceiving the models. The objective could be accomplished by either strengthening the model's robustness or eliminating the attacker process. The former can be further subdivided into proactive and passive defense algorithms. The distinction between these two categories is that proactive defense intends to enhance the performance of the model, whereas passive defense seeks to reduce or completely remove adversarial perturbations. The following sections provide the existing literature for these methodologies. The taxonomy of defensive techniques against visual adversarial attacks is shown in Fig. 10.

9.1 Proactive defensive techniques against visual adversarial attacks

Proactive defense strategies aim to improve the model's robustness and enable it to perform accurate predictions for adversarial samples. Proactive defense typically requires further training of the model or additional optimization of model parameters and structure to make it different from the original model. We focus on representative proactive defenses in this section, primarily adversarial training, optimal neural networks, and ensemble learning. The general framework of proactive defense is given in Fig. 11.

9.1.1 Adversarial training-based defense

Adversarial training is the process of injecting adversarial samples into the model's training data to make it adversarial robust. The adversarial training framework is widely considered one of the most effective principled defenses against adversarial attacks as it exposes the model to adversarial samples in training to gain some level of immunity. In [16], an adversarial training-based min-max game was introduced as a proactive measure to an adversarial attack. Training samples of the FF++ dataset were blurred with pixel-wise Gaussian blurring during the generation process. This technique helped the model to detect deepfakes in a generalizable and robust manner. The efficient and combined model achieved the best results, which was 98%. In [66], Naive Max-pooling and Feature Max-pooling defense methods are used against various attacks in both black and white box testing. To improve the robustness of the defensive method, the detector is retrained with the adversarial samples obtained after the attacks. The detector is based on ResNet-50 and trained on CelebA-HQ, with four provided attribute vectors: age, smile, pose, and gender. Without any attacks, the original detector has classified 1000 inputs with 99.6% accuracy and with an average prediction rate of 98.7%. The defense detector retrained with StyleGAN has correctly classified all 1000 inputs with an average confidence prediction rate of 99.5%. In [120] GAN-based network features a dual channel that consists of multiple modules to provide supplementary knowledge from latent space. It is proposed as a countermeasure against

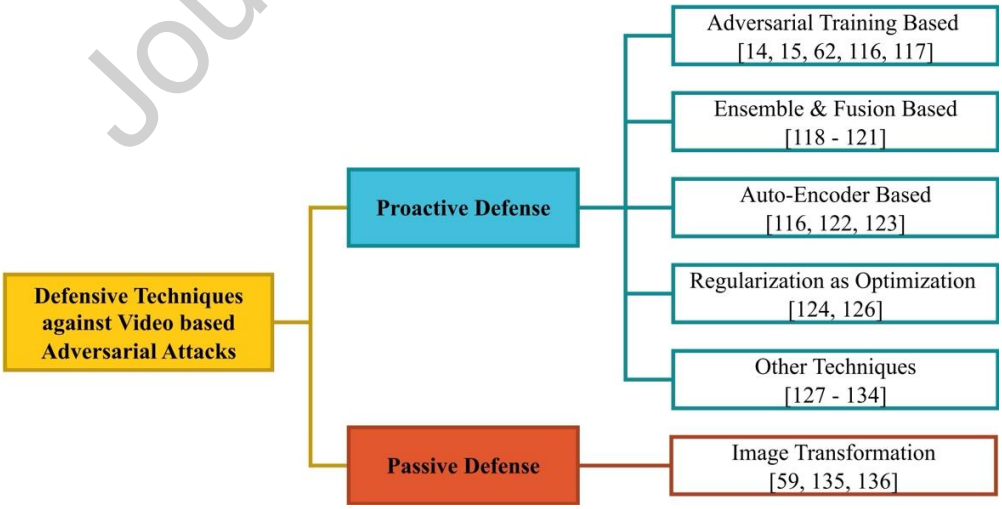

Fig. 10. Comprehensive Taxonomy of Defensive Techniques Against Visual Deepfake.

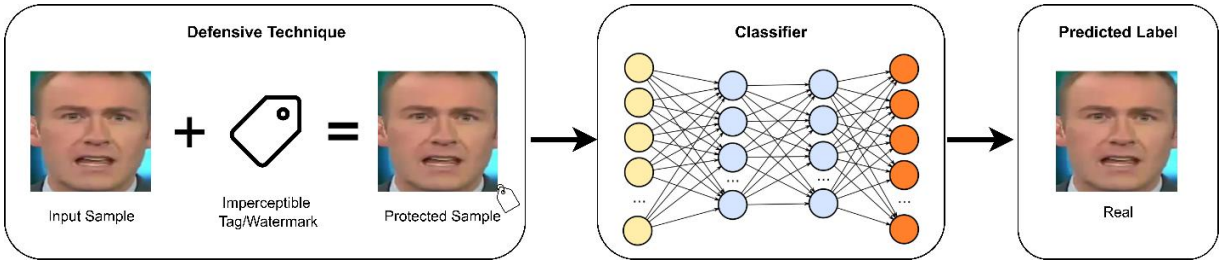

Fig. 11. Detailed schematic of the proactive defense technique in deepfake detection.

adversarial attacks on deepfakes. The model is beneficial for training forensic detectors with strong robustness against adversarial attacks. It is neutralized by adversarial perturbations by injecting new noises. The results show that the defensive method has shown good precision on DenseNet, ResNet, XceptionNet, and DefakeHop when tested against the other methods. Uddin et al. [121] introduce a novel method for identifying anti-forensic attacks on High-Efficiency Video Coding (HEVC) compressed videos, with a particular emphasis on those produced by GANs. The study enhances the efficacy of forensic tools in identifying tampered videos by employing deep learning to analyze coding patterns and detect GAN-based manipulations. However, it may not cover all attack scenarios and is dependent upon the use of extensive datasets. Furthermore, the efficacy of the method may fluctuate depending on the type of video content and the compression level. Chen et al. [17] proposed a deceptive mechanism based on statistical hypothesis testing to detect deepfake manipulation and adversarial attacks. A deceptive model with two isolated sub-networks is designed to generate random variables for deepfake detection, and a maximum likelihood loss is used to train the model. The experiments show the effectiveness of the decoy mechanism in detecting both Deepfake and adversarial attacks across different methods. A limitation of the proposed method [17] is that it assumes the attacker is unaware of the decoy mechanism and the training process. If the attacker gains knowledge of these aspects, they can develop targeted attacks that may bypass the detection system.

9.1.2 Ensemble and fusion-based defense

Ensemble learning for adversarial defense can be summed up as training an ensemble of models to achieve the same class label, but every model should be different as much as possible for improved generalization performance. In [122], a robust deepfake detection architecture was proposed called EnsembleDet. This ensemble model can detect deepfakes against the fast gradient sign method and basic iterative method. For the experiment, the FF++ dataset was used to train the classifiers. Different classifiers, including MesoNet, XceptionNet, simple ensemble, and GASEN-based ensemble, were used. The simple ensemble model achieved higher accuracy against both FGSM and BIM attacks on all subsets of FF++ datasets, whereas the DF subset of the FF++ GASEN-based ensemble achieved the highest accuracy against FGSM attacks. Although the simple ensemble model showed good accuracy, its precision rate is lower than that XceptionNet. The increasing number of parameters during the ensemble-based model training causes high computational costs. An ensemble method [123] tackles the adversarial attacks like L_2 and L_∞ infinity perturbation norms performed on multiple shallow architectures of Flickr faces HQ dataset under both black and white box testing. In this paper, Disjoint Deepfake Detection (D3) is proposed as a defensive method against adversarial attacks. D3 utilizes disjoint partitions of the input characteristics to construct a solid ensemble of models. By analyzing the results with state-of-the-art defense methods, it can be noticed that D3 maintains 100% adversarial accuracy against AutoAttack, whereas the other methods drop below 20%. Kumar et al. [124] studied modern methods for identifying generative artificial intelligence-generated multimedia content. They classified single-modal and multi-modal detection strategies as conventional or advanced techniques, using machine learning for handcrafted features and deep learning and hybrids for improved detection. Amerini et al. [125] developed an ensemble approach called D-Fence to detect deepfakes. The system classifies altered facial and vocal features and Video-Audio and Audio-Text interactions. The D-Fence layer was tested against two novel adversarial attacks: Bogus-in-the-middle and down sampling. D-Fence achieved 92% detection accuracy despite obstacles. In sophisticated multi-modal deepfake detection, D-Fence outperforms the classifiers.

9.1.3 Autoencoder-based defense

Autoencoder-based defensive techniques provide reconstruction phenomena against adversarial modified samples to remove perturbation. In [126], a residual fingerprint-based defense is introduced against BIM, FGSM, PGD, Inversion, and additive noise attacks. The reconstruction method poses strategies to degrade adversarial efficacy and extract discriminative residual fingerprints. It also reconstructs adversarial samples by removing the original deepfakes from the corresponding adversarial ones. To maximize the difference between real and fake images, another strategy of

transforming the residual fingerprints with a Discrete Cosine Transform component is used, which produces discriminative traces for further deepfake detection. The defense model shows significantly improved accuracy from 50.88% to 84.05% in classifying the FGSM-based adversarial images. In [127] MagDR, a Mask-Guided Detection and Reconstruction pipeline, is used to defend deepfakes from adversarial attacks by C&W and PGD. MagDR starts with a detection method that establishes a few criteria for assessing the abnormality of deepfake output and uses those criteria to direct a learnable reconstruction process. To record the change in specific facial regions, adaptive masks are extracted. From the experiments, it can be analyzed that MagDR defends three main types of deepfakes under both black and white box attacks. Extensive experiments are performed to demonstrate the effectiveness of the model on defense-aware and defense-unaware attacks, and it can be analyzed that the MagDR has shown the highest detection as compared to the state-of-the-art models. Ding et al. [120] proposed a method to improve the security of facial bioinformation by identifying and eliminating adversarial perturbations. This includes the preprocessing of facial images to prevent potential perturbations, thereby guaranteeing accuracy and reliability. The effectiveness of the experiment is in protecting bioinformation from attacks. However, the method's potential performance degradation and the difficulty of employing it in sophisticated techniques are among its limitations.

9.1.4 Regularization as optimized defense

According to researchers, deep neural networks are vulnerable because of their weak structure and parameters, making them susceptible to adversarial perturbations at the image level. The methods based on neural network optimization aim to increase the robustness of the model by enhancing its parameters, such as regularization, or modifying its structure. In [128], adversarial attacks FGSM and C&W L_2 are performed on ResNet and VGG in black as well as in white box settings on the fake images created through Few Shots face Translation GAN [129]. Lipschitz Regularization and Deep Image Prior (DIP) defensive techniques were used in [128] to make the model resistant to adversarial perturbation. On average, Lipschitz regularization improved the detection of adversarial perturbed deepfakes by ResNet models in the white box CW- L_2 setting, where even the regularized model correctly classified only 2.2% of the perturbed fake images. However, the DIP defense illustrates more encouraging findings. With a classification threshold of 0.25, it achieved a recall of 97.8% for both perturbed and unperturbed fake images. It can be noticed that DIP defense outperformed regularization for deepfake detection, as it improved robustness to adversarial perturbations slightly, but the performance remains impractical for real-world scenarios. In [130] a regularization based video deepfakes detector is proposed. Regularization strengthens the model's generalizability, making it robust against diverse manipulations and postprocessing attacks.

