Proceedings of the 3rd International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/117/2025.20955

A Review of Deepfake Detection Techniques

Junyi Chen[1,a,*], Minghao Yang[2,b], Kaishen Yuan[3,c]

[1]Department of School of Computer Science and Engineering, South China University of Technology, Guangzhou, China

[2]Department of School of Ande, Xi'an University of Architecture and Technology, Xi'an, China

[3]Department of School of international education, Nanjing Institute of Technology, Nanjing, China

a. cs202230140505@mail.scut.edu.cn, b. 744870774@qq.com, c. x00202221244@njit.edu.cn

**corresponding author*

Abstract: With the development of deepfake technology, the use of this technology to forge videos and images has caused serious privacy and legal problems in society. In order to solve these problems, deepfake detection is required. In this paper, the generation and detection techniques of deepfakes in recent years are studied. First, the principles of deepfake generation technology are briefly introduced, including Generative Adversarial Networks (GAN) based and autoencoder. Then, this paper focuses on the detection techniques of deepfakes, classifies them based on the principles of each method, and summarizes the advantages and limitations of each method. At the end of the paper, several key points for the future development of deepfake detection technology are proposed: enhancing the generalization ability and robustness of deepfake detection methods, developing active defensive algorithms and multimodal fusion detection, establishing research communities and data sharing platforms, and improving social legislation and judicial education. This paper argues that the future deepfake detection algorithm will be more accurate, which can further maintain the authenticity of network information and social stability.

Keywords: deepfake, deep learning, detection introduction

1. Introduction

Deepfake usually refers to the technology of using deep learning to splice a person's voice, facial expressions and body movements into false content[1]. With the development of deep fake technology, the images and videos forged using this technology are becoming more and more realistic, making it difficult to distinguish the real from the fake. At present, many APP developers use deep fake technology to create many interesting functions, such as AI face-changing and voice simulation. The emergence of these functions has attracted public attention and is extremely popular. However, many criminals have maliciously used deep fake technology to create false information and conduct fraudulent activities, resulting in many vicious incidents. In May 2023, a fake photo of an explosion near the Pentagon appeared on social media, causing widespread circulation and ultimately leading to a sharp drop in the U.S. stock market. In February 2024, an employee of the Hong Kong branch of a multinational company was invited to attend a multi-person video conference initiated by the chief financial officer of the headquarters. He made multiple transfers as required, totaling HK$200 million. He later inquired with the headquarters and found out that he had been cheated. The police investigation revealed that in the so-called video conference in this fraud case, only the victim was a real person, and the rest of the participants were fraudsters who used AI face-changing technology to disguise themselves as company insiders. Many criminals not only use deep fake technology to commit fraud, but also use photos and videos generated by the technology to interfere in political activities, causing extremely bad effects. At present, governments of various countries have taken certain regulatory measures on deep fake technology. Such as the United States' Deep Fake Reporting Act, the European Union's European Artificial Intelligence Methods Regulation, and my country's "Regulations on the Ecological Governance of Network Information Content" and "Measures for the Management of Generative Artificial Intelligence Services (Draft for Comments)", which also implement certain supervision on deep fake technology.

This article refers to relevant papers on deep fakes since 2020. It first briefly introduces deep fake generation technology, then focuses on deep fake detection technology, and divides the existing deep fake detection technology into four categories according to the principle: image forensics-based methods (for example, Cozzolino et al. proposed using the PRNU pattern to detect tampering[2]), physiological signal-based methods (Agarwal et al. use physiological signal features for analysis[3]), Generative Adversarial Network (GAN) image feature-based methods (Marra et al. use specific patterns or features left by GAN in the image generation process to uniquely identify image inconsistencies[4]), and data-driven methods (Nguyen et al. designed a method based on capsule networks to detect forged images or videos[5]).

This article will discuss the advantages and limitations of existing deep fake detection technology, point out several key points for the future development of deep fake detection technology and the challenges it may face in the future, and provide directional guidance for future research on deep fake detection technology.

