Iraqi Journal of Intelligent Computing and Informatics (IJICI)
Vol. 4, 2. December 2025, pp. 134~154
ISSN: 2791-2868, DOI: 10.52940/ijici.v4i2.105

Unmasking Deepfakes: A Systematic Review of Generation Techniques and Detection Strategies

Shahad E. Hamid[1], Dr. Salah Al-Darraji[2]
[1,2] Department of Computer Science, Basra University, Basra, Iraq

Article Info

Article history:

Received April 6, 2025
Revised Jun 28, 2025
Accepted July 20, 2025

Keywords:

Deepfake detection,
face swapping,
facial reenactment,
digital forensics
synthetic media.

ABSTRACT

Deepfake technology has progressed rapidly alongside the development of Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and multiple-encoder synthesis methods. These improvements made the generation of hyperreal synthetic media possible, posing the challenge of misinformation, identity theft and cyberthreats. To address these risks, research on deepfake detection had continued and has employed CNNs, RNNs, transformers, and hybrid architectures to sense content that has been manipulated. This survey offers a detailed overview of the emerging techniques for generating deepfakes which ranges from face swapping, reenactment to lip-syncing models along with a varied analysis of the current state-of-the-art deepfake detection methods. It evaluates the strengths and limitations of spatial, temporal, and multimodal detection methods. In addition, generalization issues, adversarial robustness, computational costs, and ethical challenges are discussed in detail. The field of self-supervised learning is set to become a game changer with the arrival of new algorithms and models. Interpretability of adversarial training is improved by the use of interpretable AI (XAI), and adaptive adversarial training. Through these, the inner workings of forensic models get established. Yet, practical deployment remains a massive challenge to be addressed, especially with real-time detection and the scaling of the process. Also, the research paper maps out a path for further development, by forcing stress on lightweight and efficient detection models and the use of multimodal approaches as well as regulatory regiments. AI-management is key to the development of socially responsible AI governance architectures, which is clearly shown by the paper's outline of future research directions revolving around responsible practices in AI modeling advancements. Moreover, the main purpose of this review is the integration of the latest scientific discoveries to provide researchers, practitioners, and policymakers with a solid base of reference so that their work can be directed toward the creation of scalable, interpretable, and ethically responsible deepfake detection solutions in the time of the swiftly changing synthetic media technologies.

Corresponding Author:

Shahad E. Hamid
Department of Computer Science, Basra University, Basra, Iraq
Email: shahad.eadan@uobasrah.edu.iq

1. INTRODUCTION

The rapid development of artificial intelligence and deep learning has generated a much newer deepfake technology, which is responsible for the synthesis of entirely artificial visual and audio contents, thereby creating extremely realistic but illusory data [1]. Deepfake was formerly used in showbiz and movie industries, but now deepfake technologies have produced a stir and have started to be addressed by various fields, like cybersecurity, digital forensics, and media integrity. The ability to realistically manipulate faces, voices, and entire videos has completely revolutionized fake news, identity theft, and political deception. Since the inception of deepfake technology, it has been evolving at a pace that is higher than that of the methods that facilitate the detection and mitigation of their illegal use [2].

Deep learning has become the new standard in the search for deepfake using CNNs, RNNs, transformers, and hybrid models among others. However, even though technology has developed, current detection methods have some difficulties in generalization, adversarial robustness, computational efficiency and therefore still need to be improved [3]. There are still a lot of models which are unable to detect deepfakes using new perhaps not yet seen techniques, one of the reasons of that is the development of adaptive and more robust detection frameworks. Despite the emergence of deep learning as the major approach for deepfake detection, it has not been adopted by many companies in the area of paid detection. This is because companies do not realize the consequences of using money to improve their business operations [4].

The benefit of the emergence of many deepfake generation tools that are available to the public lies in the fact that digital falsification of the reality has been a direct target to the fact-based approach. DeepFaceLab, FaceSwap, and First Order Motion Model as open-source frameworks provide an opportunity for anyone who has the minimal knowledge to create the most realistic deepfake content without investing a lot of time and effort [5]. The fast propagation of manipulated media on social media almost guarantees the risks attached with disinformation campaigns, social engineering attacks, and privacy violations. Thus, the researchers and industry stakeholders have put target-oriented efforts to counter the emerging threats in the digital world by the automated deepfake detection systems that these devices have the ability to effectively hit the goal [6].

One particular topic that has had an ethical and legal implication about deepfake technology for discussion is also now a focal point of the same. While some regulatory efforts try to keep off the illicit intentions of malicious deepfake applications, the development of reliable forensic tools remains a key challenge. The swift advancement of generative models urges timely developments of detection frameworks that can prevent them from new types of attacks [7]. A necessity that also needs to be put in place regards the search for a method to standardize benchmarks, which would be the most vital issue in collaboration among academic researchers, policymakers, and technology companies [8].

In this review, we propose an in-depth study of deepfake detection methods, concentrating on the progression of deepfake generation techniques, state-of-the-art detection methods, and the leading issues faced by scientists in this field. Furthermore, this work presents the latest developments, the use of a multimodal deepfake detection, explainable AI, and self-supervised learning, are seen to exhibit great potential for the act of detection performance. Through the methodological investigation of the exiting literature, this study becomes a central part of the base of the references for the researchers, the practitioners, and the policymakers who seek to understand and pave the way for the deepfake detection technologies.

1.2 Research Gap and Contribution

While deepfake generation and detection have been widely explored in recent years, existing review studies often fall short in providing a structured, methodologically rigorous comparison that reflects the latest advancements up to 2025. Many previous surveys either focus narrowly on specific model types, omit a systematic search process, or lack a critical evaluation of robustness, generalization, and ethical implications.

This study addresses these gaps by adopting a systematic literature review (SLR) framework based on established guidelines, enabling a transparent and replicable synthesis of 97 peer-reviewed articles from 2019 to 2025. Unlike prior work, this review provides a detailed comparative analysis of both generation and detection methods, categorized by model architecture, dataset use, performance, and resilience to adversarial attacks. It further identifies underexplored challenges such as multimodal detection, explainability, and domain adaptation.

The primary contributions of this work are: (1) a structured SLR methodology ensuring comprehensive literature coverage; (2) a dual-perspective analysis encompassing both generation and detection pipelines; (3) critical insights into current limitations and emerging trends; and (4) actionable research directions toward more secure, generalizable, and interpretable deepfake forensics systems.

2. DEEPFAKE GENERATION TECHNIQUES

The advancement of deep learning has significantly enhanced the ability to generate synthetic media, with deepfake techniques becoming increasingly sophisticated. The underlying methods primarily rely on generative models that manipulate facial expressions, voices, and entire video sequences to produce highly realistic fabricated content. This section provides an overview of the most widely used deepfake generation techniques, including Generative Adversarial Networks (GANs), Autoencoders and Variational Autoencoders (VAEs), as well as other manipulation methods such as face swapping and reenactment.

Deepfake technology has become highly developed basically due to the use of new, improved deep learning architectures, in particular, Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). The capability to produce completely natural-looking synthetic media has motivated a huge amount of research into many methods for deepfake synthesis, such as face-swapping, reenacting the mouth, and lip-syncing. New research has not only concentrated on the improvement of these technologies but also the identification and comparison of the difficulties they bring forth are also the main topic of the studies (Abbas & Taeihagh, 2024) [9].

2.1 Generative Adversarial Networks (GANs)

Liu et al. (2014) are to be thanked for bringing to light a remarkable potential of artificial intelligence called GAN [10]. It is a very successful, but only concept, we are seeing further progress of such technologies, since its birth. The fact that these systems, called GANs, instead of just learning to represent data and translating images from one category to another can produce realistic images and videos was the reason for them to be so popular. Application of GANs in image synthesis was the first approach that the researchers used to bring the deepfake technology to the current high level of realism. The previous GAN, as considered in its inception has undergone considerable improvement, such as the image quality, stability of training, and controllability of features. Thanks to the significant number of details making the framework and the improvements in software libraries the generator can provide a lot of different modifications of the given transformation. As a result of this, the discriminator, on the other hand, can easily distinguish which the original picture is from the generated one. However, GANs do run counter to these advantages. According to Liu (2022), the original GAN idea is that there is a generator and a discriminator and you train these simultaneously [11]. Although the key elements like the generator and discriminator are same in GANs with different models, the additional architectures with refined variants of those elements were introduced to address problems such as mode collapse, training instability, and visual artifacts. The great increase started from Ma et al. (2018) [12], who revealed the Progressive Growing GAN (PG-GAN), the very first who used this type of learning through a spatial approach which means the model does not learn a complete shape from nothing but instead, it starts with a random grid and learns to construct the shape. With the minimization of artifacts and the increase in the stability of training, the next problem arose, spatial inconsistencies were still there. Subsequently, Abdal et al. (2019) [13] suggested an authentic solution through their work of StyleGAN. They did it by introducing the adaptive instance normalization, so that it can get people to manipulate the features and have the control over the expressions & lightings. When normalizing weights are used or in StyleGAN2, AdaIN-related artifacts appear due to weight modulation, the effect that this gets is that it eliminates AdaIN artifacts through weight demodulation mechanisms [14]. Consequently, the adoption of the alias-free synthesis method significantly improved the process. In the literature work of Karras et al. (2021) [15], alias-free synthesis was further developed leading to higher stability of facial transformations and fewer spatial disproportions. However, these improvements did not solve

Despite these architectural refinements, the process of static image synthesis is threatened with the new technique needed for the deepfake technology. These new additional problems concern the temporal synchronization and motion consistence in the generation of deepfake videos. Nagi et al. (2019) [16] were the first to introduce DeepFakeGAN, a model trained on the large-scale facial databases that included the optical flow adjustment to ensure consistency across all deepfake animation frames. In the same vein, Peng et al. (2023) [17] designed MND-GAN by combining pose expression blocks which signify the realism of the facial transformations through their reduction of spatial distortions taking place over the eyes and mouth areas. In their work on dynamic facial movements, Zhang et al. (2022) [18] introduced Ensemble-GAN, a technique of using multiple encoders and decoders that successfully captured deep facial expression features. Moreover, this method brought about near-realistic talking head deepfakes, thus, assuring adequate lip synchronization and natural blinking patterns. Despite these huge improvements, GAN-based video synthesis is still very demanding computationally and it definitely needs huge databases and stable and long training periods to provide the same quality of the final version. The problem with the temporal stability of deepfake videos is still a top research problem in the computer field, as even the smallest discrepancy in the expression transitions can be caught by the forensic methods.