9.1.5 Other techniques

Many other proactive defense techniques are introduced with the aim of protecting the content from malicious attacks. Deb et al. [131] introduced a unified attack detection system (UniFAD) that can learn 25 coherent attack types from the three categories of adversarial, digital manipulation, or physical spoofs. Mostly, defensive techniques achieve perfect accuracy when tested against one of three types of attacks; however, when tested against all three types of attacks, their accuracy is downgraded. UniFAD learns joint representations for coherent attacks using a multi-task learning framework and K-means clustering, while uncorrelated attack types are learned separately. On the FakeFace dataset, having 25 different attack types of all three categories, the proposed UniFAD outperforms existing defense methods with an overall TDR of 94.73% and 0.2% FDR. Even UniFad can recognize the attack categories with 97.73%. Without affecting the significant artifact features, the DF-UDetector introduced in [132] detected the degraded deepfake images by converting the degraded feature maps into high-quality ones. The deepfake images are degraded with different strength levels of noise, blur, and compression. The attained accuracy of the model on different datasets was less than 90% for the degraded images but better than existing models. This method is robust against degradation attacks but is still vulnerable to adversarial attacks such as PGD, FGSM, etc. Uddin et al. [133] introduced a novel method to improve the resilience of digital face image classifiers against adversarial attacks. The technique enhances its capacity to manage adversarial inputs by employing open-set multi-instance learning to differentiate between known and unknown instances in an image. It evaluates the consistency of predictions across multiple instances to identify anomalies. However, it has limitations such as computational complexity and potential varying effectiveness based on attack nature.

Asha et al. [134] introduced an optical flow-based CNN with self-attention architecture that was robust against adversarial attacks while detecting deepfakes. Self-attenuated VGG16 extracted the informative facial features from which optical flow vectors were computed and then passed to the sequential CNN. This model attained a classification accuracy of 74% in the presence of adversarial attacks, which is reasonably good. Pinhasov et al. [135] introduced a defensive mechanism for the security of deepfake detectors by employing XAI to improve the identification of malicious adversarial attacks. However, the assessment depends on the FF++ dataset, and the computational expense

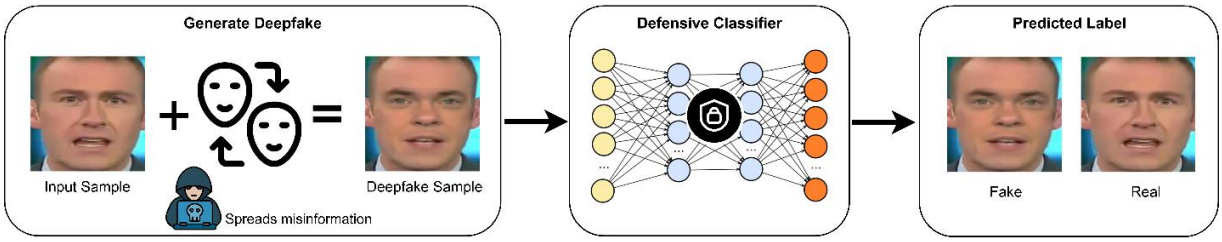

Fig. 12. Comprehensive diagram illustrating the general framework of passive defense techniques.

continues to be a hurdle for real-time implementation. Wang et al. [136] proposed a platform named DEEPFAKER, which is designed to evaluate the efficiency of deepfake creation and detection models. The platform's modular design enables effortless incorporation of new approaches, enabling researchers to include innovative detection techniques as the field progresses. However, it recognizes constraints such as the adequacy of the dataset and biases in the model. Park et al. [137] explore the simultaneous presence of deepfake disruption and detection in conventional security concerns. DDPM denoising training framework is employed to diminish defense time and minimize image distortion. The method produces deepfake images that closely resemble authentic ones, reduces defense time by 7.75%, and exhibits superior detection accuracy in disruption attacks compared to StarGAN and DiffPure.

9.2 Passive defensive techniques against visual adversarial attacks

Passive defense is intended to mitigate the damage induced by adversarial perturbation rather than to strengthen the model itself. The main advantage of this method is its ability to decouple model training and adversarial defense without modifying the already trained model, which allows defense methods to be implemented. The following section discusses passive defensive methods for protecting against adversarial attacks on visual deepfakes. The general framework of passive defense is given in Fig. 12.

9.2.1 Image transformation

Deep neural networks have been found to be generally robust when random image transformations are applied. However, perturbations are fragile under this method for adversarial samples, particularly those generated by the gradient. Based on this observation, researchers devised several defensive methods based on random image transformation techniques to counter adversarial attacks. In [63], in the first detection track, EfficientNet-B3 was used for deepfake detection of the Celeb-DF dataset. The fake image in the dataset was perturbed by PGD and MI-FGSM attacks. For better defense, the cross-entropy loss was employed with smoothing and up-sampling to prevent the model from misclassifying. In the second detection track, data augmentation, like a Gaussian blur, affine transformation, etc., was used to increase the dataset and trained with Efficientnet-b0 to improve the classification of a model against any adversarial attack. In the third detection track, FGSM was introduced in training samples to make the deepfake detector robust against adversarial attacks. EfficientNetV2 was used to train Celeb-DF original, augmented, and perturb images. Luo et al. [138] proposed bilateral filtering against three typical adversarial attacks: BIM, PGD, and FGSM on Xception and MesoInception in both black and white box settings on the subsets of the FaceForensics++ dataset. In this work, bilateral filtering is used in the passive defensive strategy, and joint adversarial training is used in the proactive defensive strategy. Leporoni et al. [139] proposed the Masked Depthfake Network (MDN), a technique that incorporates depth data into an RGB detector. For detection, the method extracts depth maps from RGB images and combines RGB features with depth information. MDN outperforms the average RGB detector, as evidenced by its 91.26% accuracy on Deepfakes and 96.86% accuracy on the FF++ dataset. In addition, this method outperforms black box attacks and exhibits an accuracy of 95.69%. All these techniques help to prevent deepfake detectors from misclassifying against adversarial attacks.

9.3 Analysis and discussion

Adversarial-based defense methods aim to make a model more robust to adversarial examples by either modifying the model's architecture or training it on adversarial examples. These defensive methods are more robust against white box attacks; however, they still fail in the presence of strong attacks. Ensemble-based defense methods involve combining multiple models to make a more robust system. Ensemble defensive methods perform very well on black box attacks but are very computationally complex. Autoencoder-based defense methods reconstruct the input data to detect and remove adversarial examples. These methods reconstruct the input and compare it to the original samples, and if the reconstruction is significantly different from the original input, it is an adversarial example. The autoencoder-based defensive methods have not shown remarkable results against adversarial attacks as compared to other defense methods. It can be seen from literature that some methods, which do not fall under a specific category, have shown remarkable results against the attacks and their subtypes. All these methodologies have their own pros

and cons; it depends on the nature of the problem, and the researcher decides which one is the best fit for their use case.

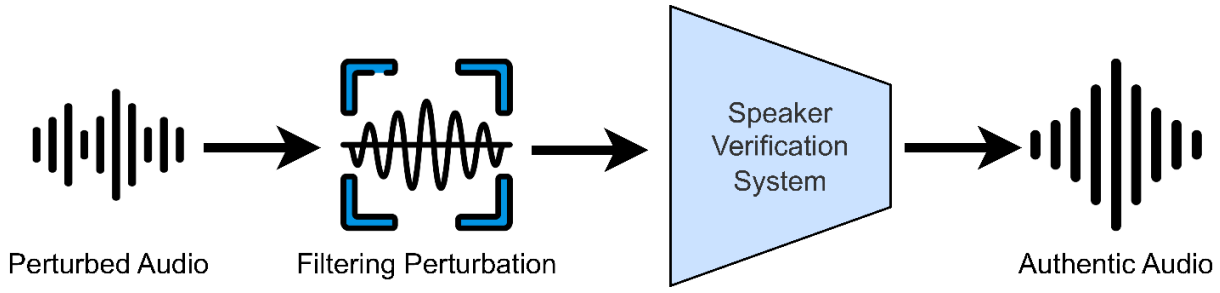

Fig. 13. Comprehensive diagram of audio defensive techniques in deepfake detection.

There is limited literature available on passive techniques against adversarial attacks. Researchers are mostly focusing on proactive defensive techniques for several reasons, such as aiming to prevent an attack from happening in the first place and being more cost-effective. As proactive defense methods have shown to be more effective in preventing attacks, there is still room for improvement in passive defense techniques. The research community can work towards making passive defense methods more efficient and effective by improving the ability of passive defense methods to detect and respond to attacks in a timely manner. Therefore, the research community has an opportunity to bridge the gap between proactive and passive defense methods by enhancing their performance of passive defense methods.

In this section, we comprehensively reviewed defensive methods against adversarial attacks and compared them in Table 9. Work reviewed under adversarial training and optimization-based methods has shown remarkable results against adversarial attacks. It is robust to adversarial attacks as it trains the model on small perturbations in the input by adding adversarial examples to the training data. This helps the model to learn to recognize and classify even adversarial examples that it has not seen before during training. This can make the model more robust to a wide range of attacks. Optimization-based methods aim to find the closest possible inputs to the adversarial example while still maintaining the desired output from the model. By finding the closest input, the optimization method can effectively reduce the impact of the adversarial perturbations and increase the model's robustness to attacks.

10 Countermeasures against audio-based adversarial attacks

In this section, defensive countermeasures against adversarial attacks on audio verification systems are presented. These systems are vulnerable to even small adversarial attacks, such as impulses and Gaussian noise, which can compromise their accuracy and reliability. To counteract these attacks, we reviewed several defensive techniques against well-known voice spoofing and adversarial attacks. We categorize these countermeasures into two groups: proactive and passive defense. The general framework of the audio defensive method is given in Fig. 13, and the literature is compared in Table 10.

Table 9. An overview of Visual Defense Techniques.

Year	Defensive Methods	Model	Dataset	Baseline Measures	Results		Limitations
					Attack	Defense	
Proactive Defense							
2020	Lipschitz Regularization, Deep Image Prior [128]	VGG-16, ResNet-18	Celeb-A	Accuracy, AUROC	FGSM, C&W	Acc = 60.0	Robust on one class, and its time taking/
2021	EnsembleDet [122]	Simple Ensemble	FF++	Accuracy, F1	FGSM	Acc = 94.45	High computational cost.
					BIM	Acc = 94.1	
2021	Residual Fingerprint-Based Defense using CNN [126]	Meso4, MesoIncpetion	Deepfake detection dataset	Accuracy, Precision	MesoInception		Computationally complex.
					BIM = 35.43	Acc = 81.65	
					FGSM = 44.32	Acc = 84.05	
					PGD = 33.30	Acc = 83.91	
					Inversion = 44.32	Acc = 85.25	

					Additive noise attack = 30.56	Acc = 85.25	
2021	MagDR, a Mask-Guided Detection and Reconstruction [127]	CycleGAN, StarGAN, GANimation	FF++, CelebA	PSNR, SSIM, MSE, Feature Similarity	**Face editing**		Leave perceptible artifices.
					C&W	Acc = 96	
					PGD	Acc = 100	
2022	Adversarial training [16]	EfficientNet, In-v3 Grad-AAT, Grad-SAT, etc	FF++	Accuracy, AUC	**PGD** Acc = 22.06 (Inc-v3)	Pixel level blurring Acc = up to 99.46	It may fail on a strong attack.
2022	Naïve Max Pooling, Feature Max Pooling [66]	Resnet-50	CelebA-HQ	Attack Success Rate	Pixel level adversarial perturbation	Acc = 75	--
2022	GAN based model [120]	DenseNet, ResNet-50, XceptionNet, DefakeHop	FF++, Celeb-DF	Accuracy, Precision	Adversarial noise	**FF++:** Acc = 99.24 **Celeb-DF:** Acc = 97.39	--
2023	Unified attack detection system UniFAD using JointCNN [131]	Joint CNN	GrandFake dataset (own)	TDR, Accuracy	25 coherent attacks	Acc = 94.73	Not generalizable on generic image manipulations.
2023	CNN with self-attenuated VGG16 [134]	VGG16	CelebDF, FF++	Accuracy, AUC, EER	FGSM, BIM, PGD	Acc = 74	---
2023	DF-UDetector [132]	EfficientNet	CelebDF, DFDC		Blur	Acc = 85	Not vulnerable to adversarial attacks such as PGD, BIM, etc.
					Noise	Acc = 78	
					Compression	Acc = 76.17	
2024	XAI based detection [135]	Xception net, EfficientNetB4ST	FF++	Precision	PGD, APGD, NES, Square Attack	Precision = 89.78	Computationally complex, Robust on one class
2024	D-Fence [125]	Ensemble VGG-16	DFDC, DF TIMIT, FakeAVCeleb, MMDFD	BERT	Acc=92	--	
Passive Defense							
2021	Cross entropy loss with image preprocessing techniques [63]	Own teacher and student models.	DFGC-21 testing dataset, FaceForensic++,	AUROC	Noise, augmentation	DFDC test AUC = 0.682 FF++-test AUC = 0.732	---
2021	Bilateral filtering (passive) Joint adversarial (proactive) [138]	XceptionNet	FF++	Accuracy	BIM, PGD, and FGSM	Acc = 90.79	Highly restrictive attack.
2024	Masked Depthfake Network [139]	MDN	FF++	Accuracy	Gaussian blur, Gaussian noise, rescaling, translation	AUC= 98.50	Restrictive to black-box attacks only.