2. Introduction to deepfake generation technology

Deepfake technology relies primarily on advanced deep learning architectures, with generative adversarial networks (GANs) and autoencoders at their core. A large number of videos and images are required to train the model and optimize it. Deepfake videos require audio processing. Speech synthesis, speech pattern analysis, and speech conversion systems are major components in this process.

2.1. GAN

In GAN, two neural networks are involved in the adversarial process - a network of generators creates synthetic content, while discriminators try to distinguish between real and fake content. Through iterative training, generators are becoming increasingly adept at making convincing fakes that trick discriminators into producing more realistic images. It is shown in Figure 1.

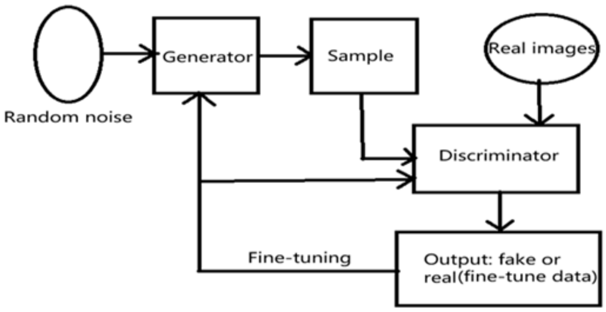

Figure 1: workflow diagram of GANs (Picture credit : Original)

2.2. Autoencoder

Autoencoders produce images by first compressing (encoding) them into compact representations of their essential features, and then reconstructing (decoding) them. When creating a deepfake, the encoder captures the basic features of the source face, and the decoder reconstructs these features onto the target face. This encoder-decoder architecture enables the system to understand and transmit facial expressions, movements, and lighting conditions. The neural network continuously learns and improves through backpropagation, adjusting its internal parameters (weights and biases) to minimize the difference between what is generated and what is needed. This is shown in Figure 2.

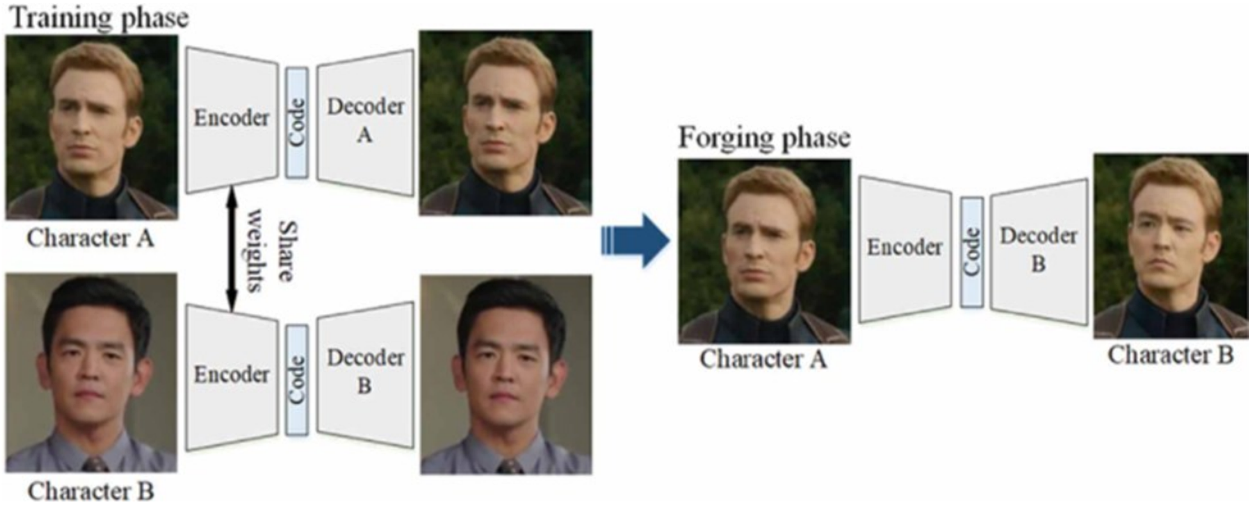

Figure 2: Autoencoders in a face swapping process [6]

The process could first train an autoencoder to learn a compact representation of facial features from real-world images. Then, use this learned representation as input for GAN.