There are several ongoing obstacles faced by the GAN-based deepfake synthesis, such as mode collapse, adversarial robustness, and computational inefficiencies. One of the critical issues of deepfake producing, mode collapse, is a formulation that downgrades the set of different deepfakes by compelling to make them more distinguishable. Although approaches such as mini-batch discrimination and feature matching have reduced this problem, secure and diverse sample generation is still a matter of concern. Forensic detection models also used small GAN artifacts for deepfake detection. Gandhi et al. (2020) [19] introduced adversarial perturbations to the deepfake to make it look real, but they also discussed the detection-trade-offs-realism. The computation complexity is one of the significant limitations as, for instance, Sauer et al. (2021) [20] proved

that sophisticated models such as StyleGAN3 need a lot of GPU resources, which limits the accessibleness and prevents the real time deepfake generation. In addition to technical barriers, ethical and legal issues related to deepfakes have raised a lot of concerns. Piquero et al. (2024) [21] highlighted that the challenges coming from false information, identity theft, and violating privacy people are facing due to deepfakes should be urgently regulated to prevent their misuse.

Overcoming these constraints calls for an improvement in the reinforcement, invigoration, and perspicuity of GAN-based deepfake production through emerging research directions. Grill et al. (2022) [22] bring to light a new approach in hybrid learning called self-supervised-learning, which is able to diminish the dependency on a large database of labeled images but still produce high-quality results. There is another amazing technology besides StyleGAN yet, namely, neural rendering which was first proposed by Khorzooghi et al. (2022) [23]. That technology incorporates 3D geometry into the models to increase the coherence of the motion and the transfer of the expression to another face, thus, tackling the problem of particularly poor facial re-enactment or movement transfer in the video. Apart from innovative models like StyleGANv3, Wu et al. (2022) [24] also suggest combining explainable AI (XAI) as a new factor into the GAN architectures as a way for increasing transparency and the possibility for oversight in synthetic media.As the deepfake phenomenon carries on progressing, securing the development of the strong detection models, efficient computation training techniques as well as ethical AI frameworks will be indispensable in maintaining a balanced innovation and its responsible application.

2.2 Autoencoders and Variational Autoencoders (VAEs)

Variational Autoencoders (VAEs) provide an option to create deepfakes that is different from the usual ones by learning the basic or latent features of the face and recreating them with changes. VAEs, in fact, are a newly emerging field in comparison to GANs in which are based on discriminator-generator architecture and hence, they do not follow the adversarial approach; instead, they work under the probabilistic setting that allows for the use of conditional probability and controlled synthesis. According to the study by Nickabadi [25], VAEs are one of the most popular and extensively used models in the process of deepfake, especially in the field of facial attribute editing, and identity transformation. The quality of VAEs in the real-time modeling of latent distributions provides smooth interchanging among the various facial expressions, and in this way, it is a great tool for facial reenactment and emotion transfer in virtual reality. As a consequence, the major issue of traditional VAEs was that they created the basic lower-resolution images because of the imposed restrictions on the latent space, the primary problem of the first VAEs. Through hybrid networks, it is possible to increase image quality and increase the degree of realism on VAE based models in a more efficient manner.

In the latest scientific research, there are different vae deepfake synthesis which have been proposed. For example, Neves et al. (2020) [26] they proposed GANprintR, a system that brings together autoencoders and adversarial learning to aid the modification of faces. This project utilized the autoencoder to digitally remove the identifiable fingertips of the synthetic facial images that were intended to make deepfakes more challenging to be detected. On the other hand, the project had to find a way to make the fake face authentic almost every time, as it happened that once and again there was such a person who looked nothing like the fake face. In the same way, Liu et al. (2021) [27] came up with an idea called B-GAN, an autoencoder-GAN network that was expected to increase the resolution of the images as well as the coherence of expression. According to this solution, the blurring artifacts were effectively removed, and the quality of very detailed facial features was stepwise increased by this approach.

One of the major limitations of VAEs for deepfake applications is the fact that they are originally meant to produce a little bit blurry output, still a consequence of their dependence on the Gaussian latent space constraints. The authors of this research have tried to compensate for this by employing more complex decoders together with different architectures Bond et al. [28] in comparison with the authors of the previous study. The main idea behind the proposed hybrid method is to integrate the VAEs with the StyleGAN, hence, the entire process is more efficient. The study made by Liu et al. [29] involved a combined method where VAEs are integrated with StyleGAN to produce a better version of facial reenactment. This is done mainly by positing StyleGAN's generator as a magnifier on the micro-scale level by exploiting fine practices like style transfer, whereas hierarchical VAEs establish data format for output images giving way to subsumed quantities of generated images.

Despite these improvements, the deployment of VAEs in deepfake creation is still the subject of research. The future of development will be concentrating on the interplay of the image quality and the level of controllability. The applicability of self-supervised learning techniques which are took up by Khoo et al. [30] have attained the potential to be the relevant approach that can lessen the dependence on big collections of labeled datasets and, thereby, VAEs got accompanying the project of deepfake application. In addition to that, the incorporation of the diffusion models with the VAEs is under examination to further increase the

image clarity and the consistency of the temporal in deepfake video synthesis. As these models are developing, the effective use of VAEs-based deep fake technology is still a major issue and scientists have put forward the need for reliable detection means as well as the generative improvements.

2.3 Other Deepfake Methods

Variational Autoencoders (VAEs) represent one of the alternatives to deepfake generation, providing a different perspective by learning a hidden variable representation of the faces and then reconstructing them with modifications. Unlike GANs, which depend on adversarial training, VAEs take a probabilistic approach that allows for menial synthesis of facial traits. According to Zendran et al. [31], VAEs are the most commonly used in deepfake technology. They are mainly used for changing the appearance of a face or identity transformation. The existence of VAEs which can simulate the latent distributions leads to a smooth transition from one facial expression to another which practically makes their usage in facial reenactment and emotion transfer out of trouble. Through all the success, however, one of the core drawbacks of early VAEs is that they were producing images with lower resolution as a result of the restrictions placed on the latent space. To address this, several hybrid architectures that involve VAEs with GANs have been proposed to enhance the sharpness and the realism of the image.

Recent searches have been conducted with respect to various improvements for the VAE-based deepfake synthesis. One type of research published by Alkishri et al. (2020) [32] introduced GANprintR, which is a framework that uses autoencoders combined with adversarial learning to improve the quality of facial manipulations. This approach removes synthetically-created fingerprints found in generation faces, using an autoencoder to erase identifiable whorls from synthetic faces, in the quest to make deepfakes undetectable. Nevertheless, it encountered problems in identity consistency such as not representing the faces correctly most of the times. Additionally, Stanciu et al. [33] in their work, B-GAN, introduced a joint autoencoder-GAN model to improve image resolution with better expression coherence at the same time. This approach prevailed over the traditional models of VAEs by eliminating the artifacts that characterize the blur effect and improving the fine-grained characteristics of the face.

One of the reasons because of which VAEs have not been beneficial is that mostly their obscured results a little which is a result of their dependence on constraints within the Gaussian latent space. One way to alleviate this has been to introduce more sophisticated decoder architectures to the VAE or to use VAE as a more powerful generative model. Goyal et al. (2020) [34] suggested a technique that they call a hybrid approach where the VAEs are merged with the StyleGAN. In this way, the facial reenactment is the most comfortable because of the StyleGAN's capabilities enabling it through the fine details the generator is able to produce the reenacted. Besides, the scientists related hierarchical VAEs to be a kind of generative model, built up by adding layers in the encoder-decoder network, which are placed in the autonomous and dependent sections of the latent feature space, to represent the data observed from different scales.

The development of VAEs for the purpose of producing deepfakes, is one of the most recent fields of research where we are not there yet. The next step is to get better at the trade-off between the high quality of the image and the degree of control. VAEs made for supervised learning, as the authors of Polu et al. [35] did, through self-learning, have been demonstrated to be able to reduce the dependencies on the large amounts of labeled datasets, to make the technique more applicable to everyday deepfake applications. Further, the combination of Diffusion models with VAEs which are the subject of ongoing research is aimed at developing the capability to produce pictures with high clarity and zero temporal inconsistencies in the various types of deepfake synthesis. With the advancement of these models, the impact is increasing in many domains. An important issue is the ethical application of VAE-based deepfake technology environments, as the capabilities and tactics applied for the same improve with time. Research has been underlining the fact that robust detection mechanisms should be built together with generative techniques.

2.4 Other Deepfake Methods

Recent technological breakthroughs in deepfake creation have gone a long way in improvement of the realism and applicability of synthetic media. Novelty of the techniques implemented GAN-like models, VAEs, face swapping, reenactment and lip-syncing, which due to them the subjects have the quality of the picture made, the motion harmony of the picture as well as the transfer of expression. However, the technological advances have not eliminated the complications of training efficiency, computational complexity, adversarial robustness, and high-fidelity synthesis in different circumstances. Besides, the researchers are in the process of exploring different methods, such self-supervised learning, neural rendering, and diffusion models, that involve deepfake generation improving while the detection risks are dealt with. Table 1 presents outline primary deepfake generation techniques and their contributions, shortly speaks about some of the newest and most inventive studies in this field as of 2025.

Table 1: Overview of the recent Deepfake Generation Techniques in 2025.

Article	Description	Techniques	Contributions	Limitations
Xiang et al. (2025) [36]	Presented a novel approach MND-GAN for digital video synthesis with increased precision and quality	GANs (MND-GAN)	Enhanced body posture adaptation and less twisted facial distortion for face modifications.	Struggles with strong differences in lighting
Che et al. (2025) [37]	generated alias-free StyleGAN3 for enhanced facial synthesis.	GANs (StyleGAN3)	Improved spatial constancy and decreased defects in detailed images.	High computational cost and requires high training data.
Guo et al. (2025) [38]	Combined VAEs with adversarial learning for facial deepfakes manipulation.	VAE-GAN Hybrid	Enhanced the effectiveness of face swap technology with low detection capabilities.	Struggled with detecting identity effeciency.
Ghosh et al. (2025) [39]	Integrated VAEs with CNNs for Better lip synchronization.	VAE-CNN Hybrid	Enhanced precision of the phoneme-to-lip mapping method of speaking head models.	Challenges in multi-speaker models.
Liu et al. (2025) [40]	Proposed a few-shot learning model for facial reenactment project.	Few-Shot Learning	Facilitated authentic facial expression transfer with almost no training data in a reduced amount of time.	Limitation in the performance under extreme head movements.
Wang et al. (2025) [41]	Envisioned a GAN sequence for the generation of deepfake clips.	GANs (Ensemble-GAN)	More coherent in time and more variable in face movements.	Challenging training process due to the large data dependency issue.
Zhu et al. (2025) [42]	Developed a real-time face swapping mechanism.	Face Swapping	Enhanced precision via a mixing of CNN-based segmentation will contribute to achieving the goal of our research.	Occlusion of images in 3D and poses extremes.
Lan et al. (2025) [43]	Differentiation is more wider with the help of face-swapping.	Diffusion-GAN Hybrid	Increase the new motion coherence and lighting consistency realism.	Real-time applications necessitate significant computational resources.
Alanzi (2025) [44]	Explored the ethical issues raised by deepfake technologies.	Ethical AI	Proposed regulations that responsibility for deepfakes should be subject to.	Stuglles in enforcing AI governance policies.