10.1 Proactive defense techniques against audio adversarial attacks

10.1.1 Adversarial training-based defense

In the adversarial training defensive technique, gradient-based perturbed voice samples were added during model training. Wang et al. [140] introduced regularization-based adversarial training and produced adversarial examples through FGSM and local distributional smoothing methods. The experimentation was performed on the TIMIT dataset on the GE2E-ASV system. This technique slightly reduces the ERR rate of adversarial attacks, but the overall results are not satisfactory. Wu et al. [141] introduced a defense against PGD and FGSM attacks to greatly boost the performance of VGG and SENet on the ASVspoof 2019 datasets. To quantify and analyze the efficacy of deep models against adversarial noise, they generate high-level representations of voice samples retrieved by the self-supervised model and a layer-wise noise-to-signal ratio (LNSR), which provides up to 90% defense against white and black box attacks as well. Pal et al. [142] proposed hybrid adversarial training based on cross-entropy and marginal loss to withstand black-box attacks and have greater adversarial robustness. The experiments were performed on 1D-CNN on several white and black box scenarios and attained good defense, but the defense is limited to one tested model. According to [143], adversarial attacks on ASV systems can reduce their accuracy, and various attacks like Gaussian noise make the detection task more sophisticated and challenging. To lessen the impact of an adversarial attack, [143] used adversarial training (adversarial samples are incorporated into the training dataset) and adversarial Lipschitz regularization (ALR). ALR relies on a function that is designed to overlook incremental changes to the input. But this defense technique [143] is only limited to PGD, FGSM, and C&W attacks with 73% accuracy on the LibriSpeech dataset. Isik et al. [144] proposed neuromorphic audio processing, which examines security vulnerabilities in neuromorphic audio, specifically addressing adversarial techniques such as FGSM and PGD. A system with an FPGA has a 94% detection rate, excellent spike encoding, and a balanced signal-to-noise ratio. Technology outperforms current methods with a 5.39 dB signal-to-noise ratio.

10.1.2 Refactoring-based defense

For data purification against adversarial attacks, a pre-processing module can be used to denoise or rearrange the data for cleansing. This will help to achieve the purpose of defense and minimize the likelihood of an attack happening. Denoising and noise addition are two components of the larger process known as input refactoring, which refers to the collection of algorithms that preprocess the input data on the input layer. Wu et al. [145] proposed spatial smoothing and adversarial training based on the Mockingjay technique [146] on LCNN and SENet detectors. Since different attacks leave distinctive patterns on the sample, this approach [145] has only been tested on PGD attacks, it may have a limited defensive capacity due to the fact that adversarial training is less effective against novel attacks. The Transformer encoder representations from the alteration (TERA) model were proposed in [147] that employed self-supervised learning techniques to predict sequences of future frames based on the input. This research was heavily influenced by the Mockingjay defense that used self-supervised learning by processing audio samples with Gaussian, mean, and median filters. This technique is useful since it cleans up the samples without requiring any prior knowledge of the adversarial sample generation process and achieves an EER rate of 22.49%.

10.2 Passive defense techniques against audio adversarial attacks

The following section discusses passive defensive methods against adversarial attacks on audio deepfakes. Wu et al. [141] proposed a passive defense known as spatial smoothing through the use of median or mean filters. This method does not modify the attack model, but it enables it to detect adversarial examples without the need for additional training. Assuming that the adversary is unaware of the implementation of spatial smoothing on the input data before it is fed into the model, however, this intentionally created perturbation will be neutralized by the spatial smoothing, rendering the adversarial attack ineffective. Li et al. [148] constructed a defensive network, a VGG-like binary classification detector that discriminates between adversarial samples and authentic ones in the convolutional layer and pools the speech sequences for decision-making in the pooling layer. The binary classification detector exhibited good adversarial robustness on cross-model SRSs but decreased recognition accuracy in the face of various attacks (the model trained on adversarial examples generated by the BIM algorithm could detect 99.83% of BIM attacks but only 48.61% of JSMA attacks). Wu et al. [149] introduced self-supervised learning models (SSLMs) to filter out noise at the surface level in the inputs and reconstruct clean samples from interrupted ones. SSLM consists of two modules: one removes inconsistencies from audio, and the other compares the unaltered sample to the one that has been cleaned up to identify potentially harmful recordings. This method achieved an R-Vector of 14.59% GenEER and an X-Vector of 11.29% GenEER against JSMA, FGSM, and BIM attacks.

Wu et al. [150] proposed neural coders to re-synthesize audio and discover discrepancies between the original sample and the re-synthesized audio samples against a BIM (PGD type) adversarial attack. Using this technique [150], adversarial noise was removed, and the authentic waveform was regenerated with less distortion. It is helpful because

it does not require knowledge of the attack algorithm. To add insult to injury, only a single neural vocoder implementation was employed. This method achieved the highest AUC of 99.94% against a BIM attack. Yoon et al.[151] introduced a new deepfake detection method, TMI-Former, which integrates features across visual, auditory, and linguistic modalities. This approach addresses challenges in sparse data environments and aims to overcome the limitations of traditional AI-based methods. TMI-Former is structured into four stages: vision feature extraction, representation, residual connections, and late-level fusion. It performs effectively in scenarios with limited identity data and single-instance deepfake examples. The model achieves accuracy rates of 18.75% to 19.5% and F1-scores from 0.2238 to 0.3561, surpassing previous multi-modal AI systems. This performance demonstrates the potential of complex, cross-modal interactions in enhancing deepfake detection systems, especially in restricted data environments. Villalba et al. [152] used representational learning to classify adversarial attacks and identify distinct attacks. Probabilistic linear discriminant analysis (PLDA) was implemented for the x-vector system and achieved a recognition accuracy of 71.8% for the classification of attacks within the training set and an error rate of 19.6% for identifying unknown attacks. Joshi et al. [153] trained the representation learning network with adversarial perturbations rather than adversarial training and used the time-domain denoiser to estimate the adversarial perturbations. This allowed us to train the network without the need for adversarial examples. In contrast to other methods, this method [153] can identify previously unknown attacks, although its detection accuracy is still poor. Uddin et al. [154] introduced a novel collaborative learning framework that is designed to identify anti-forensic manipulation generated by GANs. The authors proposed a system that employs multiple detectors to identify complex GAN-based forgeries, thereby improving the accuracy and robustness of anti-forensic detection. However, the method is limited to a substantial amount of training data, potential scalability concerns, and the expectation that all detectors are equally effective and reliable. Uddin et al. [155] proposed a robust defense framework against GAN-based anti-forensic attacks on audio deepfake detectors. Their work also presented a comprehensive empirical analysis [156] of twelve state-of-the-art audio deepfake detection methods, evaluated across five benchmark datasets under four statistical and four optimization-based anti-forensic attack scenarios.

Table 10. An overview of audio defense techniques.

Ref.	Defensive Methods	Model	Dataset	Baseline measure	Results		Limitations
					Attack	Defense	
Proactive Defense							
Adversarial Training-Based Defense							
2019	Regularization [140]	GE2E-ASV	TIMIT	ERR = 4.87	FGSM = 11.89 LDS = 9.26	FGSM = 8.31 LDS = 4.60	Slightly reduce ERR. Not robust.
2020	Adversarial training PGD [141]	VGG, SENet	ASVspoof 2019	Acc = 99.99	PGD = 37.06	Acc = 98.60	Limited to defend against PGD and FGSM
2021	Hybrid adversarial training [142]	1D-CNN	LibriSpeech	Acc = 99.95	FGSM = 6.03, PGD = 0, C&W = 0	FGSM = 90.60, PGD = 81.12, C&W = 80.12	Limited to defend one model.
2021	Adversarial Lipschitz Regularization [143]	1DCNN, TDNN	LibriSpeech	Acc=96	PGD = 43.0, FGSM = 73.0, C&W = 58.0	Acc = 93.0	Limited to defend against few attacks
2024	NEUROSEC [144]	RNN, CNN, DNN	DDR4 SDRAM	-	PGD, FGSM	Detection rate = 94	Limited to defend against few attacks
Refactoring Based Defense							
2020	Spatial smoothing and adversarial training [145]	LCNN, SENet	ASVspoof 2019	Acc= up to 90	PGD, FGSM (5-10)	Acc= (80-90)	Limited to defend against PGD attack
2021	TERA [147]	r-vector	Voxceleb1	Err rate = 8.87	BIM = 66.02	22.94	Tested with BIM attack
Passive Defense							
2020	Median and Mean filter [141]	VGG, SENet	ASVspoof 2019	ACC = 99.97	PGD = 37.06	Acc= 99.76	Limited to defend against PGD and FGSM
2020	Detection network [148]	VGG-like	Voxceleb1	-	BIM-even	ERR = 0.46	Performance degrades on other attacks.

2021	SSLM [149]	x-vector, r-vector	Voxceleb1	AdvFAR = 5.97	BIM = 87.36	16.54	--
2021	Re-synthesis [150]	Vocoder	Voxceleb1 and Voxceleb2	-	BIM (FPR=0.01)	Acc= 98.92	Limited to one attack only.
2021	Representation learning [152]	Espresso	Voxceleb1 and Voxceleb2	-	C&W- L_2	Acc= 82.9	Performance degrades on other attacks.
2022	Representation learning [153]	AdvEst	Voxceleb2	-	FGSM, PGD, CW	Acc= 96 EER= 9	
2025	GAN-based attack [155]	AI-generated audio	RawNet3, RawNet2, TSSDNet, ResNet, MS-ResNet2	Avg. Acc=95.5 Avg. Acc = 59.7	---	----	
2025	Statistical, Optimization, and GAN-based AF attacks [156]	AI-generated audio	12 SoTA deepfake detectors	AUC= 75 (raw), AUC= 86 (spectrogram)AUC = 57.6 (statistical), AUC = 38.5 (Optimization)	MSE, SSIM		Less methods explored.