3. Deepfake Detection Technology

With the advancement of deepfake technology, there is a proliferation of pseudo-videos containing altered faces, synthetic voices, and even AI-generated characters on the internet both domestically and internationally. The application of Deepfake technology has led to numerous serious social issues, including the infringement of citizens' personal privacy, the creation of fake news, and the manipulation of public opinion. These issues not only pose a threat to individual rights but may also have profound effects on political elections and social stability. Therefore, research on deepfake detection technology has become particularly crucial. This section will provide a review of some key technologies in the field of deepfake detection, focusing on the detection technology of deepfake videos in the first four parts, and in the fifth part, a comprehensive evaluation of these technologies will be conducted. The development of these technologies is essential for maintaining the authenticity and security of online information.

3.1. Traditional Image Forensics-Based Methods

In the context of the continuous advancement of deepfake technology, traditional image forensics-based methods remain effective due to their solid foundation in signal processing and statistical feature analysis. These methods rely on image frequency domain characteristics and statistical properties to identify tampering, such as detecting local noise, image quality, fingerprints, lighting shadows, and wrinkles, to identify copy-move, splicing, and removal of image tampering behaviors.

Since deepfake videos are essentially a sequence of forged images, this provides an application scenario for traditional image forensics techniques. In this field, Cozzolino et al. proposed using the unique mark left by each individual device on all its photos, known as the Photo Response Non-Uniformity (PRNU) pattern, to detect tampering [2]. More succinctly, due to imperfections in device manufacturing, each device that captures photos has a unique imprint, akin to human fingerprints, hence the method is a detection method using "device fingerprints." Figure 3 is the distribution of forged image noise traces caused by inconsistent operations as presented by Cozzolino et al. in their paper.