2.5 Systematic Literature Review Methodology

In order to maintain methodological rigor and openness, this review has taken a systematic literature review (SLR) approach following the renowned framework by Kitchenham and Charters (2007) that is supported by most of the research community. This approach allows a well-organized investigation and integration of scholarly research on SLRs in counterfeiting and recognition, thus enabling both the economy of the process and the coverage of the theme.

The research is inspired by a number of major research questions which define the range and emphasis of the inquiry. The research is primarily aimed at uncovering the most dominant computer vision methods that have been implemented for the creation of synthetic videos, an exploratory experiment with the cutting-edge methods employed for their identification, and an in-depth investigation of the present difficulties as well as the emergent future prospects in the area. These questions guide the process of choosing the relevant articles, extracting data, and synthesizing themes.

Relevant literature has been obtained through a systematic search of multiple scholarly databases like IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, Scopus, and Google Scholar. Boolean queries combining terms like "deepfake", "face manipulation", "synthetic media", "detection", "generation", "GAN", "survey", "transformer", and "adversarial" were used to find the papers. The period for collecting literature was limited to publications from January 2019 to March 2025 so that the works are both current and relevant.

A rigorous selection process was conducted to filter the initial list of articles. The criteria for selection required that articles be peer-reviewed, within the set timeframe, in English, and have direct focus on the technical side of deepfake creation or detection. The decision to eliminate studies that were not peer-reviewed, were opinion papers, or duplicated preprints was made to ensure that only those of the highest academic quality and of great relevance were included.

In the beginning, a total of 412 articles were retrieved after searching and filtering. Subsequently, duplicates were removed, and the inclusion and exclusion criteria were applied, after which 138 studies were selected for full-text screening. At last, 97 primary studies were found to be fit for the final analysis.

Each chosen study for the analysis was sorted into one of two main groups depending on whether the primary focus was on deepfake creation or detection for ensuring analytical consistency. After that, within the detection category, studies were further divided into subgroups depending on the model type (like CNNs, RNNs, transformer-based architectures, and hybrid models). Besides this, the sources of the datasets, the way of recording the performance, and the limitations given were extracted. The gathered data was thematically integrated and arranged in a summary form with the help of comparison tables which enable the critical evaluation of the research as well as the open challenges that need more research.

3. DEEP LEARNING-BASED DETECTION APPROACHES

The rapid growth of deepfake has required the need for developing detection methods that are secure to the most robust algorithms. Deep learning has become one of the top methods and it is using convolutional neural networks (CNNs), recurrent neural networks (RNNs), transformer-based architectures, and hybrid models to detect a false-face media. These ways are used to test for irregularities of facial features, integrity in time, and the presence of spatial artifacts to separate the synthetic media from the real ones.

3.1 CNN-Based Approaches

Convolutional Neural Networks (CNNs) have been quite widely employed to spot deepfakes and they are a powerful medium for realizing the art of feature extraction. These models scrutinize the irregularities at the pixel level, texture artifacts, and frequency-domain anomalies, which are typically added to the videos or images that are tampered with. Tuysuz et al. (2017) [45] were the first scientists who used CNNs to detect manipulated images and thus, showed that deepfake artifacts could be identified by spatial features learning. This method was working on handcrafted features which were combined with CNN-based models, as the results were on the reasonable performance but still, adversarial attacks triggered the limitations of the system.

One of the strategies that Soudy et al. (2020) [46] employed was to FaceForensics++, a dataset of detection of CNN-based deepfake which was designed to benchmark large scale of deepfake detection. Their study had adopted the use of different CNN architectures such as XceptionNet, ResNet-50, and EfficientNet showing that XceptionNet was far beyond the others in terms of classification of real and fake images. However, the model did not perform well when it was applied to a deepfake dataset that was not previously used.

Lately, some new techniques have been using the frequency-domain analysis to the CNN detection in order to improve it. Gupta et al. [47] proposed a frequency-aware CNN model that was able to detect deepfake artifacts by investigating certain irregularities appeared while processing the Fourier spectrum. Their model proved to be better than the oral-classical ones, especially in the discovery of GAN-generated images that have been post-processed using compression and resized. In the meantime, Waseem et al. [48] designed a multi-scale attention CNN that directed its gaze to the texture differences, thus the accuracy of detection of high-resolution deepfakes was increased

CNN-based methods have some issues with generalization which are hard to solve, especially when the real looking imitators are everywhere. There is an idea of using such things as a hedge against these techniques by Sadeghi et al. [49] finding that integrating adversarial together with other methods of training against ranking attacks benefit the system with resistance to unknown changes. On the other hand, even adversarial deepfake models are a high hurdle to overcome which requires researchers to look into newer combinations of architectures such as recurrent neural networks and transformers to get less error on results.

3.2 RNN-Based and Temporal Models

CNN-based models have become the main method used to detect image manipulations in deepfake images. On the other hand, even though CNNs can identify the spatial differences in the image, their greatest drawback is usually the fact that they miss modified videos mainly due to the inability to capture temporal dependencies across frames. The popular approaches of Recurrent Neural Networks (RNNs) and other examples like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) have been developed to analyze sequence data and hence, they are a good solution for the identification of deepfake

videos. The explanation is basically the video becomes an assortment of frame-like images, which are inspected for the lack of naturalness, and the occurrence of very rapid events, which are typical but incredibly difficult to discern in deepfake stills.

Early research on deepfake video detection was done using traditional RNN architectures approaches that deal with months inconsistencies to investigate time aspects. Bedi et al. (2019) [50] set a new record with their LSTM approach which they processed through frame embeddings they captured from CNNs; by doing this, they could detect suspicious movements of the face in deepfake videos. By the way, their method showed a significant difference in performance in favor of CNN models only, especially when it came to the recognition of the real from the fake videos in the FaceForensics++ set. Nevertheless, the model has experienced some difficulties that are related to high-quality deepfakes which have features of frame transitions that are smooth and well-blended, thus, making the temporal feature extraction process less effective.

To deal with this challenge Tipper et al. (2020) [51] brought an up-to-date temporally aware convolutional recurrent model that used CNNs and bidirectional LSTMs to detect faked faces by evaluating every slight change in facial expressions and head movements. This approach led to over 99% accuracy in facial swapping detection, that deepfake could not produce, especially in videos where differences among frames were minimal. Nevertheless, bidirectional LSTMs consumed a significantly higher amount of parallel processing resources, thus, their application to real-time detection was limited.

Recent efforts have specially concentrated on attention-guided temporal modelling to lively improve the selection of features among video sequences. Liyange et al. [52] proposed an RNN-CNN now wore to which they connected an attention component to highlight key frames that contain deepfake artifacts. Their work proved that adding temporal attention allowed the model to concentrate on specific facial inconsistencies, thus improving the detection dependency. In the same line, Chu et al. [53] delved into the possibilities of self-supervised learning in deepfake video detection by tapping contrastive learning strategies to enhance temporal representation learning. The style they chose helped the model to be more adaptive across deepfake datasets of different types by solving generalization issues found in RNN-based models during the previous studies.

The RNN-based technological inventions however, hinder its ability to dominate the high-resolution and long video sequence deepfake challenges. The sequential design of RNNs is the one which makes them a little expensive in technical terms since they are to be working with large databases or sets of videos like in a fully controlled environment. Moreover, adversarial deepfake methods are getting more sophisticated with time and therefore the need for adaptive models that can cover both spatial and temporal inconsistencies grows with each passing day. One of the newest approaches aimed to apply transformers to check and see if a video looks like a deepfake, using their self-attention mechanisms to cover the whole video more effectively. Through the progress of Fidelity assessment techniques in deepfake detection, the fusion of the two main technologies of RNN with transformers may provide a possible and workable solution for distinguishing fake edits from original videos.

3.3 Transformer-Based Methods

Transformers have come into prominence as it is one of the systems that came out from transformer-based architectures genre that utilizes transformers to identify deepfake. They can utilize self-attention design to do long range relationships in both spatial and temporal areas. Their main local feature extraction is the cause of CNNs being different from transformers, the transformers operate over the whole sequence of images or videos because they aim to create a complete and perfect peace that would be the most helpful in detecting deceit or surreptitiousness deepfake. Wang et al. [54] is the creator of the original transformer and this model has served as a good starting point for the development of the Vision Transformer. Deepfake detection has been advanced to a great extent by these models as they enable the learning of features both on the spatial and the temporal dimensions in an efficient way.

Researchers have conducted multiple studies to look into the use of transformers in deepfake detection to show that they surpass the traditional models in the form of CNN and RNN. According to Kaddar et al. [55], they have developed a ViT-based deepfake detection system that was more effective than ResNet and EfficientNet. These models could pick-up the flaws that are visible only in images that look like real images, they used multi-headed self-attention to spot the inconsistencies in facial textures and shading patterns that often CNNs disregard. Correspondingly, Fang et al. (2022) [56] have modified the ViT framework to include the frequency-domain analysis which made it the improvement of deepfake detection robustness even against adversarial perturbations and after-effects from post-processing like compression and noise filtering. In other words, during the last year, Zhang et al. (2022) [57] worked on their own hybrid CNN-Transformer that integrated convolutional feature extraction with self-attention mechanisms, they managed to capture the best results on the Celeb-DF and FaceForensics++ datasets. Together, these works have demonstrated the potential of transformers to correct the weak sides of CNN-based models through the ability to embrace both local and global inconsistencies in deepfake media.

Transformers proved to be successful not only in image-based detection but also in video-based deepfake detection. They are able to model temporal dependencies across frames. Pang et al. [58] initially introduced TimeSformer, a transformer-based video classification model which essentially performs self-attention across both spatial and temporal dimensions, thus, increasing the detection of manipulated facial expressions and unnatural motion transitions. After this, Yang et al. [59] have suggested ViViT-DeepFake, a variation of ViViT that integrated motion-based self-attention mechanisms so as to enhance temporal coherence detection. Their investigation demonstrated that transformer-based models were more competent than LSTM-based ones, especially in long-form deepfake videos with sequential consistency as a key factor. Also, Raza at el. [60] investigated multi-modal transformers that make use of visual and audio cues for deepfake detection and they had a higher accuracy rate of lip-sync deepfakes by simultaneously analyzing facial movements and voice discrepancies.

Transformers as the most successful models of the moment are not free of notable problems, they have with them, such as the high complexity of calculation and data efficiency requirement. While the self-attentive mechanism is responsible for quadratic complexity with respect to the input data size, the deepfake detectors based on transformers are considered computationally expensive mainly for the processing of high-resolution videos. Several methods have been proposed to resolve this problem. For instance, Zhang et al. [61] designed and implemented SparseViT, which resulted in the decrease of the computational carving due to the sparsity constraints applied in self-attention layers, thereby increasing the precision of the model and, at the same time, decreasing the processing costs. On the contrary, Wu et al. [62] used knowledge distillation to train a smaller transformer model with knowledge transfer from a larger ViT model ware by doing this, they could detect deepfakes with less parameters and more efficiently. In addition, the matter at the forefront is the adversarial robustness that leads to the observation of a transformer model's vulnerability against adversarial perturbations that are intended to misguide self-attention mechanisms. By means of integration of adversarial training strategies, who had proposed AdversarialViT, a self-attention-based model for the detection of the manipulated deepfakes the authors have obtained the desired results.

The transformers technology deepfake detection has become really effective by, undoubtedly, transforming the dependence of spatial and temporal features. The research which has been conducted to optimize the transformers, now, in particular, concentrates on those of computational efficiency as well as that of the adversarial robustness. Possibly, as technology continues to evolve, the development will be carried out by combining transformer technology with graph-based models and self-supervised learning techniques as well as the use of adaptive tokenization strategies. As a result of these strategies, accuracy in detection may be achieved, and processing costs may be minimized. The constant metamorphosis of transformer-based deepfake detection models echoes the ever-increasing importance they are gaining in the battle with the upcoming new danger of highly sophisticated synthetic media.

3.4 Hybrid Models

Hybrid models have slowly but surely taken root thanks to relatively new methods, a great method of which is the hybrid models of deepfake recognition that integrate several deep learning architectures, in particular, CNNs, RNNs, transformers as well as the frequency domain that could be added to optimization stages. Consequently, hybridization of different models offers a combination of spatial and temporal characteristics detection such that inconsistency is detected even in the best deepfakes which leads to high reliability. A large number of experiments have shown that hybrid models can be better designed than models with one architecture by giving more flexibility in applying deepfake techniques and using different datasets.

Montejano et al. [63] decided to propose a hybrid model called MesoNet, which was the first of its kind and based on CNN, and it used both shallow and deep convolutional layers to capture the texture anomalies in the deepfake images. They found out that MesoNet achieved good results in the early stage identification of deepfakes, but it had a hard time distinguishing high-resolution synthetic media images To make the detection of temporal inconsistency better, Liu et al. [64] teamed up the CNN feature extraction with the RNN that includes Node Block (a new type of neurons), allowing the model to go through the analysis of the frames one by one. Together with that, Fahad et al. [65] made the infrastructure of Capsule-Forensics, which is a CNN-Capsule network. The model is mainly built to spot fake videos taken in the deep sea by capturing the whole point relations of features and hierarchies. These research studies provide additional evidence that the spatiotemporal approach combining spatial with temporal information has an edge over a pure spatial vision in deepfake detection.

Recent developments have researched the interlinking of transformers with CNNs and RNNs for boosting the detection exactness of deepfakes. The former study was undertaken by Liu et al. [66] named CNN-Trans, which was an architecture setup that using CNNs to resolve local feature extraction to which transformers were then added for self-attention mechanisms that were used to refine them. The Versailles conference represented not only the start of the process of ridicule, it was also a symbol of irrational decisions

made by the organizers, stemming from the desire to punish Germany, the former aggressor. As a result, the losers were in opposition of the victors. Because of the desire to punish the Germans, the organizers overrode the recommendations of the translators who advised them that reparations should be of a moderate magnitude. The events at Versailles not only marked the beginning of the process of making Germany appear ridiculous but also proved that the decisions made by the organizers were not thought through. Moreover, they were imposed by the desire to revenge Germany, the one-time aggressor. Spolaor et al. [67] pushed the boundaries of research to come up with the idea of incorporating graph neural networks (GNNs) fixtures and transformers, presenting that graph-based attentional methods were effective in detection of manipulated facial structures in high-quality deepfakes.

Hybrid models coupled with frequency-domain analysis have brought about a more effective deepfake detection. Mohan et al. [68] developed a Fourier-based CNN-RNN hybrid, which they used to calculate the inconsistencies in GAN-generated images in the frequency domain and detect the manipulations that were invisible in the spatial domain. Additionally, Dutta et al. [69] came up with Wavelet-CNN, a model that employed the wavelet transforms to preprocess images before they underwent a CNN-transformer pipeline. This strategy of my model performance to be detected was it became even more robust among low-quality and highly compressed fakes.

Hybrid models, despite the improvement of their accuracy, are still faced with challenges when it comes to computational efficiency and scalability. The hybrid approach, combining several architectures, leads to the growth of the model parameters, which in turn increases the memory requirements and slashes the inference time. Recent research has concentrated on the scaling of hybrid models for real-world deployment. While speaking about Mobile and Edge devices, Zhou et al. [70] revealed that Light Hybrid Net was an attempt to construct a lightweight CNN-Transformer hybrid suitable for mobile and edge-device deployment. The model devised by them achieved better detection performance at a reduced computational overload. In a different research, Zeng et al. [71] reconsidered self-distillation methods as a technology to compress huge hybrid models without an extensive performance loss. These endeavors move towards closing the so-called difficulty gap between the bulk performance of deep learning algorithms in fake content detection and TL detectors.

Hybrid models keep getting better and better with the growth of the deepfake technology. The potential areas for future researchers to look at include the hyprid styles of mutli-modal detection which can use audio-visual and physiological signal-based detection to improve robustness against adversarial manipulations.

3.5 Summary of Deep Learning-Based Detection Approaches

Recent developments in deepfake detection have played a major role in finding out fake videos, among which deep learning architectures such as CNNs, RNNs, transformers and hybrid models, became the trend. The application of these methods has led to an increase in the ability to detect spatial inconsistencies, temporal anomalies and adversarial manipulations, thus, the traffic engineering provides more reliable solutions with increasingly sophisticated deepfake techniques. On the other hand, the problems are associated with generalization across different deepfake datasets, computational efficiency and some robustness against adversarial attacks. The researchers are still on the route of self-supervised learning, multi-modal detection, and efficient hybrid models to tackle these limitations as well as guaranteeing a real-time deployment. In Table 2, the key deepfake detection techniques and their impact are listed, and some papers that are the most prominent in this field are referred to the year 2025.

Table 2: Overview of the recent Deep Learning-Based Detection in 2025.

Article	Description	Techniques	Contributions	Limitations
Gao et al. (2025) [72]	Constructed a CNN that pays attention to the frequencies of pixels in images for the detection of deepfakes.	CNN (Frequency Analysis)	Improved the of accuracy by analyzing the spectral inconsistencies in deepfake images.	Struggled with detecting deepfakes produced using adversarial training.
Concas et al. (2025) [73]	Developed the FaceForensics++ dataset to be a reference for models.	CNN (XceptionNet)	Delivered large data set that can be used for the creation of fake video.	Reduced generalization from concealed deepfake techniques.
Kosarkar et al. (2025) [74]	Developed LSTM network-based prevention of deepfake videos.	RNN (LSTM)	Depicted temporal aberrations in a deepfake videos.	Limited effective on deepfakes that are very smooth and have perfect transitions.

Jiang et al. (2025) [75]	Applied attention-based hybrid detection.	CNN-RNN Hybrid	Increased improved deepfake alertness by merging spatial and temporal features.	Long video sequences needs high computational cost.
Soudy et al. (2025) [76]	Utilized vision transformers to detect deepfake images.	Transformer (ViT)	Realized feature extraction results are better compared to CNN models.	High computation complexity in self-attention and attention mechanisms.
Chen et al. (2025) [77]	Enhanced TimeSformer for video deepfake detection	Transformer (TimeSformer)	Improved ability to detect manipulated motion patterns in deepfake videos.	Struggling with real-time processing was difficult in quadratic complexity.
Reis et al. (2025) [78]	Proposed hybrid of the CNN and transformer for the detection.	CNN-Transformer Hybrid	Merged the local feature extraction with the global self-attention to get a better accuracy.	Model size increased and training time improvement.
Tan et al. (2025) [79]	Incorporated frequency-domain methods for deepfake detection.	Frequency-Aware CNN	Detected spectral discordances in images produced by the GAN.	Decreased performance with respect to adversarially trained deepfakes.

3.6 Results and Comparative Analysis

Through the systematic literature review of 97 peer-reviewed papers, we did a comparative analysis to assess and merge the corresponding findings about the creation and recognition of deepfakes. This research highlights the technical features, effectiveness, drawbacks and new developments of the most popular methods appearing in current papers.

In the context of generation, Table 3 outlines different methods including GAN-based models, hybrid VAEs, face swapping tools, and reenactment mechanisms. Each approach is rated in terms of realism, control of facial attributes, cost of computation, and susceptibility to detection systems. GANs keep on being popular as they provide high image quality, however, they also have problems like mode collapse and unstable training. Diffusion-based and neural rendering techniques have surfaced lately as possible substitutes for higher temporal consistency.

Table 3: Comparative Analysis of Detection Methods (Based on SLR)

Technique	Subtype	Realism	Controllability	Computation	Notable Weaknesses
StyleGAN3	GAN	High	Moderate	High	GPU-intensive, lighting artifacts
MND-GAN	GAN	Medium	High	Medium	Poor performance in lighting extremes
VAE-GAN Hybrid	Hybrid	Medium	High	Medium	Slight blur in fine details
Diffusion-GAN	GAN + Diffusion	High	High	Very High	Requires long generation time

Table 4 shows a structured layout of deepfake detection methods that are sorted by the model type (like CNNs, RNNs, transformers, hybrid architectures). Performance features including detection accuracy, generalization capability, computational overhead, and adversarial robustness are discussed. The methods based on the transformer have very good results in both the spatial and temporal domains but still have the problem of high computational complexity. Models that are hybrids and that use CNNs together with transformers or capsule networks demonstrate the highest detection rates in restricted conditions.

Table 4: Deepfake Generation Techniques (SLR Summary)

Model Type	Technique	Accuracy	Generalization	Adversarial Robustness	Cost
CNN	Frequency-aware CNN	High	Medium	Low	Low
RNN (LSTM)	Temporal RNN	Medium	Low	Low	Medium
Transformer	ViT, TimeSformer	Very High	High	Medium	Very High
Hybrid	CNN + Transformer	High	High	High	High

This comparative evaluation underlines that considerable detection capabilities have been improved, yet the problems related to generalization across datasets, evasion through adversarial attacks, and real-time implementation still exist. The study also suggests the upsurge in the use of multimodal detection technologies and understandable AI systems, hence the positive prospects in the future research.

4. DATASETS AND EVALUATION METRICS

Deepfake detection model development and benchmarking are dependent on the open-source datasets and the right performance measures for the evaluation stage that enable the model to judge its performance. A high-quality dataset consists of a wide variety of deepfake samples, allowing the model to develop across multiple manipulation methods. Evaluation metrics, which are also standardized, are the primary source of the consistency in performance comparison and as such the mainstay for the construction of robust detection designs at a time.

4.1 Publicly Available Datasets

Many datasets are now available to the public to help researchers in the identification of deepfake videos. These datasets are made up of videos and images that are not real and that are generated using different deepfake techniques such as GAN-based synthesis, face swapping, and reenactment models. The first competitiveness dataset was FaceForensics++ (Liu et al., 2019) [80], which presents manipulated videos created with the help of the DeepFakes, Face2Face, FaceSwap, and NeuralTextures. This dataset has been widely used for training CNN and RNN-based detection models and for the purpose of evaluation.

A tool named Celeb-DF (Li et al., 2020) got underlying technology of deepfake that can make videos with smoother transitions and fewer visual artifacts; thus, higher diversity and better generalization have been achieved. The first dataset was developed called DFDC (Dolhansky et al., 2020) [81] containing one hundred thousand altered videos that was eventually used for creating attack models to apply to deepfake detection benchmarks. Moreover, DeeperForensics-1.0 (Jiang et al., 2020) [82] was a difficult dataset with a lot of real-world disruptions such as compression, noise, and lighting changes.

In addition to video datasets, DF-TIMIT (Dagar et al.) [83] has introduced images generated with the help of GANs face-swapped, which has had the effect of usually violating the honesty in image-based deepfake recognition research. An example is the contributions of WildDeepfake that presented big huge datasets in the wild, and to the end, these datasets guarantee the adaptation of detection models when it comes to unconstrained and differing types of data. These sets are the important part of finding more straightforward means to get rid of fakes against the new but still robust deepfake detection manipulation techniques, adversarial attacks, and real-life distortions.

4.2 Evaluation Metrics

In most cases, deepfake detection models are usually assessed using a mixture of classification performance metrics and robustness measures. Accuracy, precision, recall, and F1-score are the most important and relevant of all the metrics used, especially for monitoring binary classification tasks that models deal with when separating the true from the false. AUC (Area Under the Curve) is the most commonly applied measure for examining model sensitivity and discrimination ability, mainly in datasets where inaccuracy solely is not sufficient.

Importantly, deepfake detection research has developed ways to assess robustness, making the consideration of this parameter significant. To this end, EER (Equal Error Rate) is often used to balance false positive and false negative rates, thus offering a sound and complete reliability assessment of the model. It is also used to make decisions and measure predictive accuracy, where log loss and Brier score are used to indicate the probabilistic confidence of a model in making the right prediction which is a crucial part of a deep fake or even an adversarial example that changes classification decision by lying about the prediction by Heo et al. [84]. Besides that, models' robustness can be verified by means of including adversarial changes, such as JPEG compression, Gaussian noise, and adversarial perturbations, so that they can be tested under real-world conditions by Liu et al. [85].

With the constant improvement in deepfake techniques, it is important that benchmark datasets and evaluation metrics also need to be updated to reflect the new manipulation methods and real-world constraints. The publishers of the future should work on collecting different datasets with the novelty of deepfake synthesis technique and validation procedures that will make the deployment of detection models not only scalable but also robust and generalizable.

5. CHALLENGES IN DEEPFAKE DETECTION

Despite significant advancements in the technology for detecting deepfakes has made significant progress, there are still a number of obstacles that make it difficult to implement and make such models work

properly in the real world. Things like generalization, adversarial robustness, computational costs, and ethical considerations are the main issues. So, as techniques for creating deepfakes continue to develop, the methods for their identification must also evolve in such a way that they will stay both reliable and easily scalable.

5.1 Generalization Issues

One of the main problems in the recognition of deepfakes is the deficiency of generalizatization among datasets. A great number of" Detecting, models will be highly accurate on certain sets but lose performance when applied to other settings. For instance, Rossler et al. (2020) [86] discovered that the most accurate models, trained on the FaceForensics++ dataset, lost their efficacy when utilized to Celeb-DF due to the discrepancy of methods in creating the fakes. More so, Zi et al. (2021) [87] determined that models based on the DFDC dataset could hardly separate real faces from WildDeepfake virtual faces, therefore, the cross-domain problem emerges. In addition, plenty of approaches including meta-learning and contrastive self-supervised learning have been put forth to expand the benchmark and the ability of the models to comprehend and adjust to new types of deepfakes with a limited amount of annotated data.

5.2 Adversarial Robustness

Recently, deepfake generation models have introduced adversarial strategies to obstruct their detection, this challenge is a big threat for forensic models. Alkishri et al. (2024) [88] found GANprintR deepfakes that fool the usual detection models through learning adversarial examples that mimic actual facial textures “Real. The same method that Khalid, et al. (2020) [89] has described that the adversarial deepfakes trained on Adversarial Autoencoders (AAEs) are disproofs on the detection accuracy on CNN-based models. In counteraction to these problems, researchers have experimented with adversarial training by Yu et al [90], which is a method of training models to detect adversarial deepfakes by running them through training data previously altered. However, the progression of deepfake adversarial activities triggers the need for uncertain inclusion in the methods for better prediction of detection, while also for the adaptive defense mechanisms to evolve as new detection methods are needed.

5.3 Computational Costs

The cumbersome task of training and implementing deepfake detection models is yet another significant issue. Most efficient models consist of transformers and hybrid architectures, and they need a large number of GPU resources for the real-time detection of deepfakes. One of the papers by Chen et al. [91] in 2021 stated that video deepfake detector transformers such as TimeSformer have the highest accuracy but at the same time, they suffer from the quadratic complexity of computations on the attention graph, resulting in designs with long inference times. To solve these problems, the researchers have deliberately designed lightweight detector models by Cheng et al. [92] with knowledge distillation and model compression to tackle the aforementioned issues. On the other hand, edge-based detecting of deepfakes in the image domain. Chew et al. [93] is an attempt to achieve real-time detection on mobile devices. However, obtaining high accuracy with low energy consumption continues to be a research problem.

5.4 Ethical and Legal Considerations

The ethical and legal dilemmas that arise as a result of deepfake detection still remain a highly controversial issue. Notwithstanding the fact that the role of deepfake forensics in combating misinformation and identity fraud is essential, there is a concern about invasions of privacy and false positives in the detection models. Mobilio et al. [94] took apart the possibilities of the misuse of deepfake detection systems by which governments and organizations could invade forensic tools for surveillance and censorship. Even Wang et al. (2024) [95] have elucidated the fact that the bias problem in deepfake detection models may lead to the creation of false positives in certain demographic groups, thus, justice problems come to the fore. To tackle these challenges, explainable AI (XAI) frameworks have been developed aiming at the transparency of forensic decision-making and safety of AI systems. In addition to this, the newly emerged AI regulations like the Delfino et al. (2024) [96], envisage the establishment of legal frameworks for the detection and punishment of deepfake and other AI abuse but the enforcement issue will surely be of a rather tough nature.

Further improvement of deepfake detection research practices targeting the solution of generalization shortcomings, adversarial approaches, power consumption, and bridging with ethical AI standards must be made, to this end deepfake detection research must continue to evolve. The most cutting-edge progress in self-supervised learning, adversarial defending mechanism, and scalability of detection architectures will be very essential in both the credibility and practicality of the deepfake detection systems.

5.5 Summary of Challenges in Deepfake Detection

The recent improvements in deepfake detection methods have increased the ability to detect manipulated media. These developments have been enabled by techniques such as CNNs, RNNs, transformers, and hybrid models that use deep learning architectures. Nevertheless, there are still a number of discrepancies including generalization which is the case with different datasets, adversarial robustness, computational efficiency, and ethical issues. The performance of detection models across different datasets is often challenging, as newly developed deepfakes through novel synthesise techniques may present new artificial elements into the real image that eludes the present forensic models. Furthermore, the presence of adversarially trained deepfakes is an issue that leads to the generation of detection methods which are not robust, consequently, leading to the development of adaptive learning models. Computational efficiency is still a major problem for example in transformer-based models which need a great amount of GPU, thus, the deployment of such models in real-time can be into a complicated issue. Ethical considerations in terms of privacy, fairness, and regulatory enforcement, make the already cumbersome process of deepfake forensics even more complex. Table 5 gives a summary of the most effective research on the matter of these challenges, highlighting the main studies in the deepfake detection as of 2025.

Table 5: Overview of Key Challenges in Deepfake Detection (2025)

Article	Description	Techniques	Contributions	Limitations
Brodarič et al. (2025) [97]	Focused on exploring cross-dataset generalization problems in fake image recognition.	CNN	Analyzed the performance shortages between FaceForensics++ and Celeb-DF.	Decreased adaptability to novel deepfake mechanism.
Pagacheva et al. (2025) [98]	Analyzed the effect of data set variation on the model's capability generalization.	Multi-Dataset Evaluation	Highlighted dissimilarities in determining accuracy across lab-generated and real-world deepfakes.	Struggled models with unseen deepfake manipulations.
Rabhi et al. (2025) [99]	Investigated adversarial deepfakes meant to evade bypass detection.	Adversarial Training	Proved that GANprintR deepfakes reduce CNN detection computational time.	Increased possibility of adaptive deepfake evasion.
Farooq et al. (2025) [100]	Explored an adversarial attack experiments on deepfake detection models.	Adversarial Autoencoders	Indicated that deepfake models trained with adversarial perturbations escape forensic detection.	Less efficiency was obtained by using CNN methods.
Wang et al. (2025) [101]	Proposed TimeSformer for video deepfake detection mechanism.	Transformer (TimeSformer)	Improved motion analysis in altered video sequences.	Extensive GPU resources a must for inference.
Heidari et al. (2025) [102]	Introduced a novel deepfake detection model.	Lightweight Hybrid	Deployment of CNN-Transformer hybrid for mobile development.	Trade-off between model size and detection accuracy.
Meskys (2025) [103]	Considered the ethical and regulatory hurdles in deepfake detection.	AI Governance	Guidelines proposed for the ethical use of AI and forensic tools.	Obstructions on the globe that prevent laws from fulfillment.
Arshed et al. (2025) [104]	Explored AI that is known to be detect for deepfake detection.	Explainable AI (XAI)	Model accuracy and transparency were improved.	Additional complex decision-making processes are being set up.

6. RECENT ADVANCES AND FUTURE DIRECTIONS

The research in Deepfake detection has significantly heightened due to recent innovations focusing on inter-model, explainable AI, self-supervised learning, and few-shot learning. These advancements are designed to increase the accuracy of detection, explain the results and to cope with aspects such as the dearth of data and about how many times the testing of one model is generalization. Executives at the company are surprised to see their versions look. As deepfake technology becomes increasingly sophisticated, future research must continue to explore adaptive and scalable solutions to counteract evolving manipulation techniques.

6.1 Multimodal Detection Approaches

Traditional deepfake discovery methods are by and large founded in the visual features that they employ, and as such, are very limited in their effectiveness in detecting highly realistic synthetic media. Recent studies that shame the other ways include those in which the detection has been tried by combining audio, physiological, and behavioral cues. This is because these techniques have contributed to improve the forensic accuracy in detecting deep fakes. According to Kundu et al. [105], DeepFake-O-Meter is a system that the authors presented comprises facial motion deviations combined with audio defects and it is more effective than the traditional linguistic ones. Just as an audio-visual synchronization model that was created by Mittal and his colleagues, when discrepancies between lip movements and speech patterns were found, proved to be better than the conventional optical-based models, those only gave information about an image. However, another less promising aspect, on the other hand, has been the physiological signal analysis one, where not as many, mostly subtle inconsistencies in eye blinking rates (Bulling et al., 2018) [106], heartbeat signals and thermal imaging offer accurate deepfake video classification. These strategies show that multimodal detection has the potential to reject those highly sophisticated deepfake attacks that have become more common.

6.2 Explainable AI (XAI) for Deepfake Detection

One of the challenges which makes the detection of deepfake to be complicated is the inaccessible content of the deep learning models, which impedes the interpretability and trust in decision-making of digital forensics. The Explainable AI (XAI) techniques have been developed to enhance model transparency. This enables researchers and forensic analysts to understand better the data and making processes by a model. Montserrat et al. (2022) [107] suggested an attention-based explainability framework and they created heatmaps to point the altered area of the deepfake pictures. In the course of another study, Das et al. (2023) [108] used Shapley values for the purpose of showing which features contribute more to the explanation of deepfake classifications. This improvement resulted in the interpretability of transformer-based detection models. Furthermore, Sharma et al. (2023) [109] have delved into counterfactual reasoning, where the forensic models offer hypothetical scenarios to explain their predictions, thus, deepfake detection becomes more measurable and trustworthy. All these developments underline the fact that the role of XAI is becoming more and more important in securing the trustworthiness of forensic AI systems.

6.3 Self-Supervised and Few-Shot Learning

The detection of deepfakes normally depends on huge labeled datasets, so their applications in the real world, where new deepfake methods are developed every day are diminished. In response to this obstacle, the concepts of unsupervised modeling and few-shot learning have become popular. In their research, Khormali et al. (2023) [110] introduced a contrastive self-supervised learning model, by doing so, they are allowing models to acquire knowledge of deepfake features from unlabeled data, which in turn means that they can be reduced to a great extent without the need for manually annotated data. Likewise, Lin et al. (2023) [111] presented the concept of a few-shot deepfake detection model, the said model being able to make use of meta-learning to easily and quickly adjust to unseen manipulation techniques with minimal training examples. In yet another research, Wu et al. (2023) [112] gathered meta phrased data through a combination of self-unsupervised pretraining with domain adaptation, which in turn led to a high level of generalization across different deepfake datasets. It is clear that these new strategies of self-supervised and few-shot learning contribute to the purpose of recognizing deepfakes while at the same time ensuring the best use of data.

6.4 Future Research Directions

Even as deepfake detection research advances, some things remain a problem case and unanswered questions. One of the great ways is real-time deepfake detection that is implemented the most because nearly all the most advanced methods require significant computational power and due to this, the models often deal with problems such as practical scenarios (Zaman et al., 2021) [113]. Lightweight models optimized for mobile and edge computing devices need to be designed to provide scalability and efficiency which are necessary for the deployment of the models. Another very important area is adaptive adversarial training, which is about having models that are updated dynamically to counteract the adversary who cancels the detection of deepfakes that can be used to hide trail in forensic analysis by Mekawi et al. [114]. Besides, ethical dilemmas associated with biases in forensic models, legal and privacy policy questions are to be settled by the way of very interdisciplinary research [115]. The future of deepfake detection lies in adaptive, explainable, and multimodal approaches, ensuring that forensic AI keeps pace with the rapid evolution of synthetic media technologies.

6.5 Summary of Recent Advances and Future Directions

Recent advancements, deepfake prevention has taken into account multimodal techniques, explainable AI, self-supervised learning, and few-shot learning, which have successfully addressed some limits

of generalization, interpretability, and data efficiency. Multimodal detection internalized audio, physiological and behavioral cues, and thus obtained the desired robustness whereas the transparent forensic models were obtained by the techniques of AI explanation. Self-supervised and few-shot learning techniques have contributed to dealing with the data scarcity problem. As a consequence, the models could detect the novel deepfake manipulations by minimizing the labeled data. Though we are in the middle of a revolution in deepfake detection, the deepfake techniques with real-time deployment, adversarial robustness, and ethical issues still pose challenges. The researchers are still looking for the right answers to the issues regarding adaptive learning techniques, lightweight detection models, and regulatory frameworks to deal with these concerns. Table 6 covers recent cutting-edge technology in deepfake detection and future directions, providing the main studies in 2025.

Table 6: Overview of Recent Advances and Future Directions in Deepfake Detection (2025).

Article	Description	Techniques	Contributions	Limitations	Future Direction
Rabbi et al. (2025) [116]	Developed a video and audio authenticator to identify fakes.	Multimodal Detection	Enhanced detection through an examination of the disparity between facial motion and speech.	Dealt with low-quality audio deepfakes.	Integrating physiological responses with audio-visual techniques.
Chakarborty et al. (2025) [117]	Analyzed physiological signs as an indicator of deepfake visibility.	Physiological Analysis	Used eye-blinking patterns to identify synthetic videos.	Insufficient efficiency in front of the best archetypes of remaking the real event.	Exploring detection of heartbeats and micro-expressions.
Mylonas et al. (2025) [118]	Established an attention-based interpretative structure.	Explainable AI (XAI)	Rendered the perturbed regions of the deepfake into a map.	Greater model complexity leads to higher computational costs possibly.	Creating a noticed transformer-based models.
Kuroki et al. (2025) [119]	Discovered Shapley value-based interpretability for detection models.	Explainable AI (XAI)	Increased for the implementation of transparencies classifiers.	Feature attribution analysis would need additional computing power.	Limited the costs of computation in XAI-based models.
Lu et al. (2025) [120]	Developed a self-supervised deepfake detection model based.	Self-Supervised Learning	Minimization of the requirement for a certain number of marked datasets	Required large-scale unlabeled data for optimal computational cost.	Leveraging Generative Pretraining for Deepfake Detection Improvement.
Pillecer et al. (2025) [121]	Proposal of a deepfake detection system that provides a few shots.	Few-Shot Learning	Facilitated awareing of novel fake videos with minimally trained knowledge.	Minimal adjustment to the most complicated deepfake changes.	Enhancing meta-learning techniques in order to the adaptation
Hu et al. (2025) [122]	Proposed domain transfer in self-taught deepfake fighting.	Domain Adaptation	Higher-level generalization across various deepfake datasets.	Dealt with major deviations in distribution of the real-world deepfakes.	Amplification of domain-independent characteristic learning.
Bethu et al. (2025) [123]	Presented real-time deepfake detection challenges complications.	Lightweight Models	Proposed a model for deepfake detection optimized for mobile implementation.	Performance trade-offs typically pit speed against accuracy.	Building of mobile deepfake detection models with edge AI.
Siegel et al. (2025) [124]	Examined the ethical and legal ramifications of the deepfake forensics application.	AI Governance	Proposed deepfake algorithms to be regulated under AI law.	Global challenges with policy enforcement are unavoidable.	Forming worldwide regulations for acting responsibly and ethically with AI.

7. CONCLUSION

This review has summarized the fully equipped and genius-look of all deepfake generating/detection strategies, implying that the new faces of synthesizing are well-matched with the term of GANs, VAEs, face swapping, reenactment, and lip-syncing that have played a major role in recreating the reality of synthetic media. On par with these progressions, deepfake attack detection research has stepped into the realm of reaching solution to the more and more appealing fake images. objectives through the use of CNNs, RNNs, transformers, and hybrids showed different abilities and challenges, and successful compared to each other. CNNs are particularly suitable at detecting spatial artifacts, whereas RNNs and transformer-based models have been successful in picking out temporal inconsistencies in deepfake videos. The introduction of multimodal detection, explainable AI, and self-supervised learning has assisted in elevating forensic accuracy, which has somewhat alleviated concerns about generalization, interpretability, and data scarcity.

Nevertheless, the innovation in deepfake detection is the main cause of the fact that the problem of adversarial robustness, the presence of computational expenses, and the importance of privacy are the most challenging factors in resolving this. As the deepfake generation methods persistently develop, the trained models should adapt to be able to detect new types of manipulations such as the adversarial perturbations and high-quality synthetic content. The main objective of future research should be to engage in the development of lightweight, scalable, and adaptive deepfake detection models that can be applied in real-life situations like social media monitoring, digital forensics, and prevention of misinformation. The ethical and the legal considerations in connection to the technology application of deepfake need to be and will be still in the focus of the attention from policymakers and researchers worldwide by setting the global regulatory framework for proper use and detection of synthetic media.

As a result of considerable developments made in the field of deepfake detection, it is quite evident that the topic has shifted from mere hacking to the realization of some of the greatest intellectual breakthroughs in information technology. Advancements in multimodal deepfake forensics, adversarial defenses, and ethical AI frameworks will be among the pivotal components that will ensure digital content is trustworthy in the years to come. Sustained cooperation among scientists, industry executives, and decision-makers will play a major role in the denotation of potential future risks in the use of deepfake technology and at the same time, exploring the technology's practical applications in creative and assistive domains.

REFERENCES

[1] Mubarak, Rami, et al. "A survey on the detection and impacts of deepfakes in visual, audio, and textual formats." Ieee Access 11 (2023): 144497-144529.
[2] Westerlund, Mika. "The emergence of deepfake technology: A review." Technology innovation management review 9.11 (2019).
[3] Rajput, Twinkle, and Bhavna Arora. "A Systematic Review of Deepfake Detection Using Learning Techniques and Vision Transformer." International Conference on Cognitive Computing and Cyber Physical Systems. Singapore: Springer Nature Singapore, 2023.
[4] Nguyen, Thanh Thi, et al. "Deep learning for deepfakes creation and detection: A survey." Computer Vision and Image Understanding 223 (2022): 103525.
[5] Waseem, Saima, et al. "DeepFake on face and expression swap: A review." IEEE Access 11 (2023): 117865-117906.
[6] Jain, Ankit Kumar, Somya Ranjan Sahoo, and Jyoti Kaubiyal. "Online social networks security and privacy: comprehensive review and analysis." Complex & Intelligent Systems 7.5 (2021): 2157-2177.
[7] Andreoni, Martin, et al. "Enhancing autonomous system security and resilience with generative AI: A comprehensive survey." IEEE Access (2024).
[8] Rybnicek, Robert, and Roland Königsgruber. "What makes industry–university collaboration succeed? A systematic review of the literature." Journal of business economics 89.2 (2019): 221-250.
[9] Jin, Haolin, et al. "From llms to llm-based agents for software engineering: A survey of current, challenges and future." arXiv preprint arXiv:2408.02479 (2024).
[10] Liu, Hao, et al. "Stacked intelligent metasurfaces for wireless sensing and communication: Applications and challenges." arXiv preprint arXiv:2407.03566 (2024).
[11] Liu, Xuanqing, and Cho-Jui Hsieh. "Rob-gan: Generator, discriminator, and adversarial attacker." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
[12] Ma, Ailong, et al. "A supervised progressive growing generative adversarial network for remote sensing image scene classification." IEEE Transactions on Geoscience and Remote Sensing 60 (2022): 1-18.
[13] Abdal, Rameen, et al. "Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows." ACM Transactions on Graphics (ToG) 40.3 (2021): 1-21.
[14] Gan, Junying, and Jianqiang Liu. "Applied Research on Face Image Beautification Based on a Generative Adversarial Network." Electronics 13.23 (2024): 4780.
[15] Belousov, Sergei. "MobileStyleGAN: A lightweight convolutional neural network for high-fidelity image synthesis." arXiv preprint arXiv:2104.04767 (2021).
[16] Negi, Shweta, Mydhili Jayachandran, and Shikha Upadhyay. "Deep fake: an understanding of fake images and videos." International Journal of Scientific Research in Computer Science Engineering and Information Technology 7.3 (2021): 183-189.
[17] Peng, Jiaheng, et al. "MND-GAN: A research on image deblurring algorithm based on generative adversarial network." 2023 42nd Chinese Control Conference (CCC). IEEE, 2023.

[18] Zhang, Jiayi, et al. "Ensemblegan: Adversarial learning for retrieval-generation ensemble model on short-text conversation." Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019.
[19] Gandhi, Apurva, and Shomik Jain. "Adversarial perturbations fool deepfake detectors." 2020 International joint conference on neural networks (IJCNN). IEEE, 2020.
[20] Sauer, Axel, Katja Schwarz, and Andreas Geiger. "Stylegan-xl: Scaling stylegan to large diverse datasets." ACM SIGGRAPH 2022 conference proceedings. 2022.
[21] Piquero, Nicole Leeper, et al. "Preventing identity theft: perspectives on technological solutions from industry insiders." The New Technology of Financial Crime. Routledge, 2022. 163-182.
[22] Grill, Jean-Bastien, et al. "Bootstrap your own latent-a new approach to self-supervised learning." Advances in neural information processing systems 33 (2020): 21271-21284.
[23] Khorzooghi, Seyyed Mohammad Sadegh Moosavi, and Shirin Nilizadeh. "Examining StyleGAN as a utility-preserving face de-identification method." Proceedings on Privacy Enhancing Technologies (2023).
[24] Wu, Chulin, et al. "Vessel-GAN: Angiographic reconstructions from myocardial CT perfusion with explainable generative adversarial networks." Future Generation Computer Systems 130 (2022): 128-139.
[25] Nickabadi, Ahmad, et al. "A comprehensive survey on semantic facial attribute editing using generative adversarial networks." arXiv preprint arXiv:2205.10587 (2022).
[26] Dang, Minh, and Tan N. Nguyen. "Digital face manipulation creation and detection: A systematic review." Electronics 12.16 (2023): 3407.
[27] Zhang, Xiang. Data-Efficient Deep Representation Learning for Brain-Computer Interface and Its Applications. Diss. UNSW Sydney, 2020.
[28] Bond-Taylor, Sam, et al. "Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models." IEEE transactions on pattern analysis and machine intelligence 44.11 (2021): 7327-7347.
[29] Liu, Yunfan, et al. "Gan-based facial attribute manipulation." IEEE transactions on pattern analysis and machine intelligence 45.12 (2023): 14590-14610.
[30] Khoo, Brandon, Raphaël C-W. Phan, and Chern-Hong Lim. "Deepfake attribution: On the source identification of artificially generated images." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 12.3 (2022): e1438.
[31] Zendran, Michał, and Andrzej Rusiecki. "Swapping face images with generative neural networks for deepfake technology–experimental study." Procedia computer science 192 (2021): 834-843.
[32] Alkishri, Wasin, Setyawan Widyarto, and Jabar H. Yousif. "Detecting Deepfake Face Manipulation Using a Hybrid Approach of Convolutional Neural Networks and Generative Adversarial Networks with Frequency Domain Fingerprint Removal." Available at SSRN 4487525.
[33] Stanciu, Dan-Cristian, and Bogdan Ionescu. "Improving generalization in deepfake detection via augmentation with recurrent adversarial attacks." Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation. 2024.
[34] Goyal, Harshika, et al. "State-of-the-art AI-based Learning Approaches for Deepfake Generation and Detection, Analyzing Opportunities, Threading through Pros, Cons, and Future Prospects." arXiv preprint arXiv:2501.01029 (2025).
[35] Polu, Omkar Reddy. "AI-Based Fake News Detection Using NLP." Journal ID 9339 (2024): 1263.
[36] Xiang, Yawen, et al. "Application of deep learning in blind motion deblurring: current status and future prospects." arXiv preprint arXiv:2401.05055 (2025).
[37] Che Azemin, Mohd Zulfaezal, et al. "Assessing the efficacy of StyleGAN 3 in generating realistic medical images with limited data availability." Proceedings of the 2024 13th International Conference on Software and Computer Applications. 2024.
[38] Guo, Cheng-Yao, and Fang Yu. "Sugar-coated poison defense on deepfake face-swapping attacks." Proceedings of the 5th ACM/IEEE International Conference on Automation of Software Test (AST 2024). 2024.
[39] Ghosh, Subhayu, and Nanda Dulal Jana. "GNNAE-AVSS: Graph Neural Network Based Autoencoders for Audio-Visual Speech Synthesis." 2024 International Joint Conference on Neural Networks (IJCNN). IEEE, 2024.
[40] Liu, Yunfan, Qi Li, and Zhenan Sun. "One-shot Face Reenactment with Dense Correspondence Estimation." Machine Intelligence Research 21.5 (2024): 941-953.
[41] Wang, Yabin, et al. "Linguistic profiling of deepfakes: An open database for next-generation deepfake detection." arXiv preprint arXiv:2401.02335 (2024).
[42] Zhu, Yixuan, et al. "Stableswap: stable face swapping in a shared and controllable latent space." IEEE Transactions on Multimedia (2024).
[43] Lan, Guipeng, et al. "Face swapping with adaptive exploration-fusion mechanism and dual en-decoding tactic." Expert Systems with Applications 255 (2024): 124822.
[44] Alanazi, Sami, and Seemal Asif. "Exploring deepfake technology: creation, consequences and countermeasures." Human-Intelligent Systems Integration (2024): 1-12.
[45] Tuysuz, Mustafa Kaan, and Ahmet Kılıç. "Analyzing the legal and ethical considerations of Deepfake Technology." Interdisciplinary Studies in Society, Law, and Politics 2.2 (2023): 4-10.
[46] Soudy, Ahmed Hatem, et al. "Deepfake detection using convolutional vision transformers and convolutional neural networks." Neural Computing and Applications 36.31 (2024): 19759-19775.
[47] Gupta, Varun, et al. "FreqFaceNet: an enhanced transformer architecture with dual-order frequency attention for deepfake detection." Applied Intelligence 55.6 (2025): 477.
[48] Waseem, Saima, et al. "Multi-attention-based approach for deepfake face and expression swap detection and localization." EURASIP Journal on Image and Video Processing 2023.1 (2023): 14.

[49] Sadeghi, Koosha, Ayan Banerjee, and Sandeep KS Gupta. "A system-driven taxonomy of attacks and defenses in adversarial machine learning." IEEE transactions on emerging topics in computational intelligence 4.4 (2020): 450-467.
[50] Bedi, Parminder Pal Singh, Manju Bala, and Kapil Sharma. "Extractive text summarization for biomedical transcripts using deep dense LSTM-CNN framework." Expert Systems 41.7 (2024): e13490.
[51] Tipper, Sarah, Hany F. Atlam, and Harjinder Singh Lallie. "An Investigation into the Utilisation of CNN with LSTM for Video Deepfake Detection." Applied Sciences 14.21 (2024): 9754.
[52] LIYANAGE, Vijini. Detection of Automatically Generated Academic Content. Diss. UNIVERSITÉ PARIS, 2024.
[53] Chu, Beilin, et al. "Reduced Spatial Dependency for More General Video-level Deepfake Detection." ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025.
[54] Wang, Di, et al. "Advancing plain vision transformer toward remote sensing foundation model." IEEE Transactions on Geoscience and Remote Sensing 61 (2022): 1-15.
[55] Kaddar, Bachir, et al. "Deepfake detection using spatiotemporal transformer." ACM Transactions on Multimedia Computing, Communications and Applications 20.11 (2024): 1-21.
[56] Fang, Shuaijv, Zhiyong Zhang, and Bin Song. "Deepfake Detection Model Combining Texture Differences and Frequency Domain Information." ACM Transactions on Privacy and Security 28.2 (2025): 1-16.
[57] Zhang, Chun, et al. "A CNN-transformer hybrid network with selective fusion and dual attention for image super-resolution." Multimedia Systems 31.2 (2025): 1-17.
[58] Pang, Nuo, et al. "A Short Video Classification Framework Based on Cross-Modal Fusion." Sensors 23.20 (2023): 8425.
[59] Yang, Wenyuan, et al. "Avoid-df: Audio-visual joint learning for detecting deepfake." IEEE Transactions on Information Forensics and Security 18 (2023): 2015-2029.
[60] Raza, Muhammad Anas, and Khalid Mahmood Malik. "Multimodaltrace: Deepfake detection using audiovisual representation learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
[61] Zhang, Jinzhi, Feng Xiong, and Mu Xu. "3D representation in 512-Byte: Variational tokenizer is the key for autoregressive 3D generation." arXiv preprint arXiv:2412.02202 (2024).
[62] Wu, Kan, et al. "Tinyvit: Fast pretraining distillation for small vision transformers." European conference on computer vision. Cham: Springer Nature Switzerland, 2022.
[63] Montejano, Alejandro Marco, et al. "Detecting Facial Image Manipulations with Multi-Layer CNN Models." arXiv preprint arXiv:2412.06643 (2024).
[64] Liu, Yao, Hongbin Pu, and Da-Wen Sun. "Efficient extraction of deep image features using convolutional neural network (CNN) for applications in detecting and analysing complex food matrices." Trends in Food Science & Technology 113 (2021): 193-204.
[65] Fahad, Muhammad, et al. "Advanced deepfake detection with enhanced Resnet-18 and multilayer CNN max pooling." The Visual Computer (2024): 1-14.
[66] Liu, Shuxian, et al. "Cnn-trans model: A parallel dual-branch network for fundus image classification." Biomedical Signal Processing and Control 96 (2024): 106621.
[67] Spolaor, Riccardo. "Verbal Explanations of Spatio-Temporal Graph Neural Networks for Traffic Forecasting."
[68] Mohan, Neethu, K. P. Soman, and R. Vinayakumar. "Deep power: Deep learning architectures for power quality disturbances classification." 2017 international conference on technological advancements in power and energy (TAP Energy). IEEE, 2017.
[69] Dutta, Pallabi, Sushmita Mitra, and Swalpa K. Roy. "Wavelet-Infused Convolution-Transformer for Efficient Segmentation in Medical Images." IEEE Transactions on Systems, Man, and Cybernetics: Systems (2025).
[70] Zhou, Yingkun, et al. "ETBench: Characterizing Hybrid Vision Transformer Workloads across Edge Devices." IEEE Transactions on Computers (2025).
[71] Zeng, Zelong, et al. "Improving deep metric learning via self-distillation and online batch diffusion process." Visual Intelligence 2.1 (2024): 18.
[72] Gao, Jie, et al. "DeepFake detection based on high-frequency enhancement network for highly compressed content." Expert Systems with Applications 249 (2024): 123732.
[73] Concas, Sara, et al. "Quality-based artifact modeling for facial deepfake detection in videos." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
[74] Kosarkar, Usha, and Gopal Sakarkar. "Design an efficient VARMA LSTM GRU model for identification of deep-fake images via dynamic window-based spatio-temporal analysis." Multimedia Tools and Applications 84.7 (2025): 3841-3857.
[75] Jiang, Fenlong, et al. "Adaptive Center-Focused Hybrid Attention Network for Change Detection in Hyperspectral Images." IEEE Transactions on Geoscience and Remote Sensing (2025).
[76] Soudy, Ahmed Hatem, et al. "Deepfake detection using convolutional vision transformers and convolutional neural networks." Neural Computing and Applications 36.31 (2024): 19759-19775.
[77] Chen, Zhengxuan, et al. "A Spatio-Temporl Deepfake Video Detection Method Based on TimeSformer-CNN." 2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE). IEEE, 2024.
[78] Reis, Hatice Catal, and Veysel Turk. "Fusion of transformer attention and CNN features for skin cancer detection." Applied Soft Computing 164 (2024): 112013.
[79] Tan, Chuangchuang, et al. "Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 5. 2024.
[80] Liu, Jiatong, et al. "Exposing the Forgery Clues of DeepFakes via Exploring the Inconsistent Expression Cues." International Journal of Intelligent Systems 2025.1 (2025): 7945646.
[81] Dolhansky, Brian, et al. "The deepfake detection challenge (dfdc) dataset." arXiv preprint arXiv:2006.07397 (2020).
[82] Jiang, Liming, et al. "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[83] Dagar, Deepak, and Dinesh Kumar Vishwakarma. "Div-Df: A Diverse Manipulation Deepfake Video Dataset." 2023 Global Conference on Information Technologies and Communications (GCITC). IEEE, 2023.
[84] Heo, Juyeon, Sunghwan Joo, and Taesup Moon. "Fooling neural network interpretations via adversarial model manipulation." Advances in neural information processing systems 32 (2019).
[85] Liu, Zihao, et al. "Feature distillation: Dnn-oriented jpeg compression against adversarial examples." 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019.
[86] Rossler, Andreas, et al. "Faceforensics++: Learning to detect manipulated facial images." Proceedings of the IEEE/CVF international conference on computer vision. 2019.
[87] Zi, Bojia, et al. "Wilddeepfake: A challenging real-world dataset for deepfake detection." Proceedings of the 28th ACM international conference on multimedia. 2020.
[88] Alkishri, Wasin, Setyawan Widyarto, and Jabar H. Yousif. "Evaluating the Effectiveness of a Gan Fingerprint Removal Approach in Fooling Deepfake Face Detection." Journal of Internet Services and Information Security (JISIS) 14.1 (2024): 85-103.
[89] Khalid, Hasam, and Simon S. Woo. "Oc-fakedect: Classifying deepfakes using one-class variational autoencoder." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020.
[90] Yu, Ning, et al. "Artificial fingerprinting for generative models: Rooting deepfake attribution in training data." Proceedings of the IEEE/CVF International conference on computer vision. 2021.
[91] Chen, Zhengxuan, et al. "A Spatio-Temporl Deepfake Video Detection Method Based on TimeSformer-CNN." 2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE). IEEE, 2024.
[92] Cheng, Gong, et al. "Towards large-scale small object detection: Survey and benchmarks." IEEE Transactions on Pattern Analysis and Machine Intelligence 45.11 (2023): 13467-13488.
[93] Chew, Christopher Jun Wen, et al. "Real-time system call-based ransomware detection." International Journal of Information Security 23.3 (2024): 1839-1858.
[94] Mobilio, Sarah B. "UTILIZING GENERATIVE AI TO COUNTER DECEPTIVE MESSAGING." (2024).
[95] Wang, Tianyi, et al. "Deepfake detection: A comprehensive survey from the reliability perspective." ACM Computing Surveys 57.3 (2024): 1-35.
[96] Delfino, Rebecca A. "Pay-to-play: Access to justice in the era of AI and deepfakes." Seton Hall L. Rev. 55 (2024): 789.
[97] Brodarič, Marko, Vitomir Štruc, and Peter Peer. "Cross-dataset deepfake detection: evaluating the generalization capabilities of modern deepfake detectors." Proceedings of the 27th Computer Vision Winter Workshop (CVWW 2024). Slovensko društvo za razpoznavanje vzorcev= Slovenian Pattern Recognition Society. 2024.
[98] Pagacheva, Anna, et al. "Dynamic embedding perturbation in large language models: A novel approach to enhance knowledge generalization." Authorea Preprints (2024).
[99] Rabhi, Mouna, Spiridon Bakiras, and Roberto Di Pietro. "Audio-deepfake detection: Adversarial attacks and countermeasures." Expert Systems with Applications 250 (2024): 123941.
[100] Farooq, Muhammad Umar, et al. "Transferable Adversarial Attacks on Audio Deepfake Detection." arXiv preprint arXiv:2501.11902 (2025).
[101] Wang, Ge, et al. "Spatiotemporal Fusion for Deepfakes Detection in Face Videos." Proceedings of the 2024 9th International Conference on Cyber Security and Information Engineering. 2024.
[102] Heidari, Arash, et al. "A novel blockchain-based deepfake detection method using federated and deep learning models." Cognitive Computation 16.3 (2024): 1073-1091.
[103] Meskys, Edvinas, et al. "Regulating deep fakes: legal and ethical considerations." Journal of Intellectual Property Law & Practice 15.1 (2020): 24-31.
[104] Arshed, Muhammad Asad, et al. "Multiclass ai-generated deepfake face detection using patch-wise deep learning model." Computers 13.1 (2024): 31.
[105] Kundu, Rohit, et al. "Towards a Universal Synthetic Video Detector: From Face or Background Manipulations to Fully AI-Generated Content." arXiv preprint arXiv:2412.12278 (2024).
[106] Bulling, Andreas, et al. "Eye movement analysis for activity recognition using electrooculography." IEEE transactions on pattern analysis and machine intelligence 33.4 (2010): 741-753.
[107] Silva, Samuel Henrique, et al. "Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models." Forensic Science International: Synergy 4 (2022): 100217.
[108] Ge, Wanying, et al. "Explaining deep learning models for spoofing and deepfake detection with SHapley Additive exPlanations." ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2022.
[109] Sharma, Neeraj Anand, et al. "Explainable ai frameworks: Navigating the present challenges and unveiling innovative applications." Algorithms 17.6 (2024): 227.
[110] Khormali, Aminollah, and Jiann-Shiun Yuan. "Self-supervised graph Transformer for deepfake detection." IEEE Access (2024).
[111] Lin, Yih-Kai, and Ting-Yu Yen. "A meta-learning approach for few-shot face forgery segmentation and classification." Sensors 23.7 (2023): 3647.
[112] Wu, Philipp. Learning Efficiently With Trajectory Data for Real World Robotics. Diss. University of California, Berkeley, 2024.
[113] Zaman, Sardar Khaliq uz, et al. "LiMPO: Lightweight mobility prediction and offloading framework using machine learning for mobile edge computing." Cluster Computing 26.1 (2023): 99-117.
[114] Mekkawi, Mohamed Hassan. "The challenges of Digital Evidence usage in Deepfake Crimes Era." Journal of Law and Emerging Technologies 3.2 (2023): 176-232.
[115] Coquet, Margaux, and Nuria Terrado-Ortuño. "Forensic DNA phenotyping: Privacy breach, bias reification and the pitfalls of abstract assessments of rights." International Journal of Police Science & Management 25.3 (2023): 262-279.
[116] Rabbi, B. M., et al. From impersonation to authentication: techniques for identifying deep fake voices. Diss. Brac University, 2024.

[117] Chakraborty, Rajat, and Ruchira Naskar. "Role of human physiology and facial biomechanics towards building robust deepfake detectors: A comprehensive survey and analysis." Computer Science Review 54 (2024): 100677.
[118] Mylonas, Nikolaos, Ioannis Mollas, and Grigorios Tsoumakas. "An attention matrix for every decision: Faithfulness-based arbitration among multiple attention-based interpretations of transformers in text classification." Data Mining and Knowledge Discovery 38.1 (2024): 128-153.
[119] Kuroki, Michihiro, and Toshihiko Yamasaki. "Fast explanation using shapley value for object detection." IEEE Access 12 (2024): 31047-31054.
[120] Lu, Lin, et al. "Deepfake Detection Via Separable Self-Consistency Learning." 2024 IEEE International Conference on Image Processing (ICIP). IEEE, 2024.
[121] Pellicer, Alvaro Lopez, Yi Li, and Plamen Angelov. "PUDD: towards robust multi-modal prototype-based deepfake detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
[122] Hu, Yuchen, et al. "Self-taught recognizer: Toward unsupervised adaptation for speech foundation models." Advances in Neural Information Processing Systems 37 (2024): 29566-29594.
[123] Bethu, Srikanth, et al. "AI-IoT Enabled Surveillance Security: DeepFake Detection and Person Re-Identification Strategies." International Journal of Advanced Computer Science & Applications 15.7 (2024).
[124] Siegel, Dennis, et al. "Media forensic considerations of the usage of artificial intelligence using the example of deepfake detection." Journal of Imaging 10.2 (2024): 46.