*(Acc=Accuracy, ASR=Attack Success Rate, EER= Equal Error Rate)

10.3 Analysis and discussion

This section presents the analysis and discussion of defensive approaches to audio deepfake methods, which cover both proactive and passive defense in audio. Proactive defense entails taking measures to prevent adversarial attacks from ever occurring. This involves adversarial training [140-143], in which the model is trained on adversarial examples to improve its robustness. These techniques include known adversarial examples during adversarial training [141] or hybrid adversarial instances [141]. In addition, the use of a regularization [140, 143] also provides detection models to better protect them from adversarial attacks. Other proactive defense strategies include refactoring and data preprocessing to eliminate hostile perturbations. It includes a feature-based pre-processing [145-147], but in comparison, adversarial training provides better defense than refactoring-based techniques. In contrast, passive defense entails detecting and mitigating adversarial attacks after they have been launched. This may involve employing anomaly detection or post-processing techniques to determine and eliminate adversarial perturbations from the inputs [141, 148-150, 152, 153]. Both proactive and passive defenses have benefits and drawbacks. Proactive defense has the benefit of being able to prevent adversarial attacks, but its implementation may be time- and resource-intensive. Passive defense, on the other hand, can be more effective and less resource-intensive, but it does not prevent adversarial attacks and may only mitigate their effects. In general, the choice between proactive and passive defense is determined by the needs and resources of the system. Passive defense may be more suitable for systems with limited resources and low overhead, whereas proactive defense may be more suitable for systems requiring high security and robustness.

11 Evaluation and perceptual similarity measures

This section describes the perceptual similarity measures used in the reviewed papers to quantify the similarity between perturbed and original images. Most of the reviewed papers utilized the L-Norm (L_0, L_2, L_∞) distance metrics to measure the extent to which the adversarial example is different from the original image. However, a few papers also used other metrics, including distortion score, perceptual image patch similarity (LPIPS), Fréchet Inception Distance (FID), etc.

11.1 Attack success rate (ASR)

The attack success rate [157] is a metric used to assess an adversarial attack's effectiveness as the percentage of times the attack successfully changes the model's prediction. This metric is a useful tool for providing insight into the performance of adversarial attacks, as well as potential measures that can be taken to mitigate the threat of such an attack. ASR can be computed as:

$$ASR = \frac{(Number\ of\ Adversarial\ Examples\ Correctly\ Classified)}{(Total\ Number\ of\ Adversarial\ Examples)} \quad (19)$$

11.2 Equal error rate (EER)

The EER [158] is the intersection of the false acceptance rate (FAR) and the false rejection rate (FRR) on the ROC curve. The EER quantifies the compromise between the FAR and FRR. The greater the performance of the biometric

system, the lower the EER. The EER can be calculated as the value at which the FAR and FRR are equal or by interpolating between the two closest points on the ROC curve.

$$EER = \frac{(FAR+FRR)}{2} \quad (20)$$

11.3 L-Norm

L-norm is used in adversarial learning to measure the dissimilarity between the original and adversarial samples as well as to evaluate the robustness of a model [159]. Choosing an appropriate L-norm can effectively minimize the perturbation while still producing an incorrect prediction. The L-norm is easy to optimize and can be used to limit the size of adversarial perturbations to make them more realistic and less likely to be found. The formula for the L-norm in adversarial learning is:

$$\text{minimize } ||x' - x||p, s.t. f(x') != y \quad (21)$$

Where x is the original input, x' is the adversarial input, $f(x)$ is the model's prediction for input x, y is the target label for the adversarial example, and p is the order of the L-norm. This formula is the mathematical expression of the trade-off between model accuracy and the degree to which an adversarial example can be generated from a given input. This equation tries to find the smallest change (as measured by the L-norm) that can be made to the original input x to make an adversarial input x'that is different from the target label y.

11.4 Similarity score

The similarity score is determined by calculating the difference between the original and adversarial inputs. In other words, a higher similarity score implies that the adversarial attack has been successful in fooling the model, and a lower score indicates that the model has been robust against the attack [160]. The equation for the similarity score can take many forms, such as L-norm distance, cosine similarity, etc., depending on the specific task and the type of input. Consequently, choosing the appropriate similarity score for each task and input type is important. The L-norm distance between the original input and the adversarial input is often used as a basis for the similarity score equation and can be computed as:

$$Similarity\ Score = \frac{1 - ||x\prime - x||^2}{||x||^2} \quad (22)$$

where x is the true input, x' is the adversarial input, and L_2 (x) is the norm of x. Another example is the cosine similarity, which is calculated as:

$$Similarity\ Score = \frac{(x\prime * x)}{(||x|| * ||x\prime||)} \quad (23)$$

where x and x' are inputs and L_2 norms of the original and adversarial inputs, respectively; * is the dot product; and $||x||$ and $||x'||$ are the L_2 norms of x and x'.

11.5 Distortion score

The distortion score in adversarial machine learning is a measure of how much the original input has been changed to create the adversarial input [161]. It is used to evaluate the success of an adversarial attack and to measure the robustness of a model against adversarial examples. Typically, the distortion score is calculated as a scalar value, with higher values indicating more distortion and lower values indicating less. One common equation for the distortion score is based on the L_2-norm distortion score, which is calculated as:

$$Distortion\ Score = \frac{||x\prime - x||^2}{||x||^2} \quad (24)$$

where x is the original input, x' is the adversarial input, and $||x||^2$ is the L_2-norm of the original input. Another example is the L_1-norm distortion score, which is calculated as:

$$Distortion\ Score = \frac{||x\prime - x||}{||x||} \quad (25)$$

where x is the original input, x' is the adversarial input, and $||x||^1$ is the L_1-norm of the original input.

11.6 Fréchet inception distance (FID)

Fréchet Inception Distance (FID) is a metric used to measure the similarity between two sets of images in adversarial machine learning [162]. It is based on the idea that a pre-trained network can be used to compare the feature representations of the images in the two sets. The Fréchet distance is a measure of the similarity between two

multivariate Gaussian distributions. FID only compares the feature representations of the images and does not account for the content or semantic information of the images.

$$FID = \left|\left|x_1 - x_2\right|\right|_2^2 + Tr((C_1C_2) - 2(C_1C_2)^{\frac{1}{2}}) \quad (26)$$

11.7 Mean squared error

Mean Squared Error (MSE) is a metric used to measure the difference between original images and adversarial images [163]. This loss function is frequently employed in regression analyses. The MSE can be calculated as:

$$(MSE) = \frac{1}{n} \sum_{i-1}^{n} (\mathrm{x}' - \mathrm{x})^2 \quad (27)$$

The square of the difference between the original and adversarial data is denoted by $(\mathrm{x}' - \mathrm{x})^2$, where *N* is the total number of data points and *x* and *x′* are the original and adversarial data, respectively.

11.8 Peak signal-to-noise ratio

In adversarial machine learning, the peak signal-to-noise ratio (PSNR) is a method to compare the quality of a reconstructed image to the quality of the original image [164]. The PSNR is determined by dividing the maximum possible signal power by the maximum possible noise power. PSNR can be computed as:

$$Average\ PSNR = 20 * log10(MAX) - 10 * log10\ (MSE) \quad (28)$$

Where MAX is the pixel's maximum value, MSE is the mean squared error between the original and adversarial image, and log10 is the base 10 logarithm.

11.9 Structural similarity index measures

In adversarial machine learning, the Structural Similarity Index measures (SSIM) is the similarity between two images in terms of structural information and luminance, contrast, and structure [165]. It is a perception-based method that considers image degradation as a perceived change in structural information, as opposed to pixel value changes. SSIM can be computed as follows.

$$SSIM = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(2\mu_x^2 + \mu_y^2 + C1)(2\sigma_x^2 + \sigma_y^2 + C_2)} \quad (29)$$

SSIM formula computes brightness, contrast, and structural similarities by using the means (μ), standard deviations (σ), and covariance (σ_{xy}) of two images x and y. To avoid division by zero, constants (C) values are also added.

11.10 Normalized mean error

Normalized Mean Error (NME) is a metric used to measure the dissimilarity between two images, it compares the normalized pixel values of the two images and calculates the mean error between them [166]. The NME is calculated as the average of the absolute differences between the pixel values of the two images divided by the maximum value of the pixel as shown in Eq. (26).

$$(NME) = \frac{1}{n} \sum_{i-1}^{n} \frac{||\mathrm{x}\prime - \mathrm{x}||^2}{d} \quad (30)$$

NME compares the pixel values of two images by determining the sum of absolute differences divided by the highest pixel value (d). NME values range from 0 to 1, with 0 representing perfect similarity and 1 representing total dissimilarity.

12 Open issues, future direction, and recommendations

Adversarial attacks, anti-forensics using GAN based approaches and defenses play a crucial role in the endless war between deepfake generation and detection. While considerable progress has been made in both the generation and detection of deepfakes, there are still limitations and challenges that must be overcome. Deepfakes, for instance, may be generated with high precision, especially given access to a massive amount of diverse training data, but even a small watermark may disrupt the generative process. Additionally, deepfake detection methods can be deceived by small perturbations or changes in parameters. The employment of adversarial attacks can easily fool deepfakes generators and detectors. In this section, we discuss the limitations of existing research and recommendations for continued efforts.

12.1 Robustness of detectors against combined adversarial and anti-forensics attacks

As technology advances, adversarial attacks pose a significant threat to the security and integrity of deepfake generators and detectors. To mitigate this threat, researchers introduce anti-forensics and adversarial attacks to their

systems as penetration testing to enhance the generalizability of these systems. However, existing adversarial attacks have various limitations, i.e., transferability, high dependency on model knowledge, limited multi-modal capability, high computational cost, limited realism and interceptability, domain dependency, lack of defense awareness, temporal inconsistency in video, vulnerability to detection, lack of adaptability to GAN-based generation, etc. Some of the limitations [67] and potential future directions of adversarial attacks on deepfake detectors and generators include adversarial training, GANs, and transfer learning. The detectors need to be robust against the combined effect of anti-forensics and adversarial attacks, and multi-order anti-forensics attacks.

12.1.1 Evade generalization of GANs

GANs and VAEs are powerful tools to generate deepfakes, however, adversarial attacks can be launched to undermine their generalizability by changing the parameters of the generator or discriminator module, especially in white-box scenarios. White box attacks can alter the generalizability of these models and lead to a generation of disrupted deepfakes. Additionally, adversarial attacks on GANs [83] are susceptible to adaptability and transferability problems. Attacks designed for one specific GAN architecture often fail when applied to a different type of GAN, limiting their broader applicability. This architectural dependency reduces the real-world effectiveness of such attacks, especially in diverse environments where multiple generative models may be deployed. Therefore, a key future direction in adversarial research is the development of transferable attack strategies that can degrade the performance of a wide range of GANs regardless of their specific design. Such advancements could be instrumental in building more resilient detection systems, as understanding how different models fail under pressure is critical to anticipating and mitigating their weaknesses. Ultimately, improving the adaptability and reach of adversarial techniques not only tests the robustness of current generative models but also guides the creation of more secure and trustworthy deepfake generation and detection frameworks.

12.1.2 Watermarks and tags

Watermarks and tags [10, 11, 167] on input samples can also cause the deepfake generators to create undesirable output. One possible direction can be the use of imperceptible watermarks, nearly invisible markers that are embedded in visual or audio data in a way that is undetectable to human perception but can disrupt deepfake generation models. These watermarks are particularly valuable because they are difficult for the generator to detect and remove, making it significantly harder for the model to produce convincing forgeries. It would be more difficult for the deepfake generator to avoid the watermark and produce realistic outcomes. As a result, imperceptible watermarks provide a powerful instrument in the continuing battle against deepfake technology. Unlike traditional watermarks, imperceptible versions are designed to blend seamlessly with the input data while still affecting the internal workings of generative models. Their subtle interference can degrade the quality of synthesized content, making it easier to flag manipulated media. This approach not only hinders the generator's ability to create realistic outputs but also offers a more robust layer of defense in the arms race between deepfake generation and detection. Furthermore, the integration of imperceptible watermarks could enhance existing detection mechanisms by acting as passive markers for identifying synthetic media. Their presence can support forensic analysis and contribute to the development of more trustworthy media authentication systems. Overall, imperceptible watermarks represent a strategic, low-cost, and scalable method for disrupting deepfake generation while reinforcing the integrity of digital content. They hold great potential as a defensive tool in both research and real-world applications aimed at combating synthetic media threats.