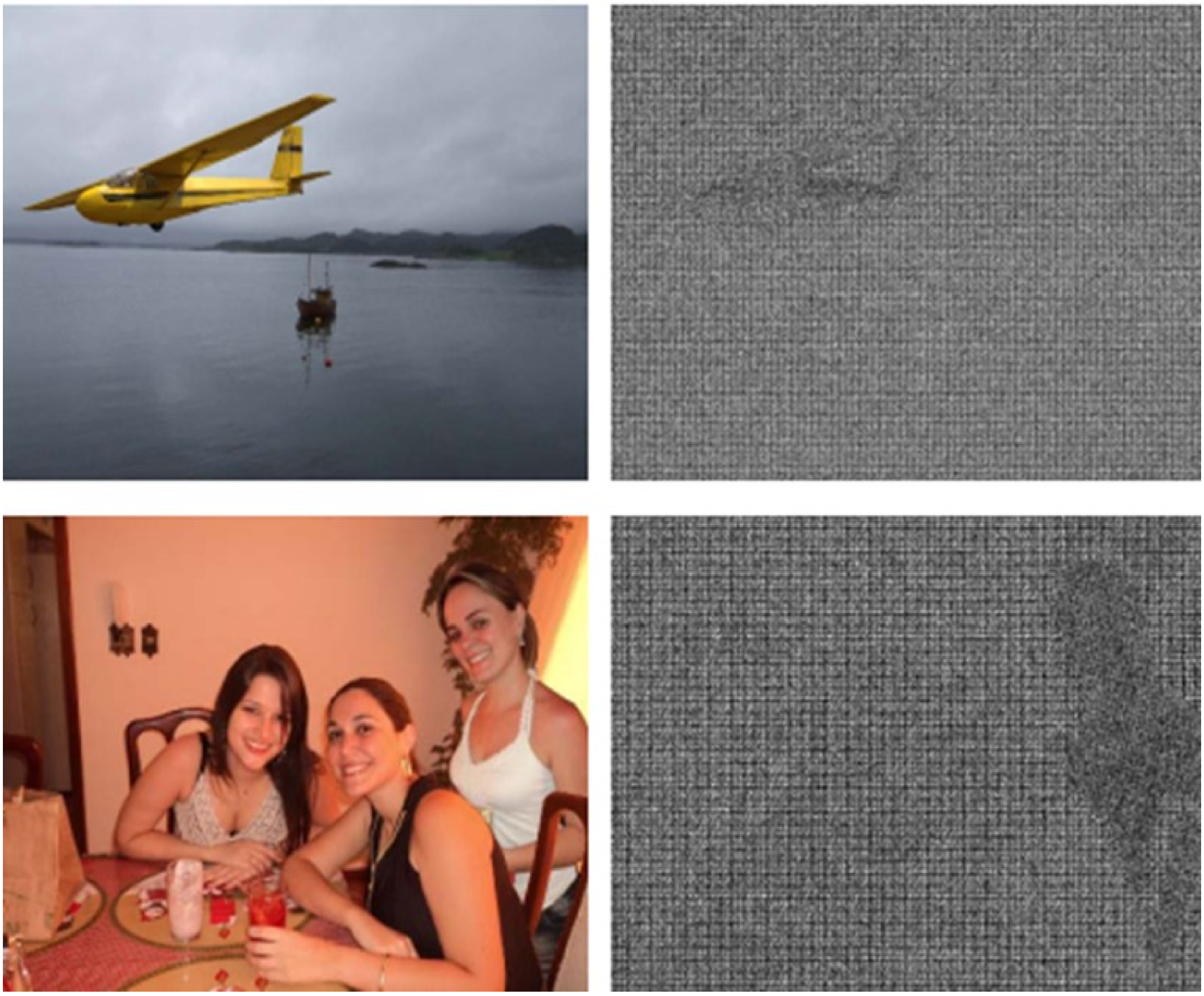

Figure 3: Different extracted noise patterns due to inconsistent operations [2]

Additionally, early-generation networks did not handle details well, leading to images that needed to be forged often exhibiting flaws such as inadequate resolution. For instance, after a Deepfake algorithm generates a human face, it often needs to be artificially synthesized and replaced with a new face due to its insufficient resolution. This leaves traces of human intervention, and image tampering recognition technology is developed based on this vulnerability. Li et al.'s team mainly focuses on the fact that due to limitations in computational resources and production time, Deepfake algorithms can only synthesize faces with limited resolution and must undergo affine transformations to match the configuration of the source face and be integrated into the source video[7]. The transformed face will inevitably have some inconsistencies with the environment in the original video, and this distortion leaves unique artifacts in the generated Deepfake videos. Li et al. detect Deepfakes by identifying these artifacts, and Figure 4 will illustrate the method by which general Deepfake methods produce images [7].

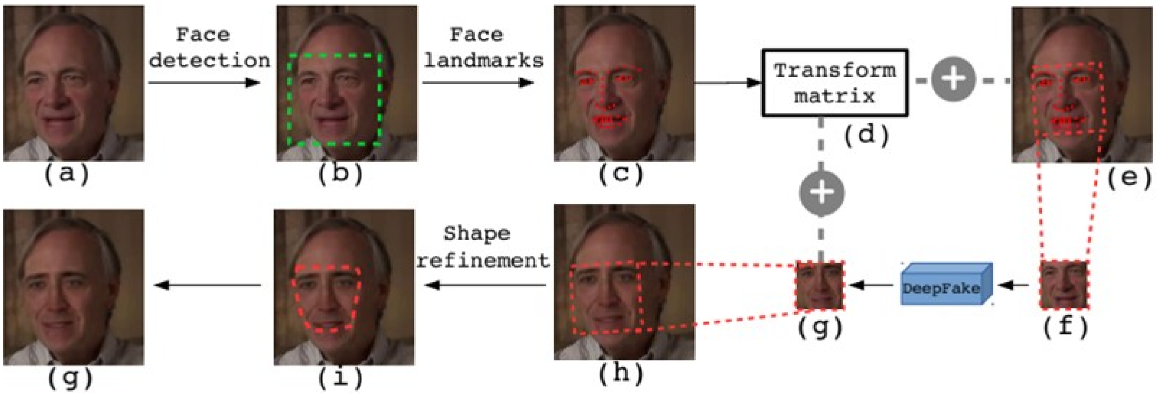

Figure 4: Overview of the DeepFake production line diagram[7]. (a) The source image. (b) The green box is the detected facial area. (c) The red dots are facial landmarks. (d) Computing the transformation matrix to distort the facial area in (e) to the normalized area (f). (g) The synthesized face from the neural network image. (h) Synthesizing the distorted face using the same transformation matrix. (i) Post-processing including boundaries to smooth the synthesized image. (j) The final synthesized image.