12.1.3 Adversarial training as an attack

Adversarial training as an attack has shown promising results, but it has some limitations, such as the possibility of overfitting, in which the model becomes extremely specific to the training data and cannot generalize effectively to new data. This inability to transfer adversarial examples generated during training [141] to attack other models reduces its impact. The adversarial examples generated through training often lack transferability, they may succeed against the model they were trained on but fail to deceive other architectures, reducing their practical effectiveness in black-box or real-world settings [35]. Additionally, adversarial training is computationally expensive. It requires iterative optimization processes like Projected Gradient Descent (PGD), significantly increasing training time and resource demands [168] . Despite these challenges, adversarial training remains a viable direction for creating more potent attack strategies. Future research could focus on developing unsupervised adversarial training methods that do not rely on labeled data, potentially expanding the attack surface and applicability across domains [169].

Furthermore, hybrid techniques that combine adversarial training with generative models such as GANs or diffusion models may offer new pathways for crafting robust and highly transferable attacks [170].This underscores the importance of continued exploration into adversarial training as both a threat vector and a benchmarking tool for robustness.

12.1.4 Model-specific attacks

Model-based attacks can be created to exploit weaknesses in the architecture and settings of the deepfake detection model. These attacks can evade the model's security and cause it to inaccurately classify input samples. In a model-based attack, the attacker constructs a surrogate model [171] using the target system and then produces input data intended to reclassify or attack the target system; however, these models are not transferable. The development of transferable surrogate models is a potential future direction in adversarial attacks. In addition to model transfer, key areas of emphasis include addressing domain adaptation challenges and improving attack optimization efficiency.

12.1.5 Black box attack

Black box attacks [99] test a model's robustness against adversarial cases or generate novel attacks for real-world scenarios. In these attacks, the attacker typically has no information about the target model other than its output in a query-based setup. Several current techniques calculate attack transfer rates on pre-trained models to generate perturbations in those models. This suggests that they possess full awareness of the target model's training data and thus violates the definition of a black box setup, which requires that the target model be trained on unobserved data and have an unknown number of output labels. True black box testing remains an unexplored area.

12.1.6 Multi-modal attack

Adversarial attacks on multimodal systems, process and integrate information from more than one modality, pose unique challenges and risks amplifies [172]. Although they provide enhanced performance by leveraging complementary signals, their multimodal nature inherits vulnerabilities of all the modalities, and ultimately, the adversarial threat.

Future multi-modal adversarial attacks [173] against multi-model systems will become increasingly frequent and complex, and attackers will be capable of exploiting flaws in existing models to deceive those systems. For multi-model (audio-visual) detectors, perturbation can be added to either audio or video to fail the detection capability. As a result, the multi-model system will become less dependable and trustworthy, and organizations will need to be more cautious to ensure their systems are adequately protected against adversarial threats.

12.2 Defense

A feasible way to protect deepfake generators and detectors from adversarial attacks is to develop more generalizable and resilient defensive methods. The future of defense against adversarial attacks on deepfake generators and detectors is likely to involve a combination of multiple approaches. Some potential strategies include adversarial training, ensemble methods, which are discussed in this section.

12.2.1 Adversarial training as defense

Adversarial training has emerged as a defense mechanism against adversarial attacks. In this method, the model is trained on both clean and adversarial instances to increase its resistance to future attacks. The following are potential future approaches for adversarial training [140-143] as a defense, as the field continues to progress. As adversarial attacks become increasingly complex, the efficacy of present adversarial training approaches will decline. Additionally, Bai et al. [174] highlighted the issue of generalization in adversarial trained systems on both unperturbed and perturbed test data scenarios, and on unseen attacks as well. Thus, future research can focus on developing more sophisticated adversarial training systems that can protect against a broader range of attacks with improved generalizability. Adversarial training has shown potential, but its efficacy in real-world conditions remains undetermined. Future studies should evaluate the efficacy of adversarial training in practical applications.

12.2.2 Deep reinforcement learning and game-theoretic approaches

Deep reinforcement learning can be used to defend against adversarial threats by developing a model that can withstand such malicious inputs. In [175], an adversarial attacks resistant reinforcement learning framework, coupled with Xception and Inception-ResNetv2, was proposed to counter deepfakes. In addition, game theory [16] can be used as a strategy that utilizes rewards or penalties as feedback signals to learn the optimal response to adversarial attacks. This strategy can be used to effectively recognize disrupted inputs and defend the system from harm or disruption caused by attackers. Another direction is the creation of more interpretable and explicable deep RL and adversarial ML models, which is another key direction. Interpretable models can provide insights into how they make decisions and enhance trustworthiness and credibility. Additionally, explainable models can aid in identifying and mitigating potential weaknesses in adversarial attacks.

12.2.3 Proactive defense

A proactive defense technique [135] can also be introduced before training models to detect adversarial examples, preventing the production of perturbed instances. Existing works have employed proactive defense methods using identity watermarks [83, 176, 177]. One potential future direction of proactive defense can be the use of explainable

AI to address this issue by revealing how models make decisions and highlighting flaws that attackers can exploit. This can be useful to protect the system from any attack. Another future direction can be the use of hybrid strategies, which include the implementation of various defense mechanisms, including feature fusion, decision fusion, and adversarial training, to protect against adversarial attacks better than using individual defensive mechanisms. More advanced strategies for integrating and optimizing these mechanisms may be developed in the future.

12.2.4 Decoy mechanism as defense

Deepfake detector attack resilience can be increased through the use of decoy techniques [17]. Adversarial training, defensive distillation [18], gradient masking, input transformation, ensemble approaches, and feature squeezing can all be used as decoy mechanisms. Using multiple deepfake detectors in parallel and combining their outputs as an ensemble method [105] can be used to enhance the system's accuracy and robustness. Another direction can be an exploration of feature fusion and decision fusion, as these methods can be explored as an effective approach to defending against adversarial attacks, as they mix several models with diverse architectures and parameters. This builds more robust protection than any single model can achieve by making it difficult for adversaries to identify the same flaws in all of them.

12.2.5 Attack-resistant automatic speaker verification and face recognition systems

Several end-to-end and unified countermeasures have been developed against audio [141] and visual deepfakes [127], to protect automatic speaker verification and face recognition systems. However, these countermeasures protect against specific types of attacks and are susceptible to other adversarial attacks [166-168]. There is a need to develop a unified attack-resistant system that not only protects against audio-visual deepfakes but is also resistant to adversarial attacks. Future research should focus on the development of more effective countermeasures to enhance the robustness of anti-spoofing systems, face recognition systems, and deepfake detectors against adversarial attacks.

12.2.6 Fusion-Based Defensive Techniques

Fusion-Based defensive approaches are emerging as a promising strategy for improving the robustness and generalizability of deepfake detectors. Instead of relying on a single model or modality, these frameworks defend against adversarial attacks by combining complementary spatial, temporal, and frequency cues or fusing output from several networks. Recent studies [54, 155] show that generative models as defense, and multi-model fusion and knowledge distillation can improve resilience to unseen manipulations. Future research should look into hybrid fusion of visual, aural, and semantic modalities to develop adaptive and trustworthy deepfake defense systems.

12.2.7 Spatiotemporal Defense Techniques

Spatiotemporal and holistic deepfake defense methods [178] have emerged as highly effective strategies for detecting both traditional 2D GAN-generated forgeries and advanced 3D neural-rendered deepfakes. Unlike spatial-only detectors that analyze individual frames, these approaches integrate complementary spatial, temporal, and joint spatiotemporal features to capture subtle inconsistencies in motion, dynamics, and facial geometry that generative models struggle to reproduce consistently. Recent works [178-181] demonstrate that combining embeddings from multiple representation dimensions significantly improves robustness, generalization, and resistance to unseen manipulations. Transformer-based fusion architectures further enhance detection accuracy while remaining lightweight and computationally efficient. Future research should explore holistic multi-modal fusion including visual, motion, audio, and semantic signals to build adaptive, scalable, and resilient deepfake defense frameworks capable of countering rapidly evolving generative models.

12.2.8 Testbed against adversarial attack

It is necessary to develop a testbed for gauging the robustness of deepfake models against adversarial attacks. This can be a powerful resource for analyzing the integrity of detectors against adversarial attacks, but so far there is no testbed or penetration testing tool available to even individuals, let alone multiple types of adversarial attacks. By replicating actual attack scenarios, penetration tests can reveal system vulnerabilities that can be exploited by malicious actors. This knowledge can then be utilized to strengthen existing defenses or develop new ones to protect against these threats more effectively. Additionally, penetration testing helps developers become more aware of the possible threats posed by attackers and allows them to rehearse response processes before actual crises.

13 Conclusion

This paper has presented a detailed review of existing adversarial attacks and defenses on audio-visual deepfake generation and detection approaches, analyzing their strengths and shortcomings. The field of deepfake creation and detection is in a continuous process of evolution, with new methodologies and strategies being created to address the

most recent deepfake-related challenges, but adversarial attacks on these methods hinders the trustworthiness of these methods. In addition, the development of effective and robust deepfake detection methods will continue to be a crucial area of research. Lastly, a comprehensive study on the limitations and potential for future research in this field is also presented. Addressing these challenges can only increase the efficacy of deepfake generative and detection models.

This thorough research has highlighted the rapid advancement and complex issues of adversarial attacks and defences in deepfake technology. The domain continues in its expansion, driven by the uncompromising momentum of innovation and the essential need for reliable detection techniques. We anticipate that the insights and criticisms included in this study will enhance existing research and stimulate the interest of emerging researchers in this intriguing field. We aspire that the diverse array of options presented will stimulate an additional influx of research that expands the limits of deepfake generation and detection, enhancing the robustness and dependability of future systems.

CRediT authorship contribution statement

Qurat Ul Ain, Hafsa Ilyas and Fatima Khalid: Writing the original manuscript, Validation, Software, and Methodology. Khalid Mahmood Malik and Aun Irtaza: Validation, Methodology, and Writing – review and editing. Ali Javed and Khan Muhammad: Validation, Writing – review and editing, Supervision and Project Administration.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Science Foundation, USA, award # 2409577, in part by the Punjab Higher Education Commission of Pakistan under Award No. (PHEC/ARA/PIRCA/20527/21), and by the "Regional Innovation System & Education (RISE)" through the Seoul RISE Center, funded by the Ministry of Education (MOE) and the Seoul Metropolitan Government (2025-RISE-01-018-04). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the PHEC ACT.

References

[1] Roe, J., & Perkins, M. (2024) Deepfakes and Higher Education: A Research Agenda and Scoping Review of Synthetic Media. *arXiv preprint arXiv:2404.15601*.

[2] Masood M, Nawaz M, Malik KM, Javed A, Irtaza A, Malik H, (2023) Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Appl Intel 53(4):3974-4026. https://doi.org/10.1007/s10489-022-03766-z

[3] Khan A, Malik KM, Ryan J, Saravanan M, (2022) Voice Spoofing Countermeasures: Taxonomy, State-of-the-art, experimental analysis of generalizability, open challenges, and the way forward. arXiv preprint arXiv:2210.00417. https://doi.org/10.48550/arXiv.2210.00417

[4] Abdelkader, S., Amissah, J., Kinga, S., Mugerwa, G., Emmanuel, E., Mansour, D. E. A., ... & Prokop, L. (2024). Securing modern power systems: Implementing comprehensive strategies to enhance resilience and reliability against cyber-attacks. *Results in engineering*, 102647.

[5] Javed A, Malik KM, (2022) Faceswap Deepfakes Detection using Novel Multi-directional Hexadecimal Feature Descriptor. In: 2022 19th International Bhurban Conference on Applied Sciences and Technology (IBCAST). pp. 273-278

[6] Khalid, F., Javed, A., Ilyas, H., & Irtaza, A. (2023) DFGNN: An interpretable and generalized graph neural network for deepfakes detection. *Expert Systems with Applications*, *222*, 119843.

[7] Narayan, K., Agarwal, H., Thakral, K., Mittal, S., Vatsa, M., & Singh, R. (2022). Deephy: On deepfake phylogeny. In *2022 IEEE International Joint Conference on Biometrics (IJCB)* (pp. 1-10). IEEE.

[8] Javed, A., Malik, K. M., Irtaza, A., & Malik, H. (2021). Towards protecting cyber-physical and IoT systems from single-and multi-order voice spoofing attacks. *Applied Acoustics*, *183*, 108283.

[9] Huang Y, Juefei-Xu F et al (2009) Dodging DeepFake detection via implicit spatial-domain notch filtering. arXiv preprint arXiv:2009.09213.

[10] Wang R, Juefei-Xu F et al (2020) Deeptag: Robust image tagging for deepfake provenance. arXiv preprint arXiv:2009.09869.

[11] Wang R, Juefei-Xu F, Luo M, Liu Y, Wang L (2021) Faketagger: Robust safeguards against deepfake dissemination via provenance tracking. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 3546-3555. https://doi.org/10.1145/3474085.3475518

[12] Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology*, *159*, 107197.

[13] Haider, J. (2023). Deconstructing Deepfakes: Ethical Implications and Mitigating Strategies in a Post-Truth World. *Journal of Media Horizons, ISSN (E) 2710-4060 (P) 2710-4052*, *4*(4), 1-14.

[14] Ethics Guidelines for Trustworthy. Available: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html

[15] What is trustworthy AI?. Available: https://www.ibm.com/think/topics/trustworthy-ai#:~:text=Trustworthy%20AI%20refers%20to%20artificial,among%20stakeholders%20and%20end%20users

[16] Wang Z, Guo Y, Zuo W (2022) Deepfake forensics via an adversarial game. IEEE Trans on Image Proces 31:3541-52

[17] Chen GL, Hsu CC (2023) Jointly Defending DeepFake Manipulation and Adversarial Attack using Decoy Mechanism. IEEE Tran on Patt Anal and Mach Inteli

[18] Papernot, N., & McDaniel, P. (2016). On the effectiveness of defensive distillation. *arXiv preprint arXiv:1607.05113*.

[19] Liz-Lopez, H., Keita, M., Taleb-Ahmed, A., Hadid, A., Huertas-Tato, J., & Camacho, D. (2024). Generation and detection of manipulated multimodal audiovisual content: Advances, trends and open challenges. *Information Fusion*, *103*, 102103.

[20] Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE access*, *10*, 25494-25513.

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

[22] Croitoru, F. A., Hiji, A. I., Hondru, V., Ristea, N. C., Irofti, P., Popescu, M., ... & Shah, M. (2024). Deepfake Media Generation and Detection in the Generative AI Era: A Survey and Outlook. *arXiv preprint arXiv:2411.19537*.

[23] Li, M., Ahmadiadli, Y., & Zhang, X. P. (2025). A Survey on Speech Deepfake Detection. *ACM Computing Surveys*.

[24] Pham, L., Lam, P., Nguyen, T., Tang, H., Tran, D., Schindler, A., ... & Vu, C. (2024). A Comprehensive Survey with Critical Analysis for Deepfake Speech Detection. *arXiv preprint arXiv:2409.15180*.

[25] Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *14*(2), e1520.

[26] Gambín, Á. F., Yazidi, A., Vasilakos, A., Haugerud, H., & Djenouri, Y. (2024). Deepfakes: Current and future trends. *Artificial Intelligence Review*, *57*(3), 64.

[27] *Google Trends*. Available: https://trends.google.com/trends/explore?date=2018-01-01%202024-11-27&q=deepfake&hl=en.

[28] Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572. https://doi.org/10.48550/arXiv.1412.6572

[29] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2017) Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083. https://doi.org/10.48550/arXiv.1706.06083

[30] Li, D., Wang, W., Fan, H., & Dong, J. (2021). Exploring adversarial fake images on face manifold. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5789-5798).

[31] Ivanovska, M., & Struc, V. (2024). On the vulnerability of deepfake detectors to attacks generated by denoising diffusion models. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 1051-1060).

[32] Szegedy C, Zaremba W et al (2013) Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199. https://doi.org/10.48550/arXiv.1312.6199

[33] Kurakin A, Goodfellow I, Bengio S (2016) Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236. https://doi.org/10.48550/arXiv.1611.01236

[34] Kurakin A, Goodfellow IJ, Bengio S (2018) Adversarial examples in the physical world. In: Artificial intelligence safety and security, pp. 99-112

[35] Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A (2016) The limitations of deep learning in adversarial settings. In: 2016 IEEE European symposium on security and privacy (EuroS&P), pp. 372-387

[36] Carlini N, Wagner D (2017) Towards evaluating the robustness of neural networks. In: 2017 IEEE symposium on security and privacy, pp. 39-57

[37] Moosavi-Dezfooli SM, Fawzi A, Fawzi O, Frossard P (2017) Universal adversarial perturbations. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1765-1773

[38] Moosavi-Dezfooli SM, Fawzi A, Frossard P (2016) Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2574-2582

[39] Chen PY, Zhang H, Sharma Y, Yi J, Hsieh CJ (2017) Zoo: Zeroth order optimization based black box attacks to deep neural networks without training substitute models. In: Proceedings of the 10th ACM workshop on artificial intelligence and security, pp. 15-26. https://doi.org/10.1145/3128572.3140448

[40] Su J, Vargas DV, Sakurai K (2019) One pixel attack for fooling deep neural networks. IEEE Tran on Evolu Comp 4;23(5):828-41. doi: 10.1109/TEVC.2019.2890858

[41] Wu S (2021) Universal Watermark Attack in Image Classification. In: 2021 International Conference on Intelligent Computing, Automation and Systems (ICICAS), pp. 401-406

[42] Hou, Y., Guo, Q., Huang, Y., Xie, X., Ma, L., & Zhao, J. (2023). Evading deepfake detectors via adversarial statistical consistency. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12271-12280).

[43] Ain, Q. U., Javed, A., Malik, K. M., & Irtaza, A. (2024, October). Exposing the Limits of Deepfake Detection using novel Facial mole attack: A Perceptual Black-Box Adversarial Attack Study. In *2024 IEEE International Conference on Image Processing (ICIP)* (pp. 3820-3826). IEEE.

[44] Ain, Q. U., Javed, A., & Irtaza, A. (2025). DeepEvader: An evasion tool for exposing the vulnerability of deepfake detectors using transferable facial distraction blackbox attack. *Engineering Applications of Artificial Intelligence*, *145*, 110276.

[45] Laidlaw C, Feizi S (2019) Functional adversarial attacks. Adva in neur inform proc sys. 2019;32

[46] Yuan Z, Zhang J, Jia Y, Tan C, Xue T, Shan S (2021) Meta gradient adversarial attack. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7748-7757

[47] Duan R, Chen Y, Niu D, Yang Y, Qin AK, He Y (2021) Advdrop: Adversarial attack to dnns by dropping information. Proceedings of the IEEE/CVF International Conference on Computer Vision

[48] Cilloni T, Walter C, Fleming C (2022) Focused Adversarial Attacks. arXiv preprint arXiv:2205.09624. https://doi.org/10.48550/arXiv.2205.09624

[49] Yang S, Yang Y, Zhou L, Zhan R, Man Y (2022) Intermediate-Layer Transferable Adversarial Attack With DNN Attention. IEEE Access 10:95451-61

[50] Wang X, Ni R, Li W, Zhao Y (2021) Adversarial attack on fake-faces detectors under white and black box scenarios. In:2021 IEEE International Conference on Image Processing (ICIP), pp. 3627-3631

[51] Zhao, X., & Stamm, M. C. (2021). Making GAN-generated images difficult to spot: a new attack against synthetic image detectors. *arXiv preprint arXiv:2104.12069*, *2*.

[52] Zhao, X., Chen, C., & Stamm, M. C. (2021). A transferable anti-forensic attack on forensic CNNs using a generative adversarial network. *arXiv preprint arXiv:2101.09568*.

[53] Liu, C., Chen, H., Zhu, T., Zhang, J., & Zhou, W. (2023). Making DeepFakes more spurious: evading deep face forgery detection via trace removal attack. *IEEE Transactions on Dependable and Secure Computing*, *20*(6), 5182-5196.

[54] Uddin, K., Tasnim, N., Saeed, M. S., & Malik, K. M. (2025). GUARD: Generative Unmasking and Adversarial-Resistant Deepfake Detection using Multi-Model Knowledge Distillation. *Authorea Preprints*.

[55] Huang, Y., Juefei-Xu, F., Wang, R., Guo, Q., Ma, L., Xie, X., ... & Pu, G. (2020, October). Fakepolisher: Making deepfakes more detection-evasive by shallow reconstruction. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 1217-1226).

[56] Ding F, Zhu G, Li Y, Zhang X, Atrey PK, Lyu S (2021) Anti-forensics for face swapping videos via adversarial training. IEEE Trans on Multi 24:3429-41.

[57] Carlini, N., & Farid, H. (2020). Evading deepfake-image detectors with white-and black-box attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 658-659).

[58] Huang R, Fang F, Nguyen HH, Yamagishi J, Echizen I (2020) Security of facial forensics models against adversarial attacks. In:2020 IEEE International Conference on Image Processing (ICIP) pp. 2236-2240

[59] Li Y, Lyu S (2021) Obstructing DeepFakes by Disrupting Face Detection and Facial Landmarks Extraction. Deep Learning-Base Face Ana 2021:247-67. https://doi.org/10.1007/978-3-030-74697-1_12

[60] Fan L, Li W, Cui X (2021) Deepfake-image anti-forensics with adversarial examples attacks. Futu Inte 13(11):288. https://doi.org/10.3390/fi13110288

[61] Neekhara P, Dolhansky B, Bitton J, Ferrer CC (2021) Adversarial threats to deepfake detection: A practical perspective. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 923-932

[62] Liao Q, Li Y, Wang X et al (2021) Imperceptible adversarial examples for fake image detection. arXiv preprint arXiv:2106.01615, https://doi.org/10.48550/arXiv.2106.01615

[63] Peng B, Fan H, et al (2021) DFGC 2021: A deepfake game competition. In: 2021 IEEE International Joint Conference on Biometrics (IJCB), pp. 1-8

[64] Cao X, Gong NZ (2021) Understanding the security of deepfake detection. In: International Conference on Digital Forensics and Cyber Crime, pp. 360-378

[65] Zhang H, Chen B, Wang J, Zhao G (2022) A local perturbation generation method for gan-generated face anti-forensics. IEEE Trans on Circ and Sys for Vid Techn

[66] Ngoc NH, Chan A, Binh HT, Ong YS (2022) Anti-Forensic Deepfake Personas and How To Spot Them. In:2022 International Joint Conference on Neural Networks (IJCNN), pp. 1-8

[67] Jia S, Ma C, Yao T, Yin B, Ding S, Yang X (2022) Exploring frequency adversarial attacks for face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4103-4112

[68] Lim NT, Kuan MY, Pu M, Lim MK, Chong CY (2022) Metamorphic Testing-based Adversarial Attack to Fool Deepfake Detectors. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 2503-2509

[69] Liu, J., Zhang, M., Ke, J., & Wang, L. (2024). AdvShadow: Evading DeepFake Detection via Adversarial Shadow Attack. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4640-4644). IEEE.