Although traditional image forensics-based techniques have made progress in identifying image tampering, they still face challenges when dealing with the new generation of deepfake videos. Deepfake content often undergoes complex post-processing, such as compression and resizing, which increases the difficulty of detection. Image-level forensics techniques primarily identify local anomalies, but they may be insufficient in deepfake video detection. When the forged content has no significant differences from the synthesized images, the further synthesized videos are more likely to evade detection. Moreover, although detection methods based on tampering traces perform well on some datasets, these datasets often contain products of early-generation technologies and are not suitable for training with modern techniques. With the advancement of image generation technology, the resolution and detail of modern forged images have improved, and adversarial processing measures such as adding noise have reduced the effectiveness of detection methods. Therefore, this technology is not always directly applicable to identifying the ever-evolving deepfake videos.

3.2. Methods Based on Physiological Signal Characteristics

In addition to the traditional image forensics and basic image signal processing methods mentioned earlier, individuals' physiological signals are also often used to identify deepfakes. During the production of forged videos, the precise simulation of real human physiological responses is often not achieved, leading to differences between fake and real human behaviors. Therefore, researchers have begun to explore the use of physiological feature signals as a basis for detecting the authenticity of videos.

Agarwal et al., when conducting physiological signal feature analysis, first categorized existing AI-synthesized face forgery methods into the following three types: face swap (replacing the face of a person appearing in a video with another person's face, usually aligning and replacing the entire face), lip-sync (making the person in the video move their lips according to predetermined audio, usually forging the target's lip area), and puppet-master (making the person in the video make a given facial expression, including head movement, usually requiring the establishment of a 3D model of the person's face and forging the lip area) [3]. Based on the above research, Agarwal's team found that different people have distinct patterns of facial expressions and head movements when speaking, and the three forgery methods mentioned above disrupt this pattern, resulting in facial tampering. After tampering, the facial muscle movements are inconsistent, leading to unnatural expressions or even making significantly unreasonable expressions that do not belong to this class (such as national leaders), which can be used to further determine deepfakes. Similarly, Ciftci et al. asserted that biological signals hidden in portrait videos, such as heartbeat, pulse, and blood volume patterns, can be used as implicit descriptions of authenticity because they are neither spatially nor temporally preserved as fake content but are constantly changing according to human metabolism[8]. Therefore, this technology captures facial temperature information through infrared imaging or other temperature-sensing devices and then uses machine learning algorithms to analyze these temperature distribution patterns. By comparing the facial temperature distribution of the person in the video with known patterns of real people, inconsistencies can be identified, thus detecting deepfake videos. This can be considered a fusion method. Figure 5 will show the physiological signal feature sample frames of original and forged images analyzed by Ciftci et al.

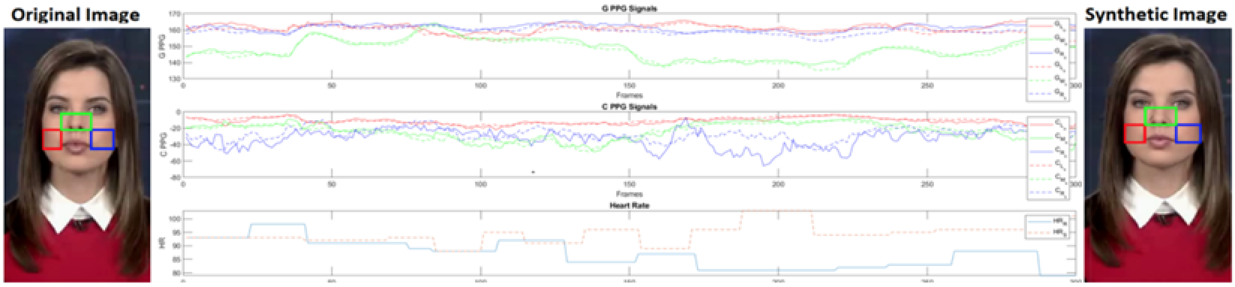

Figure 5: Biometric Signal Analysis Chart. Green (G* - top) and chrom-PPG (C* - middle) from left (*L - red), middle (*M - green), and right (R - blue) regions. Heart rate (HR - bottom) as well as original (left, *O - solid line) and synthesized (right, **S - dashed line) [8].

From the current perspective, many detection methods based on physiological signal characteristics target the shortcomings of deepfake technology, such as the inability to truly synthesize a "human." However, as deepfake technology continues to advance, it begins to incorporate more complex physiological features, such as more natural blinking patterns, head movements, lip swing amplitude, and speech consistency, and can even analyze human gaze points to complete reasonable predictive analysis of human eye movements. This makes detection methods based on original physiological signal characteristics gradually lose their effectiveness. In addition, detection technologies that rely on biological signals such as pulse and heart rate, which are less likely to be simulated, may have their accuracy reduced due to compression and other processing steps that videos undergo during transmission. This means that to effectively combat deepfakes, physiological signal detection technology needs to be continuously updated to adapt to the new developments in forgery technology.

3.3. Methods Based on GAN Image Features

Deepfake detection methods based on Generative Adversarial Networks (GANs) primarily revolve around the inherent characteristics of images generated by GANs. Firstly, GAN feature recognition technology focuses on identifying unique patterns or features that GANs may introduce during the image generation process. These features may be reflected in visual content, pixel distribution, or frequency characteristics, providing a basis for distinguishing GAN images from real images. Secondly, intermediate layer feature analysis methods study specific features captured by the intermediate layers of GANs to identify differences between real and GAN-generated images. Additionally, image quality assessment techniques compare differences in quality metrics such as texture, clarity, or noise levels between GAN-generated images and real images to detect deepfake content.

According to research by Marra et al., the specific patterns or features left by GANs during the image generation process can be referred to as "artificial fingerprints," used to uniquely identify image inconsistencies, including but not limited to[4]:

Pixel-level features: Images generated by GANs may exhibit different distribution characteristics at the pixel level compared to natural images, which may be reflected in color, brightness, or contrast.

Frequency features: In frequency domain analysis, GAN images may show anomalies in certain frequency components, which can be used to differentiate GAN images from real images.

Texture features: GAN-generated images may have limitations in texture generation, leading to differences in detailed textures compared to real images, which can serve as a basis for detection.

Pattern consistency: GANs may reuse certain patterns or components when generating images, and this consistency may appear in multiple generated images, becoming a recognizable feature.

Training data bias: If a GAN is trained on a specific dataset, it may learn the biases within the dataset and replicate these biases in the generated images.

Artifacts in the generation process: GANs may introduce artifacts during the image generation process, such as unnatural edges or transitions, which can serve as clues for detection.

3.4. Data-driven Methods

The development of data-driven deepfake detection technology has benefited from the availability of big data, advancements in computing power, progress in deep learning algorithms, and the increasing demand for automated feature extraction. The rapid growth of the internet and digital media has provided a vast amount of image and video data, offering resources for the training of deep learning models. The development of modern computing hardware such as GPUs and TPUs has enhanced the ability to process large-scale datasets, making the training of deep learning models more efficient. The success of convolutional neural networks (CNNs) in image recognition and classification tasks has prompted researchers to apply them to deepfake detection.

Nguyen et al. designed a method based on capsule networks for detecting forged images or videos[5]. Capsule networks can learn features at different levels in images, such as lighting and wrinkles, thus more accurately capturing the inconsistencies of forged content. The advantage of this method lies in its ability to simulate the human visual system's processing of the relationships between parts and the whole of objects, enhancing the robustness of detection. Figure 6 is an overview of Nguyen et al.'s capsule network method, and Figure 7 is the basic design of this team's capsule network.

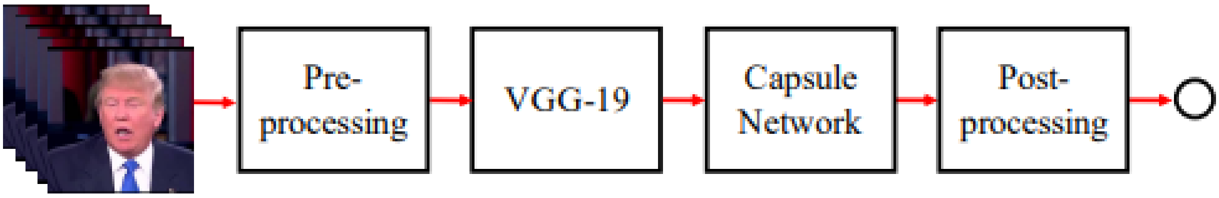

Figure 6: Capsule Network Method Overview Diagram [5]

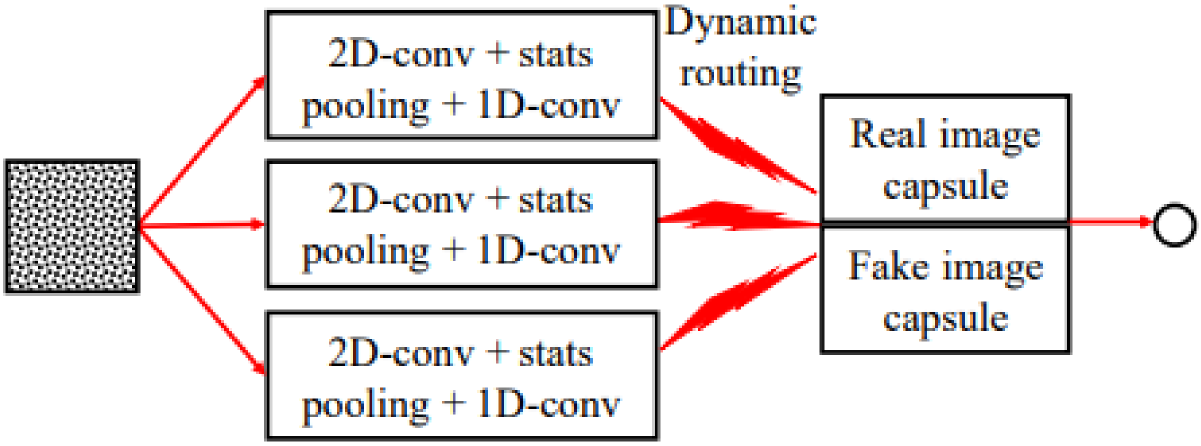
2D-conv + stats pooling + 1D-conv
2D-conv + stats pooling + 1D-conv
2D-conv + stats pooling + 1D-conv
Dynamic routing
Real image capsule
Fake image capsule

Figure 7: Basic Design Diagram of Capsule Network [5]

At the same time, Rossler et al. proposed Faceforensics++, a deep learning-based model specifically designed for detecting facial manipulation in deepfake videos. By training on a large dataset of both manipulated and real facial data, the model can effectively distinguish between manipulated and non-manipulated facial images. This method optimizes the model to minimize the distances between real samples while maximizing the distances between fake samples, thereby achieving effective classification.

Data-driven deepfake detection technology, leveraging its ability to learn from and extract useful features from large amounts of data, shows great potential for future development. As deep learning algorithms continue to be refined, computing power significantly increases, and large-scale, diverse datasets become more abundant, these technologies are gradually becoming a powerful tool for identifying and preventing deepfake content. It is foreseeable that future data-driven models will achieve rapid and accurate detection of deepfake videos and images through more efficient learning algorithms, more advanced model architectures, and more powerful computing resources.

3.5. Summary of Detection Techniques

As previously discussed, deepfake video detection technology has evolved into four major categories of detection algorithms, each with its own advantages and limitations depending on the application scenario. Table 1 will present the known strengths and weaknesses of these algorithms.

Table 1: Summary of Technologies Based on the Previous Discussion

Methods	Strengths	Weaknesses
Traditional Image Forensics-Based Methods	These methods are technologically mature and rely on interpretable features such as local noise analysis and image quality assessment.	While these methods are primarily oriented toward images and may not fully consider the dynamic characteristics of video content, they are also sensitive to image preprocessing such as compression.

Table 1: (continued).

Methods Based on Physiological Signal Characteristics	These methods can capture real human physiological characteristics, such as blink frequency, head posture, eye saccades, and pulse, which are difficult to simulate in deepfake videos.	With the advancement of deepfake technology, including the incorporation of more natural blinking patterns, the effectiveness of such methods may be diminished.
Methods Based on GAN Image Features	Focusing on identifying image features generated by Generative Adversarial Networks (GANs), these features may be common in images generated by GANs.	Depending on specific GAN structures, there may be insufficient generalization capabilities for different GAN generators.
Data-driven Methods	Through training on large-scale datasets, complex feature representations can be learned, which leads to high detection accuracy for various types of deepfake videos.	Sensitive to the distribution of training data, these methods may lack robustness against unseen types of forgeries and are sensitive to video compression and quality variations.

4. Challenges and Prospects

With the development of deepfake technology, research, and application of detection technology are facing unprecedented challenges. Future research directions need to be expanded in multiple dimensions to cope with the evolving forgery techniques and the ever-expanding training methods and forgery databases. Here are several key points for the future development of deepfake detection technology that this article believes in:

1. Enhance Generalization Ability: Research should explore various types of deepfakes to find common features, such as generator fingerprints, facial and lip consistency differences, and lighting differences on body parts, to enhance the model's adaptability to unknown forgery types.
2. Strengthen Robustness: Detection algorithms need to adapt to complex real-world conditions such as compression, noise, and lighting. Model robustness can be improved through data preprocessing and adversarial training.
3. Proactive Defense Algorithms: Research on using adversarial sample technology and video tracking technology to achieve proactive defense against unknown forged data.
4. Multimodal Fusion Detection: Develop detection technologies that can handle the fusion of audio and image data to cope with more realistic forgery effects.
5. Establish Research Communities and Data-Sharing Platforms: Concentrate data resources, and establish unified communities and data-sharing platforms to promote resource-sharing and academic cooperation.
6. Judicial Legislation and Social Education: Establish a legal system to punish malicious creators and distributors, train journalists to identify fake videos, and reduce the spread of forged videos.

5. Conclusion

This paper has reviewed deepfakes and their detection technologies, analyzing detection methods based on image forensics, physiological signals, GAN characteristics, and data-driven approaches. Despite certain advancements, current technologies still need to be enhanced in terms of generalization and robustness. The paper suggests that future research on deepfake detection technology requires efforts on multiple levels, including technical, legal, and social dimensions. It is necessary to focus on extracting common features, enhancing model adaptability, developing multimodal detection technologies, and building data-sharing platforms. It is anticipated that with algorithm optimization, increased computing power, and enriched datasets, detection technologies will become more accurate and robust. Strengthening legal frameworks and social education will support technological development, build a comprehensive defense against deepfakes, and maintain the authenticity of online information and social stability. The authors of this paper also hope that in the near future, more accurate and robust deepfake detection technologies can be realized.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Qu, Z., Yin, Q., Sheng, Z., et al. (2024). A Review of Active Defense Technologies for Deepfake Facial Recognition. Journal of Image and Graphics, 29(2): 318-342.

[2] Cozzolino, D., Verdoliva, L. (2019). Noiseprint: A CNN-Based Camera Model Fingerprint. IEEE Transactions on Information Forensics and Security, PP(99):1-1.

[3] Boháek, M., Farid, H. (2022). Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms.Proceedings of the National Academy of Sciences of the United States of America. 119(48): e2216035119-e2216035119.

[4] Marra, F., Gragnaniello, D., Verdoliva, L., et al. (2018). Do GANs leave artificial fingerprints.

[5] Nguyen, H. H., Yamagishi, J., Echizen, I. (2019). Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos. IEEE.

[6] Zhao, Z., Hui, P., Lu, W. W. (2021). Multi-Layer Fusion Neural Network for Deepfake Detection. International Journal of Digital Crime and Forensics, 13(4): 26-39.

[7] Li, Y., Lyu, S. (2018). Exposing DeepFake Videos By Detecting Face Warping Artifacts.

[8] Ciftci, U. A., Demir, I., Yin, L. (2024). Fakecatcher: Detection of Synthetic Portrait Videos Using Biological Signals: US 202117143093. US2021209388A1[2024-11-25].