[70] Hussain, S., Neekhara, P., Jere, M., Koushanfar, F., & McAuley, J (2021) Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 3348-3357

[71] Shahriyar SA, Wright M (2022) Evaluating Robustness of Sequence-based Deepfake Detector Models by Adversarial Perturbation. In: Proceedings of the 1st Workshop on Security Implications of Deepfakes and Cheapfakes, pp. 13-18

[72] Hussain S, Neekhara P et al (2022) Exposing vulnerabilities of deepfake detection systems with robust attacks. Digital Threats: Research and Practice (DTRAP), 3(3):1-23. https://doi.org/10.1145/3464307

[73] Liu H, Zhou W, Chen D et al (2023) Coherent adversarial deepfake video generation. Signal Processing 203:108790. https://doi.org/10.1016/j.sigpro.2022.108790

[74] Gowrisankar, B., & Thing, V. L. (2024). An adversarial attack approach for eXplainable AI evaluation on deepfake detection models. *Computers & Security*, *139*, 103684

[75] Ding, F., Shen, Z., Zhu, G., Kwong, S., Zhou, Y., & Lyu, S. (2022). ExS-GAN: Synthesizing anti-forensics images via extra supervised GAN. *IEEE Transactions on Cybernetics*, *53*(11), 7162-7173.

[76] Peng, F., Yin, L., & Long, M. (2022). BDC-GAN: Bidirectional conversion between computer-generated and natural facial images for anti-forensics. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(10), 6657-6670.

[77] Frank J, Eisenhofer T, Schönherr L, Fischer A, Kolossa D, Holz T (2020) Leveraging frequency analysis for deep fake image recognition. In: International conference on machine learning, pp. 3247-3258

[78] Nguyen HH, Fang F, Yamagishi J, Echizen I (2019) Multi-task learning for detecting and segmenting manipulated facial images and videos. In:2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1-8

[79] Afchar D, Nozick V, Yamagishi J, Echizen I (2018) Mesonet: a compact facial video forgery detection network. In: 2018 IEEE international workshop on information forensics and security (WIFS),11pp. 1-7

[80] Cozzolino D, Thies J, Rössler A, Riess C, Nießner M, Verdoliva L (2018) Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510, https://doi.org/10.48550/arXiv.1812.02510

[81] Güera D, Delp EJ (2018) Deepfake video detection using recurrent neural networks. In: 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS), pp. 1-6

[82] Sohrawardi SJ, Chintha A, Thai B, Seng S, Hickerson A, Ptucha R, Wright M. Poster: Towards robust open-world detection of deepfakes. In: Proceedings of the 2019 ACM SIGSAC conference on computer and communications security, pp. 2613-2615

[83] Zhao, Y., Liu, B., Ding, M., Liu, B., Zhu, T., & Yu, X. (2023). Proactive deepfake defence via identity watermarking. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 4602-4611)

[84] Lv L (2021) Smart watermark to defend against deepfake image manipulation. In: 2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS), pp. 380-384

[85] Huang H, Wang Y, Chen Z et al (2022) Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 1, pp. 989-997

[86] Neekhara P, Hussain S et al (2022) FaceSigns: semi-fragile neural watermarks for media authentication and countering deepfakes. arXiv preprint arXiv:2204.01960. https://doi.org/10.48550/arXiv.2204.01960

[87] Liu P, Sun L, Mao X, Dai L, Guo S, Yang Y (2021) A CycleGAN Adversarial Attack Method Based on Output Diversification Initialization. In: Journal of Physics: Conference Series, Vol. 1948, No. 1, p. 012041

[88] Ruiz N, Bargal SA, Sclaroff S (2020) Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, pp. 236-251. https://doi.org/10.1007/978-3-030-66823-5_14

[89] Fang Z, Yang Y, Lin J, Zhan R (2020) Adversarial attacks for multi target image translation networks. In: 2020 IEEE International Conference on Progress in Informatics and Computing, pp. 179-184

[90] Yeh CY, Chen HW, Tsai SL, Wang SD (2020) Disrupting image-translation-based deepfake algorithms with adversarial attacks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, pp. 53-62

[91] Segalis E, Galili E (2020) OGAN: Disrupting deepfakes with an adversarial attack that survives training. arXiv preprint arXiv:2006.12247. https://doi.org/10.48550/arXiv.2006.12247

[92] He Z, Wang W, Guan W, Dong J, Tan T (2022) Defeating DeepFakes via Adversarial Visual Reconstruction. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 2464-2472. https://doi.org/10.1145/3503161.3547923

[93] Ruiz N, Bargal SA, Sclaroff S (2022) Protecting against image translation deepfakes by leaking universal perturbations from black box neural networks. arXiv preprint arXiv:2006.06493

[94] Wang X, Huang J, Ma S, Nepal S, Xu C (2022) Deepfake disrupter: The detector of deepfake is my friend. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14920-14929

[95] Wang R, Huang Z, Chen Z, Liu L, Chen J, Wang L (2022) Anti-Forgery: Towards a Stealthy and Robust DeepFake Disruption Attack via Adversarial Perceptual-aware Perturbations. arXiv preprint arXiv:2206.00477. https://doi.org/10.48550/arXiv.2206.00477

[96] Dong J, Xie X (2021) Visually maintained image disturbance against deepfake face swapping. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), pp. 1-6

[97] Qiu H, Du Y, Lu T (2022) The framework of cross-domain and model adversarial attack against deepfake. Futu Inte 14(2):46. https://doi.org/10.3390/fi14020046

[98] Yeh CY, Chen HW, Shuai HH, Yang DN, Chen MS (2021) Attack as the best defense: Nullifying image-to-image translation gans via limit-aware adversarial attack. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16188-16197

[99] Dong J, Wang Y, Lai J, Xie X (2023) Restricted black box adversarial attack against deepfake face swapping. IEEE Trans on Infor Foren and Sec. doi: 10.1109/TIFS.2023.3266702

[100] Li Q, Gao M, Zhang G, Zhai W (2023) Defending Deepfakes by Saliency-Aware Attack. IEEE Trans on Compu Soci Sys. doi: 10.1109/TCSS.2023.3271121

[101] Das RK, Tian X, Kinnunen T, Li H (2020) The attacker's perspective on automatic speaker verification: An overview. arXiv preprint arXiv:2004.08849. https://doi.org/10.48550/arXiv.2004.08849

[102] Tan H, Wang L, Zhang H, Zhang J, Shafiq M, Gu Z (2022) Adversarial attack and defense strategies of speaker recognition systems: A survey. Electronics; 11(14):2183. https://doi.org/10.3390/electronics11142183

[103] Liu, S., Wu, H., Lee, H. Y., & Meng, H. (2019, December). Adversarial attacks on spoofing countermeasures of automatic speaker verification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 312-319). IEEE.

[104] Li Z, Shi C, Xie Y, Liu J, Yuan B, Chen Y (2020) Practical adversarial attacks against speaker recognition systems. In: Proceedings of the 21st international workshop on mobile computing systems and applications, pp. 9-14

[105] Zhang, Y., Jiang, Z., Villalba, J., & Dehak, N. (2020, October). Black-Box Attacks on Spoofing Countermeasures Using Transferability of Adversarial Examples. In *Interspeech* (pp. 4238-4242).

[106] Wang Q, Guo P, Xie L (2021) Inaudible adversarial perturbations for targeted attack in speaker recognition. arXiv preprint arXiv:2005.10637. https://doi.org/10.48550/arXiv.2005.10637

[107] Nakamura, T., Saito, Y., Takamichi, S., Ijima, Y., & Saruwatari, H. (2019). V2S attack: building DNN-based voice conversion from automatic speaker verification. *arXiv preprint arXiv:1908.01454*.

[108] Xie, Y., Shi, C., Li, Z., Liu, J., Chen, Y., & Yuan, B. (2020, May). Real-time, universal, and robust adversarial attacks against speaker recognition systems. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1738-1742). IEEE.

[109] Li, X., Zhong, J., Wu, X., Yu, J., Liu, X., & Meng, H. (2020, May). Adversarial attacks on GMM i-vector based speaker verification systems. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6579-6583). IEEE.

[110] Chen, G., Chenb, S., Fan, L., Du, X., Zhao, Z., Song, F., & Liu, Y. (2021, May). Who is real bob? adversarial attacks on speaker recognition systems. In *2021 IEEE Symposium on Security and Privacy (SP)* (pp. 694-711). IEEE.

[111] Chang, Y., Ren, Z., Zhang, Z., Jing, X., Qian, K., Shao, X., ... & Schuller, B. W. (2024). STAA-net: A sparse and transferable adversarial attack for speech emotion recognition. *IEEE Transactions on Affective Computing*.

[112] Marras, M., Korus, P., Memon, N. D., & Fenu, G. (2019, September). Adversarial Optimization for Dictionary Attacks on Speaker Verification. In *Interspeech* (pp. 2913-2917).

[113] Tian, X., Das, R. K., & Li, H. (2019). Black-box attacks on automatic speaker verification using feedback-controlled voice conversion. *arXiv preprint arXiv:1909.07655*.

[114] Chen, G., Zhao, Z., Song, F., Chen, S., Fan, L., & Liu, Y. (2021). SEC4SR: A security analysis platform for speaker recognition. *arXiv preprint arXiv:2109.01766*.

[115] Abdullah, H., Rahman, M. S., Garcia, W., Warren, K., Yadav, A. S., Shrimpton, T., & Traynor, P. (2021, May). Hear" no evil", see" kenansville": Efficient and transferable black-box attacks on speech recognition and voice identification systems. In *2021 IEEE Symposium on Security and Privacy (SP)* (pp. 712-729). IEEE.

[116] Chen, G., Zhao, Z., Song, F., Chen, S., Fan, L., & Liu, Y. (2022). AS2T: Arbitrary source-to-target adversarial attack on speaker recognition systems. *IEEE Transactions on Dependable and Secure Computing*.

[117] Rabhi, M., Bakiras, S., & Di Pietro, R. (2024). Audio-deepfake detection: Adversarial attacks and countermeasures. *Expert Systems with Applications*, *250*, 123941.

[118] He, R., Cheng, Y., Ze, J., Ji, X., & Xu, W. (2024, May). Understanding and Benchmarking the Commonality of Adversarial Examples. In *2024 IEEE Symposium on Security and Privacy (SP)* (pp. 1665-1683). IEEE.

[119] Farooq, M. U., Khan, A., Uddin, K., & Malik, K. M. (2025). Transferable adversarial attacks on audio deepfake detection. In *Proceedings of the Winter Conference on Applications of Computer Vision* (pp. 1640-1649).

[120] Ding, F., Fan, B., Shen, Z., Yu, K., Srivastava, G., Dev, K., & Wan, S. (2022). Securing facial bioinformation by eliminating adversarial perturbations. *IEEE Transactions on Industrial Informatics*, *19*(5), 6682-6691.

[121] Uddin, K., Yang, Y., & Oh, B. T. (2023). Deep learning-based counter anti-forensic of GAN-based attack in HEVC compressed domain using coding pattern analysis. *Expert Systems with Applications*, *233*, 120912.

[122] Dutta, H., Pandey, A., & Bilgaiyan, S. (2021). EnsembleDet: ensembling against adversarial attack on deepfake detection. *Journal of Electronic Imaging*, *30*(6), 063030-063030.

[123] Hooda, A., Mangaokar, N., Feng, R., Fawaz, K., Jha, S., & Prakash, A. (2022). Towards adversarially robust deepfake detection: an ensemble approach. *arXiv preprint arXiv:2202.05687*.

[124] Kumar, A., Singh, D., Jain, R., Jain, D. K., Gan, C., & Zhao, X. (2025). Advances in DeepFake detection algorithms: Exploring fusion techniques in single and multi-modal approach. *Information Fusion*, 102993.

[125] Amerini, I., & Menon, V. G. (2024). D-Fence layer: an ensemble framework for comprehensive deepfake detection. *Multimedia Tools and Applications*, *83*(26), 68063-68086.

[126] Jiang, J., Li, B., Yu, S., Liu, C., An, S., Liu, M., & Yu, M. (2021, December). A Residual Fingerprint-Based Defense Against Adversarial Deepfakes. In *2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)* (pp. 797-804). IEEE.

[127] Chen, Z., Xie, L., Pang, S., He, Y., & Zhang, B. (2021). Magdr: Mask-guided detection and reconstruction for defending deepfakes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9014-9023).

[128] Gandhi, A., & Jain, S. (2020, July). Adversarial perturbations fool deepfake detectors. In *2020 International joint conference on neural networks (IJCNN)* (pp. 1-8). IEEE.

[129] Shaoanlu. (2019). *Fewshot Face Translation GAN*. Available: https://github.com/shaoanlu/fewshot-face-translation-GAN

[130] Ain, Q. U., Javed, A., Malik, K. M., & Irtaza, A. (2024). Regularized forensic efficient net: a game theory based generalized approach for video deepfakes detection. *Multimedia Tools and Applications*, 1-44.

[131] Deb, D., Liu, X., & Jain, A. K. (2023, January). Unified detection of digital and physical face attacks. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)* (pp. 1-8). IEEE.

[132] Ke, J., & Wang, L. (2023). DF-UDetector: An effective method towards robust deepfake detection via feature restoration. *Neural Networks*, *160*, 216-226.

[133] Uddin, K., Yang, Y., Jeong, T. H., & Oh, B. T. (2023). A robust open-set multi-instance learning for defending adversarial attacks in digital image. *IEEE Transactions on Information Forensics and Security*, *19*, 2098-2111.

[134] Asha, S., Vinod, P., & Menon, V. G. (2023). A defensive framework for deepfake detection under adversarial settings using temporal and spatial features. *International Journal of Information Security*, *22*(5), 1371-1382.

[135] Pinhasov, B., Lapid, R., Ohayon, R., Sipper, M., & Aperstein, Y. (2024). Xai-based detection of adversarial attacks on deepfake detectors. *arXiv preprint arXiv:2403.02955*.

[136] Wang, L., Meng, X., Li, D., Zhang, X., Ji, S., & Guo, S. (2024). DEEPFAKER: a unified evaluation platform for facial deepfake and detection models. *ACM Transactions on Privacy and Security*, *27*(1), 1-34.

[137] Park, J., Park, L. H., Ahn, H. E., & Kwon, T. (2024). Coexistence of Deepfake Defenses: Addressing the Poisoning Challenge. *IEEE Access*, *12*, 11674-11687.

[138] Luo Y, Ye F, Weng B, Du S, Huang T (2021) A novel defensive strategy for facial manipulation detection combining bilateral filtering and joint adversarial training. Secu and Comm Networks 2021:1-0. https://doi.org/10.1155/2021/4280328

[139] Leporoni, G., Maiano, L., Papa, L., & Amerini, I. (2024). A guided-based approach for deepfake detection: RGB-depth integration via features fusion. *Pattern Recognition Letters*.

[140] Wang, Q., Guo, P., Sun, S., Xie, L., & Hansen, J. H. (2019, September). Adversarial Regularization for End-to-End Robust Speaker Verification. In *Interspeech* (pp. 4010-4014).

[141] Wu H, Liu S, Meng H, Lee HY (2020) Defense against adversarial attacks on spoofing countermeasures of ASV. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6564-6568

[142] Pal M, Jati A, Peri R, Hsu CC, AbdAlmageed W, Narayanan S (2021) Adversarial defense for deep speaker recognition using hybrid adversarial training. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6164-6168

[143] Jati, A., Hsu, C. C., Pal, M., Peri, R., AbdAlmageed, W., & Narayanan, S. (2021). Adversarial attack and defense strategies for deep speaker recognition systems. *Computer Speech & Language*, *68*, 101199.

[144] Isik, M., Vishwamith, H., Sur, Y., Inadagbo, K., & Dikmen, I. C. (2024, March). Neurosec: Fpga-based neuromorphic audio security. In *International Symposium on Applied Reconfigurable Computing* (pp. 134-147). Cham: Springer Nature Switzerland.

[145] Wu, H., Liu, A. T., & Lee, H. Y. (2020). Defense for black-box attacks on anti-spoofing models by self-supervised learning. *arXiv preprint arXiv:2006.03214*.

[146] Liu, A. T., Yang, S. W., Chi, P. H., Hsu, P. C., & Lee, H. Y. (2020, May). Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6419-6423). IEEE.

[147] Wu H, Li X, Liu AT, Wu Z, Meng H, Lee HY (2021) Adversarial defense for automatic speaker verification by cascaded self-supervised learning models. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6718-6722

[148] Li X, Li N, Zhong J, Wu X, Liu X, Su D, Yu D, Meng H (2020) Investigating robustness of adversarial samples detection for automatic speaker verification. arXiv preprint arXiv:2006.06186. 2020 Jun 11. https://doi.org/10.48550/arXiv.2006.06186

[149] Wu H, Li X, Liu AT, Wu Z, Meng H, Lee HY (2021) Improving the adversarial robustness for speaker verification by self-supervised learning. IEEE/ACM Tran on Aud, Spe, and Lang Proc, 30:202-17.

[150] Wu, H., Hsu, P. C., Gao, J., Zhang, S., Huang, S., Kang, J., ... & Lee, H. Y. (2021). Spotting adversarial samples for speaker verification by neural vocoders. *arXiv preprint arXiv:2107.00309*.

[151] Yoon, J., Panizo-LLedot, A., Camacho, D., & Choi, C. (2024). Triple-modality interaction for deepfake detection on zero-shot identity. *Information Fusion*, *109*, 102424.

[152] Villalba J, Joshi S, Żelasko P, Dehak N (2021) Representation learning to classify and detect adversarial attacks against speaker and speech recognition systems. arXiv preprint arXiv:2107.04448. https://doi.org/10.48550/arXiv.2107.04448

[153] Joshi S, Kataria S, Villalba J, Dehak N (2022) Advest: Adversarial perturbation estimation to classify and detect adversarial attacks against speaker identification. arXiv preprint arXiv:2204.03848. https://arxiv.org/abs/2204.03848

[154] Uddin, K., Jeong, T. H., & Oh, B. T. (2024). Counter-act against GAN-based attacks: A collaborative learning approach for anti-forensic detection. *Applied Soft Computing*, *153*, 111287.

[155] Uddin, K., Khan, A., Farooq, M. U., & Malik, K. M. (2025). SHEILD: A Secure and Highly Enhanced Integrated Learning for Robust Deepfake Detection against Adversarial Attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1502-1511).

[156] Uddin, K., Farooq, M. U., Khan, A., & Malik, K. M. (2025). Adversarial Attacks on Audio Deepfake Detection: A Benchmark and Comparative Study. *arXiv preprint arXiv:2509.07132*.

[157] Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., & Li, J. (2018). Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9185-9193).

[158] Agrawal, P., Kapoor, R., & Agrawal, S. (2014, May). A hybrid partial fingerprint matching algorithm for estimation of equal error rate. In *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies* (pp. 1295-1299). IEEE.

[159] Rony, J., Hafemann, L. G., Oliveira, L. S., Ayed, I. B., Sabourin, R., & Granger, E. (2019). Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4322-4330).

[160] Luo, C., Lin, Q., Xie, W., Wu, B., Xie, J., & Shen, L. (2022). Frequency-driven imperceptible adversarial attack on semantic similarity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15315-15324).

[161] Jordan, M., Manoj, N., Goel, S., & Dimakis, A. G. (2019). Quantifying perceptual distortion of adversarial examples. *arXiv preprint arXiv:1902.08265*.

[162] Yu, Y., Zhang, W., & Deng, Y. (2021). Frechet inception distance (fid) for evaluating gans. *China University of Mining Technology Beijing Graduate School*, *3*(11).

[163] Kong, X., & Ge, Z. (2021). Adversarial attacks on neural-network-based soft sensors: Directly attack output. *IEEE Transactions on Industrial Informatics*, *18*(4), 2443-2451.

[164] Qin, Z., Zhang, X., & Li, S. (2023). A robust adversarial attack against speech recognition with uap. *High-Confidence Computing*, *3*(1), 100098.

[165] Xu, B., Li, X., Hou, W., Wang, Y., & Wei, Y. (2021). A similarity-based ranking method for hyperspectral band selection. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(11), 9585-9599.

[166] Hodson, T. O., Over, T. M., & Foks, S. S. (2021). Mean squared error, deconstructed. *Journal of Advances in Modeling Earth Systems*, *13*(12), e2021MS002681.

[167] Prabu, R. T., Kaliyamoorthy, C., Saral, G. B., Rajasekaran, M., & Xavier, B. M. Hybrid watermarking scheme: Walsh-Hadamard transform and SVD integration. In *Hybrid and Advanced Technologies* (pp. 390-395). CRC Press.

[168] Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2018, July). Synthesizing robust adversarial examples. In *International conference on machine learning* (pp. 284-293). PMLR.

[169] Zhao, W., Alwidian, S., & Mahmoud, Q. H. (2022). Adversarial training methods for deep learning: A systematic review. *Algorithms*, *15*(8), 283.

[170] Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., & Yuille, A. L. (2019). Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2730-2739).

[171] Feldsar, B., Mayer, R., & Rauber, A. (2023). Detecting adversarial examples using surrogate models. *Machine Learning and Knowledge Extraction*, *5*(4), 1796-1825.

[172] Dou, Z., Hu, X., Yang, H., Liu, Z., & Fang, M. (2023, November). Adversarial attacks to multi-modal models. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis* (pp. 35-46).

[173] Yin, Z., Ye, M., Zhang, T., Du, T., Zhu, J., Liu, H., ... & Ma, F. (2023). Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *Advances in Neural Information Processing Systems*, *36*, 52936-52956.

[174] Bai, T., Luo, J., Zhao, J., Wen, B., & Wang, Q. (2021). Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*.

[175] Barik, P. K., Pandya, P., & Gaur, S. S. (2024, November). Efficient Deepfake Detection Using AI. In *2024 IEEE 4th International Conference on Applied Electromagnetics, Signal Processing, & Communication (AESPC)* (pp. 1-6). IEEE.
[176] Nadimpalli, A. V., & Rattani, A. (2024). Proactive deepfake detection using gan-based visible watermarking. *ACM Transactions on Multimedia Computing, Communications and Applications*, *20*(11), 1-27.
[177] Jiang, T., Yu, H., Meng, W., & Qi, P. (2023, November). Towards Retentive Proactive Defense Against DeepFakes. In International Conference on Testbeds and Research Infrastructures (pp. 139-153). Cham: Springer Nature Switzerland.
[178] Raza, M. A., Malik, K. M., & Haq, I. U. (2023). Holisticdfd: Infusing spatiotemporal transformer embeddings for deepfake detection. Information Sciences, 645, 119352.
[179] Wang, J., Lei, J., Li, S., & Zhang, J. (2025). STA-3D: Combining Spatiotemporal Attention and 3D Convolutional Networks for Robust Deepfake Detection. Symmetry, 17(7), 1037.
[180] Wang, X., Song, W., Hao, C., & Liu, F. (2025). Deepfake Detection Method Based on Spatio-Temporal Information Fusion. Computers, Materials & Continua, 83(2).
[181] Zafar, F., Khan, T. A., Akbar, S., Ubaid, M. T., Javaid, S., & Kadir, K. (2025). A Hybrid Deep Learning Framework for Deepfake Detection Using Temporal and Spatial Features. IEEE Access.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: