A Chronological Review of Deepfake Detection: Techniques and Evolutions

JINGJING RAO[1] and TETSUTARO UEHARA[2]

[1] Graduate School of Information Science and Engineering.
Ritsumeikan University, Japan
[2] College of Information Science and Engineering
Ritsumeikan University, Japan

Abstract. Since the development of deep learning technology, various new technologies have emerged one after another, greatly facilitating our daily lives. However, the development of these technologies has also brought some troubles, among which Deepfake technology is a typical example. Deepfake technology is mainly used to generate false pictures and videos, or modify real pictures and videos to achieve the purpose of deception. In the early days of this technology, people could often distinguish the authenticity with the naked eye. However, as the technology matures, the generated pictures and videos become more and more realistic, and many criminals have begun to use this technology to commit economic fraud, produce illegal pornographic content, distort political facts and other illegal acts. In order to better understand the importance of Deepfake detection and its related technologies, this article sorts out the main Deepfake detection technologies from 2018 to 2024. We briefly explain the various methods mentioned in the work and organize them into a table form. At the same time, we also set up a series of Q&A sessions, the purpose of which is to comprehensively introduce Deepfake technology and its detection methods from multiple perspectives, so as to help readers fully understand the latest developments and challenges in this field.

Keywords: Deepfake, Detection, State-of-the-Art,GANs, Deeplearning, Dataset, Traditional method

1 INTRODUCTION

In recent years, the rapid development of artificial intelligence has brought numerous opportunities to society and challenges and risks, becoming a focal point of attention for countries worldwide. Mainly, Deepfake technology, a significant branch of artificial intelligence, has rapidly evolved and found widespread application. This technology has demonstrated immense potential in creative media, film production, and personalized entertainment and has sparked extensive ethical and legal controversies.

Deepfake technology combines deep learning and fake images, videos, audio, people, or scenes. As shown in Fig.1, Deepfake can be traced back to the MCP model in 1943[1], an early milestone in artificial intelligence. The development of Deepfake includes face replacement technology in the 1994 movie "Forrest Gump," which not only promoted technological innovation but also laid the foundation for public acceptance. By 2012, deep learning technology caused a research boom and brought revolutionary changes to image processing. In 2014, generative adversarial networks (GANs)[2] introduced a key technology for generating realistic deepfake content. In 2017, the first appearance of the term "deepfake" began to attract widespread attention from the public and academia. In 2018, the first political news about deepfakes highlighted the application of this technology in the political field and triggered numerous ethical and legal discussions. By 2024, the Deepfake incident in South Korea pushed these discussions to the world, causing people to think deeply about issues such as politics, pornography, privacy, and security, highlighting the far-reaching impact and challenges brought by deepfakes.

In the face of the challenges brought by deepfakes, many researchers have proposed a series of countermeasures. This work will focus on the detection methods for deepfake

David C. Wyld et al. (Eds): SIGML, CNSA, MoWiN, CCSEIT, NIAI, AIAP – 2025
pp. 23-49, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2024.150202

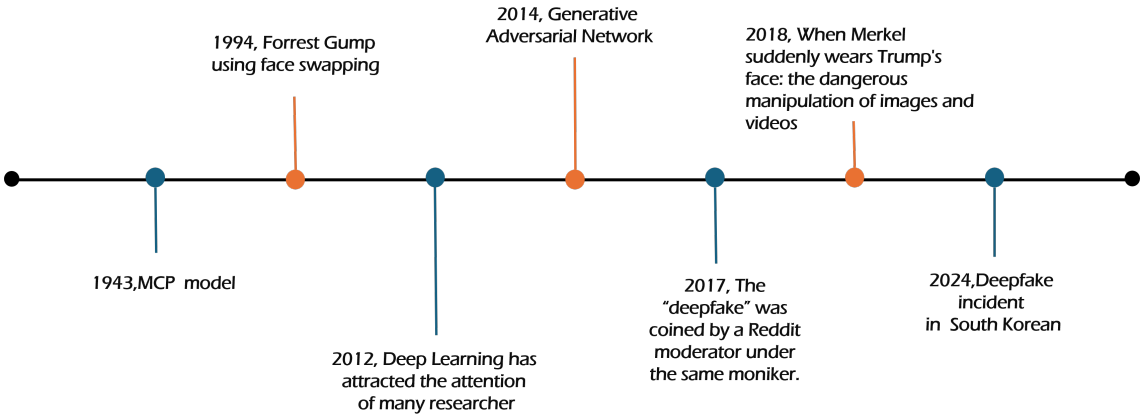

Fig. 1. The development of deepfake

images and videos. By analyzing the deepfake detection technology from 2018 to the present, we can observe the progress of technology, the evolution of methods, and the changes in data. Since 2018, research on deepfake detection has increased yearly, and 2024 has witnessed a booming development in this field. In addition, this work summarizes the database of deepfake images and videos for reference. This review will show how these detection technologies have developed to cope with the increasingly complex Deepfake generation technology.

This work is organized as follows: Section 1 introduces the background of Deepfake and its impact on society. Section 2 details the different types of Deepfake and the process of forged information generation. Section 3 lists the databases currently available for Deepfake detection, covering video and image materials. Section 4 is a method review that summarizes the research results of Deepfake detection from 2018 to the present. Section 5 discusses legal issues. Section 6 sets up a series of Q&As about deepfakes. Section 7 discusses and outlines the development process in recent years and explores future research directions. This structure aims to comprehensively analyze the development of Deepfake technology and evaluate its social impact while guiding future research.

2 DEEPFAKE

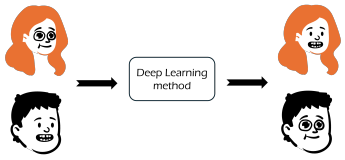

Fig. 2. Face swapping

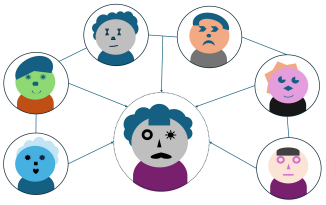

Fig. 3. Face synthesis (Solid circles represent existing face images, and dotted circles represent new face synthesized by deep learning methods)

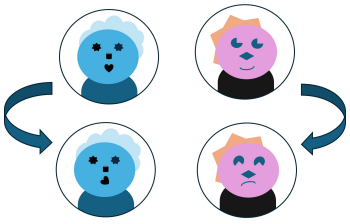

Fig. 4. Facial attribute and expression manipulation

In [3] 2018, Deepfake is defined as a technology that allows the face of one individual to be replaced with that of another in visual content called *Face swapping* as illustrated in Fig.2. As technology has advanced, the applications of Deepfake have expanded beyond merely replacing faces; it now encompasses the creation of entirely new images and videos. Fig.3 presents an instance of *Face synthetic* generation, where the solid circle indicates an actual face image, and the dotted circle depicts a newly generated face crafted by deep learning techniques from a composite of features derived from various individuals. This synthesis includes not only facial features but also hairstyles and body structures.

Another common category of Deepfakes is *facial attribute and expression manipulation* as shown in Fig.4. This technique adjusts an individual's expression by modifying specific facial features, such as eyes and eyebrows, thereby changing the emotion conveyed by an image or video. For example, an image that originally showed a happy expression could be adjusted to express anger, changing the emotional context of the content and the audience's perception.

Deepfake technology usually involves a combination of multiple neural networks, among which the Generative Adversarial Networks (GANs) [2] that emerged in 2014 have become the core technology of Deepfake. GAN consists of two parts: one is the generator (G), and the other is the discriminator (D). The task of the generator is to create realistic fake samples, while the task of the discriminator is to determine whether the sample is real. The two networks play an adversarial game during the training process. The generator constantly learns how to make images that are difficult to distinguish from real samples, while the discriminator strives to improve its ability to identify these fake samples. After the training is completed, the discriminator is usually discarded, and only the generator is used to produce high-quality, realistic images. This method is effective in generating detailed and highly realistic visual content.

Due to the powerful capabilities of Generative Adversarial Networks (GANs), researchers have developed a variety of improved GAN methods, including conditional generative adversarial networks (CGANs)[4], cycle-consistent generative adversarial networks (CycleGANs)[5], ProGANs[6], StyleGANs[7], BigGANs[8], and GauGANs[9]. Each variant is optimized for specific applications and challenges, such as finer image detail control, cross-domain image conversion, high-resolution image generation, and highly realistic natural scene rendering.

In addition, there is a Deepfake technology based on the encoder-decoder architecture, which is used to manipulate or generate highly realistic visual and audio content. The encoder is a neural network that processes input data (such as images or video frames) and compresses it into a low-dimensional representation, namely the latent space or hidden state. This process involves extracting essential features from the input while discarding redundant information. The encoder is able to capture the identity, expression, and other relevant attributes of the source material through learning.

The decoder is another neural network that receives the compressed data from the encoder and reconstructs it into the desired output form. In the Deepfake generation process, the decoder is responsible for creating the final manipulated image or video by applying the learned features to the target face or scene. The main task of the decoder is to ensure that the output content is seamless and naturally retains the characteristics of the target.

In encoder-decoder architecture, variational autoencoders (VAEs)[10] are particularly popular. VAEs introduce the concept of a probability distribution, allowing the generator to explore the latent space, thereby producing new and diverse outputs, which makes the generated Deepfake content more realistic and varied.

At the same time, the successive emergence of diffusion models[11] also provides significant support for Deepfake technology. This model generates visual content by gradually guiding the transformation from a noise distribution to a data distribution. Its excellent detail capture and high-quality image generation capabilities significantly improve the naturalness and realism of synthesized images, making the generated false content more difficult to distinguish.

3 DATASET

We screened and introduced some of the more popular and featured Deepfake datasets from 2018 to 2024. These datasets are widely used in the research of deepfake detection and generation technology, and each dataset has its unique contributions and application scenarios. As shown in Table 1, these datasets provide a wide range of coverage from video face swapping to audio tampering.

For example, FaceForensics++[15] and DFDC (DeepFake Detection Challenge)[16] provide a large number of high-quality video samples, especially for training and testing more advanced detection algorithms. Datasets such as Celeb-DF[17] and DeeperForensics[18] focus on generating high-quality Deepfake videos that are difficult to detect to test the boundaries of detection technology. ForgeryNet[19] and DeepFake MNIST+[20] provide a platform for multi-task learning, aiming to improve algorithms' versatility by identifying multiple types of tampering.

In addition, with the continuous advancement of technology, culturally and language-specific datasets such as KoDF[21] have begun to appear, focusing on generating and detecting Deepfakes for specific groups. Audio tampering detection has also gradually gained attention. The WaveFake[22] and FakeAVCeleb[28] datasets are explicitly designed for audio deepfake detection. To meet more complex challenges, datasets such as LAV-DF[27] and GOTCHA[29] focus on capturing subtle traces in generating deepfakes, further enhancing the depth and accuracy of detection technology.

New datasets such as WildDeepfake[23], OpenForensics[24], FFIW10K[25], DeePhy[26], DFDM[30], AV-Deepfake1M[31], CIFAKE[32], and DF40[33] have further broadened the depth and breadth of research, each contributing to the solution of challenges and the advancement of technology in its specific way. Overall, the design of these datasets reflects the diverse challenges and application scenarios faced by deepfake technology. They not only promote the development of technology but also provide researchers with a platform to experiment and verify new methods.

As shown in Table 1, many studies use three databases, FF++, Celeb-DF, and DFDC, for Deepfake detection testing. The widespread use of these datasets not only shows their popularity in academia, but also reflects that they provide a standardized and comparable evaluation platform for Deepfake detection algorithms. Although some studies have performed very well on FF++, even achieving near-perfect detection rates (such as 99.99% or even 100%), this high performance indicates that existing detection methods can effectively identify Deepfake content in this dataset.

However, this high performance may also imply that these datasets are relatively outdated and may not fully represent the latest developments in current Deepfake generation technology. For example, with the advancement of generative adversarial network (GAN) technology, newly generated Deepfakes may have fewer recognizable artifacts and higher visual quality, making them more difficult to detect than samples in earlier datasets such as FF++. In addition, these datasets may fail to cover all possible Deepfake application scenarios, such as different cultural backgrounds, different lighting conditions, and complex background noise, which are extremely common in real-world applications.

Therefore, while existing datasets are critical to developing and testing Deepfake detection techniques, the research community should also focus on developing more datasets that include modern Deepfake techniques and more challenging scenarios. This will not only better evaluate the actual effectiveness of existing techniques, but also promote the continued advancement of detection technology to address the evolving Deepfake threat.

4 Literature review

In this section, we will introduce and discuss the relevant works from 2018 to the present in detail, arranged in chronological order. We will show the main progress in technical methods, tool applications, and theoretical frameworks over the years according to the year of publication of the research. In addition, the contribution of each study and its position in the existing technology are analyzed to provide a deep understanding of the evolution of Deepfake detection technology. In this way, we hope to grasp the overall trend of the development of this field.

4.1 2018-2020

2018 As shown in Table 2, several representative methods emerged for the initial deepfake image detection task in 2018. Tariq, Shahroz, et al.[34] proposed a neural network (Ensemble ShallowNet) method based on ensemble learning to detect fake face images generated by GAN and artificially produced at different resolutions. McCloskey, Scott, et al.[35] observed the characteristics of the Deepfake images at the time and used color clues to distinguish between GAN-generated and camera-created images.

Meanwhile, in the field of deepfake video detection, Li, Y.[36] Because the images generated by deepfake technology at the time could only reach a limited resolution, the generated images needed to be distorted to fit the faces in the source video, and the image distortion process would leave recognizable artifacts in the video. By using a convolutional neural network (CNN) to capture these artifacts, Deepfake videos were distinguished from the original videos. Similarly, Li, Yuezun, et al.[37] found that deepfake videos were insufficient in simulating the natural biological phenomenon of blinking, so they developed a method based on blinking detection. Using the local convolutional recurrent network (LCRN) model, the naturalness of blinking movements is analyzed to identify Deepfake videos. The disadvantage of this method is that it can only identify missing blinks. Compared with these two methods, Güera, David, et al.[38] proposed an automatic detection method that combines convolutional neural networks (CNN) and recurrent neural networks (LSTM), utilizing the image processing capabilities of CNN and the time series data processing capabilities of LSTM, significantly improving the recognition rate of Deepfake videos and achieving an accuracy of 97.1%. In Afchar, Darius, et al.[39], it is believed that traditional image forensics techniques are usually not suitable for videos because compression can seriously reduce data quality. They proposed two detection networks, Meso-4 and MesoInception-4, both of which have fewer layers and parameters and focus on using the mesoscopic features of images for analysis. These networks have shown up to 98% and 95% accuracy in detecting Deepfake and Face2Face. In addition, there are detection methods based on photo response non-uniformity (PRNU) analysis[40] and CNN[13] [41] [42].

2019 In 2019, there were some new developments in the field of Deepfake detection research. First, the study by Matern et al.[43] reviewed the facial processing technology at the time and its characteristic artifacts during the processing, such as eyes, teeth, and

Table 1. List of Deepfake Dataset (From 2018 to 2024)

DATASET	YEAR	TYPE	SIZE	WORKS
FaceForensics[12]	2018	Video	1,004 Real videos 2,008 Fake videos	[45][53]
Fake Faces in the Wild(FFW)[13]	2018	Image	53000 Real/Fake images	[13]
DeepfakeTIMIT[14]	2018	Audio-Video	320 Real videos 640 Fake videos	[23][49][68][97]
FaceForensics++ [15]	2019	Video	1000 Real videos 4000 Fake videos	[23][46][47][60][61] [62][64][66][67][68][70] [71][72][73][74][75][76] [77][78][79][81][82][84] [86][87][88][89][90][91] [92][93][94][95][96][97] [100][101][102][105][106] [107][108][109][110][111] [112][113][114][115][116] [117][118][120][122]
DFDC [16]	2019	Video	23,654 Real videos 104,500 Fake videos	[54][61][62][68][69][73] [75][77][85][87][88][89][90] [91][92][94][95][97][102] [103][105][106][107][108][109] [110][111][115][116][117][118] [119][120][122]
Celeb-DF [17]	2020	Video	590 Real videos 5,639 Fake videos	[54][58][61][66][67][68] [70][71][73][74][75][77] [78][82][84][85][86][88] [89][90][91][92][93][94] [95][96][100][102][103][105] [106][107][108][109][110][111] [112][113][114][115][116][117] [118][122]
DeeperForensics [18]	2020	Video	1000 Real/Fake videos	[67][89][91][92][106][107] [109][111][112][113][114]
ForgertNet [19]	2021	Video/Image	99,630 Real videos 121,617 Fake videos	-
DeepFake MNIST+ [20]	2021	Video	10,000 Real/Fake videos	-
KoDF[21]	2021	Audio-Video	62,166 Real videos and 175,776 Fake videos	[99]
WaveFake[22]	2021	Audio-Video		-
WildDeepfake[23]	2021	Video/Image	3805 Real videos 3509 Fake videos	[23][75][82][85][90][91] [95][96][111][117][120]
OpenForensics[24]	2021	Image	45,473 Real image 70,325 Fake image	-
FFIW10K[25]	2021	Video	10,000 Real/Fake videos	[118]
DeePhy[26]	2022	Video	100 Real videos 5,040 Fake videos	-
LAV-DF[27]	2022	Audio-Video	36,431 Real videos 99,873 Fake vides	-
FakeAVCeleb[28]	2022	Audio-Video	500 Real videos 19,500 Fake videos	[97][99]
GOTCHA[29]	2022	Video	56,247 Real/Fake videos	-
DFDM[30]	2022	Video	6,450 Real and Fake videos	-
AV-Deepfake1M[31]	2023	Audio-Video	286,721 Real videos 860,039 Fake videos	-
CIFAKE[32]	2024	Image	60,000 Real/Fake images	-
DF40[33]	2024	Video/Image	1M+ Real/Fake videos	-

facial contours. The study claimed that relatively simple visual artifacts could be very effective in exposing such operations, including Deepfakes and Face2Face. This method is similar to the study by Li, Y.[36], both observing artifacts to identify Deepfakes, but Matern et al.[43] have more human involvement. Yang, Xin, et al. [44] from the same team as Li, Y.[36] proposed a new method using SVM classifiers to identify erroneous artifacts in synthetic face areas, which relatively requires less human involvement.

In applying convolutional neural networks, Nguyen, HH., et al.[45] used capsule networks to detect Deepfakes. This network was used for detection after extracting the latent features of VGG-19. The study also compared the performance of Capsule-Forensics-Noise containing random noise and standard Capsule-Forensics, showing that it is better than the method of Afchar, Darius et al.[39] on the FF[12] database. In October of the same year, an extended article on this method was published [46], and several modifications and two regularizations were introduced to enhance its performance; Sabir, Ekraam, et al.[47] combined the circular convolution model with face alignment technology. Experiments found that the face alignment based on landmarks combined with a bidirectional circular dense network performed best in facial manipulation detection in videos. D. Cozzolino et al. [48] proposed the ForensicTransfer (FT) neural network for image-level Deepfake detection, aiming to solve the problem of CNN's performance degradation in the detection of unknown forgery methods, showing an efficient performance of 95%.

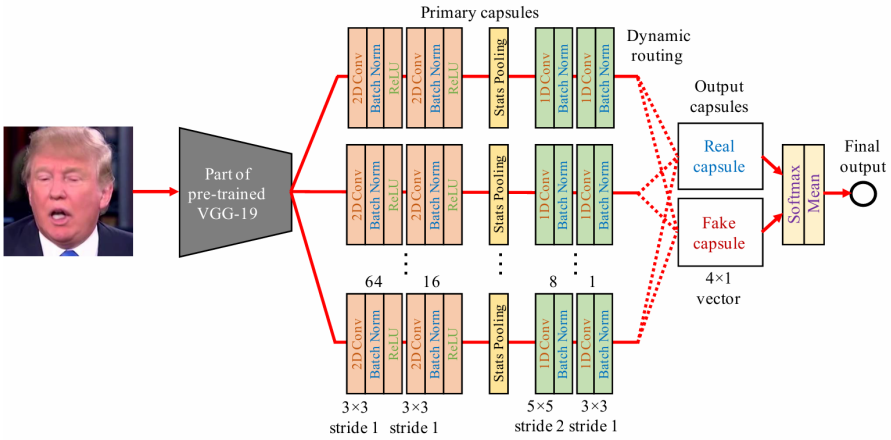

Fig. 5. Proposed method of Nguyen, HH., et al.[46]

In addition to traditional feature descriptors, Akhtar et al.[49] conducted a critical study to evaluate the potential of local feature descriptors in face-swapping detection.

Work	Deepfake	Method
Tariq, Shahroz, et al. [34]	Image	Ensemble ShallowNet
McCloskey, S, et al.[35]	Image	Color clues
Li, Y.[36]	Video	Artifacts;CNN
Li, Yuezun, et al.[37]	Video	Eye blinking
Güera, David, et al. [38]	Video	CNN and LSTM
Afchar, Darius, et al.[39]	Video	Meso-4 and MesoInception-4
Koopman, M, et al [40]	Video	PRNU analysis
Do, Nhu-Tai, et al. [41]	Video	CNN-Based
Badale, Anuj, et al.[42]	Video	CNN-Based

Table 2. Deepfake detection research in 2018

Work	Deepfake	Method
Matern, et al.[43]	Image	Artifacts
Yang, Xin, et al [44]	Image	Artifacts+SVM
Nguyen, HH., et al.[45]	Video	Capsule-Forensics
Nguyen, HH., et al.[46]	Video	Enchanced Capsule-Forensics
Sabir, Ekraam et al.[47]	Video	RNN+Face align
D. Cozzolino et al. [48]	Image	ForensicTransfer
Akhtar, et al.[49]	Video	LBP, FDLBP etc.
Kharbat, et al. [50]	Video	HOG,ORB etc.+ SVM
Dorević, et al.[51]	Video	SIFT
Zhang, et al.[52]	Video	ELA and DL

Table 3. Deepfake detection research in 2019

They tested ten widely used local descriptors, including LBP, FDLBP, QLRBP, BGP, LPQ, BSIF, CENTRIST, PHOG, SIFT, and SURF. The experiment was conducted on the DeepfakeTIMIT[14] database, and the results showed that these local descriptors are very effective in identifying low-quality manipulated faces. Still, their performance decreases when processing high-quality manipulated samples. At the same time, Kharbat et al. [50] also used local descriptors and SVM classifiers for Deepfake video detection. Unlike Matern et al.[43], the Kharbat team chose feature point extraction methods, including HOG, ORB, BRISK, KAZE, SURF, FAST, etc. Their research results show that the SVM method trained with feature detector descriptors can effectively detect fake videos, among which HOG performs best in Deepfake detection; Dorević et al.[51] used SIFT features to analyze Deepfake videos. By matching key points between consecutive frames, this study verified the effectiveness of traditional feature descriptors in distinguishing original videos from forged videos. Zhang et al.[52] proposed a new model based on deep learning and error level analysis (ELA) detection. The ELA method can significantly improve the training efficiency of the CNN model, and the detection accuracy can reach more than 97%. The above techniques are briefly summarized in Table 3.

2020 By 2020, the application of deep learning in deepfake detection will have increased. As shown in Table 4, Kumar et al. [53] proposed a method to detect Deepfake videos by learning regional artifacts. The study used five parallel ResNet-18 models, four of which focused on learning local and regional artifacts, and one model was used to understand the overall effect of facial reproduction. On the uncompressed FaceForensics database, the detection accuracy of this method reached 99.96%. Ranjan et al. [54] analyzed the performance of the transfer learning convolutional neural network framework in the Deepfake detection task. The results showed that transfer learning significantly improved single-domain classification accuracy and generalization ability. Guarnera, Luca, et al. [55] extracted local features specific to the deep learning convolution generation process through the expectation maximization (EM) algorithm to simulate the "fingerprints" left in the image generation process.

In addition, some studies have adopted traditional methods to detect Deepfake videos. For example, Younus et al. [56] used the Haar wavelet transform to analyze the sharpness and type of edge of the facial area in the video to quickly and effectively detect Deepfake videos. Kawa et al. [57] proposed an enhancement method based on [39] image-level detection research and designed a new activation function Pish. Experimental results show that MesoNet equipped with the Pish activation function outperforms the baseline model under resource-constrained conditions. At the same time, De Lima, Oscar, et al. [58] eval-

uated the effects of different action recognition networks in Deepfake detection and found that the R3D network outperformed other models in performance.

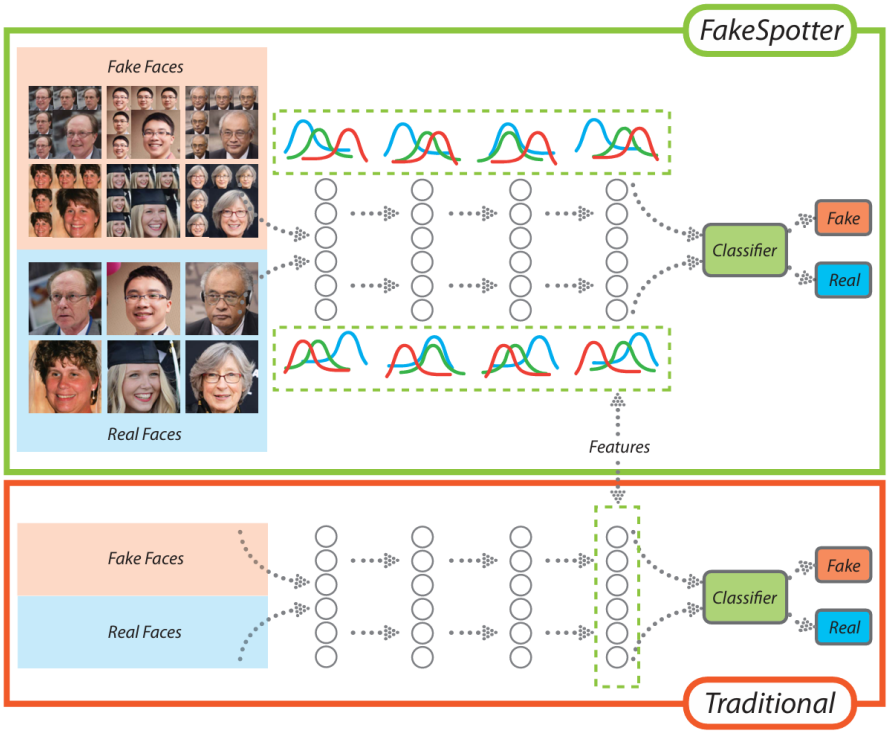

Fig. 6. Proposed method of Wang, R. et al.[59]

Meanwhile, researchers have proposed innovative networks. Wang, R. et al.[59] and Rana et al. [60] proposed FakeSpotter and DeepfakeStack, respectively. FakeSpotter is the first method to detect AI-generated fake faces based on monitoring neuronal behavior, showing strong robustness against common perturbation attacks. DeepfakeStack is an ensemble learning method with a detection accuracy of 99.65% and F1-SCORE and accuracy close to 1. In addition, Du, Mengnan, et al.[63] proposed the Locality-Aware AutoEncoder (LAE) method, while Zi, Bojia, et al.[23] proposed the Attention-Based Deepfake Detection Network (ADDNets). Dong, Xiaoyi, et al.[65] introduced the identity-driven "Outerface" algorithm.

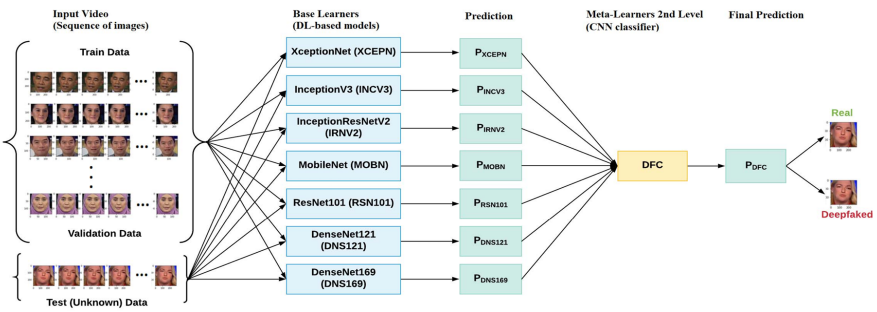

Fig. 7. Proposed method of Rana et al. [60]

Work	Deepfake	Method
Kumar et al.[53]	Video	Five parallel ResNet-18
Ranjan et al.[54]	Video	Transfer learning CNN
Luca et al. [55]	Video	EM +DL
Younus et al.[56]	Video	Haar wavelet transform
Kawa et al. [57]	Image	MesoNet with Pish
Lima, Oscar et al. [58]	Video	Action recognition networks
Wang, R. et al.[59]	Video	FakeSpotter
Rana et al. [60]	Video	DeepfakeStack
Nirkin, Y., et al. [61]	Video	XceptionNet
Chintha, Akash, et al.[62]	Video	modified XceptionNet
Du, Mengnan, et al. [63]	Video	LAE
Zi, Bojia, et al.[23]	Video	ADDNets
Pan, Deng, et al.[64]	Video	Xception and MobileNet
Dong, Xiaoyi, et al. [65]	Video	Outerface
Xie, Daniel, et al.[66]	Video	AlexNet

Table 4. Deepfake detection research in 2020

Regarding detection technology using XceptionNet, Nirkin, Y., et al. [61] significantly improved the performance of traditional classifiers by training two different signals based on the Xception network to detect the difference between faces and backgrounds. Chintha, Akash, et al.[62] modified the XceptionNet architecture and introduced edge and optical flow maps to help isolate Deepfakes at the instance and video levels, significantly improving detection accuracy. Pan, Deng, et al.[64] compared the performance of Xception and MobileNet in Deepfake detection, and the results showed that Xception performed better. Xie, Daniel, et al.[66] used an improved light-weight version of AlexNet to identify real and fake videos and performed exceptionally well in the Celeb-DF database, with an accuracy of 98.85%.

The trend of deepfake detection technology development from 2018 to 2020 shows that with the advancement of deepfake generation technology, the artifact problem has received widespread attention in early research. However, such research has gradually decreased over time. This reflects that Deepfake technology is constantly evolving to cope with increasingly sophisticated detection technology. In 2019, we observed a unique phenomenon in which some researchers began to try to apply traditional technologies in image forgery detection to deepfake detection. The practice has proved that these conventional methods are, in fact, feasible. Still, with the development of deep learning technology, the technique of active feature extraction has gradually become less efficient than automatic feature extraction. The technological progress in 2020 has verified the advantages of deep learning in improving detection efficiency and accuracy.

4.2 2021-2023

2021 In 2021, innovations in artifact detection, traditional image processing techniques, application of Vision Transformer, unsupervised learning methods, and methods based on spatial and temporal features. As shown in Table 5.

Trinh, Loc, et al.[67] proposed a Dynamic Prototype Network (DPNet) that uses dynamic representations (i.e., prototypes) to explain deepfake temporal artifacts. It consists of the feature encoder, the prototype layer, the fully connected layer, and the temporal logic verifier to improve the interpretability and credibility of the model. Based on traditional image processing techniques, Xu, Bozhi, et al.[68] combined traditional image processing techniques such as gradient, standard deviation, gray-level co-occurrence matrix, and wavelet transform and detected Deepfake videos through SVM.

Since the advent of ViT, some researchers have tried to apply it to deepfake detection. Wodajo et al. [69] proposed a convolutional visual transformer combining the advantages of CNN and ViT. In this model, CNN is first used as a front-end to extract learnable features, and then these features are input into ViT, which uses its attention mechanism to analyze these features further and classify them. This shows that combining CNN with ViT is an effective strategy for capturing the nuances in Deepfake-generated content. Kaddar, Bachir, et al.[79] also proposed the HCiT method, which combines CNN and ViT. CNN is used to efficiently extract features from video frames, while ViT uses its self-attention mechanism to process these features and perform deep classification. This combination takes advantage of the powerful ability of CNN in feature extraction and the ability of ViT to focus on essential parts in processing sequence data, thereby providing higher accuracy and better result interpretation in detecting Deepfake videos.

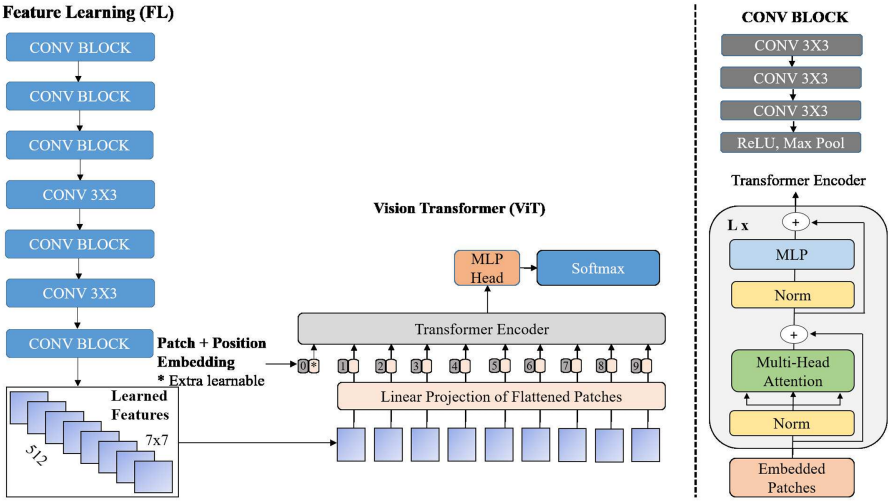

Fig. 8. Proposed method of Wodajo et al.[69]

Fung, Sheldon, et al.[70] designed a novel deepfake detection method based on unsupervised methods. First, two different transformed versions of the image are generated and input into two consecutive sub-networks, namely the encoder and the projection head. Unsupervised training is achieved by maximizing the correspondence of the projection head output.

Based on the space-based methods, Chen, Hong-Shuo, et al.[71] use the principle of successive subspace learning (SSL) to extract features from various parts of face images automatically. Ismail, Aya, et al.[74] proposed a new model, InceptionResNetV2-XGBoost, to learn spatial information and then detect the authenticity of the video. Gu, Zhihao, et al.[75] proposed the spatiotemporal inconsistency learning (STIL) process and instantiated it as a new STIL block, which consists of a spatial inconsistency module (SIM), a temporal inconsistency module (TIM), and an information supplementation module (ISM). A new temporal modeling paradigm is proposed in TIM by exploiting the temporal differences of adjacent frames and horizontal and vertical directions.

There are other detection methods. Kim, Minha, et al.[72] adopt the representation learning (ReL) and knowledge distillation (KD) paradigms to introduce the feature representation transfer adaptation learning (FReTAL) method based on transfer learning. Zhao, Hanqing, et al.[73] formulate deepfake detection as a fine-grained classification problem

Work	Deepfake	Method
Trinh, Loc, et al.[67]	Video	DPNet
Xu, Bozhi, et al.[68]	Video	Traditional method+ SVM
Wodajo et al. [69]	Video	CNN+ViT
Fung, Sheldon, et al.[70]	Video	Unsupervised method
Chen, Hong-Shuo, et al.[71]	Video	SSL
Kim, Minha, et al.[72]	Video	FReTAL
Zhao, Hanqing, et al.[73]	Video	Multi-attention method
Ismail, Aya, et al.[74]	Video	InceptionResNetV2-XGBoost
Gu, Zhihao, et al.[75]	Video	STIL
Zhao, Tianchen, et al.[76]	Video	PCL
Das, Sowmen, et al.[77]	Video	Face-Cutout
Zhao, Lei, et al.[78]	Video	MFF-Net
Bachir, et al.[79]	Video	HCiT

Table 5. Deepfake detection research in 2021

and propose a new multi-attention deepfake detection network. Zhao, Tianchen, et al.[76] introduce a new representation learning method called pairwise self-consistent learning (PCL) to train ConvNets to extract these source features and detect deepfake images. Das, Sowmen, et al.[77] propose a simple data augmentation method called Face-Cutout. This method uses facial landmark information to cut out image regions dynamically. Zhao, Lei, et al.[78] propose a deepfake detection network that fuses RGB features and texture information extracted by neural networks and signal processing methods, namely MFF-Net.

2022 As shown in Table 6, Jeong, Yonghyun et al.[80] proposed a Deepfake image detection method called Bilateral High-Pass Filter (BiHPF). This method uses two high-pass filters (HPF) to amplify the frequency-level artifacts that are commonly found in images synthesized by generative models. Among them, the frequency-level HPF is used to enhance the artifact amplitude in the high-frequency component, while the pixel-level HPF is used to emphasize the changes in background pixels in the pixel domain. This method mainly relies on Fourier transform and its inverse transform. Based on frequency artifacts, Jeong, Yonghyun et al.[81] further proposed a strategy of using GAN to treat GAN in a subsequent paper[81], by generating frequency-level perturbation maps (Frequency-Level Perturbation Maps), making the generated images indistinguishable from real images in the frequency domain, and integrating this information into detector training to improve the sensitivity of detection. Chen, Liang et al.[83] also adopted a similar strategy of using GAN to treat GAN, using synthesizers and adversarial training frameworks to dynamically generate forgeries. By training to recognize generated fakes, the network can learn more powerful feature representations and produce a more general deepfake detector.

The application of ViT continues to make progress in the field of Deepfake detection. Khormali et al[82] is different from ([69][79]) who also use ViT. This method does not rely on CNN and proposes a transformer model developed mainly for deepfake detection tasks. This method has achieved excellent performance of 99.41%, 99.31% and 81.35% in FaceForensics++, Celeb-DF (V2), and WildDeepfake. Wang, Junke, et al.[84] proposed the Multi-modal Multi-scale TRansformer (M2TR) method, which uses a multi-scale transformer based on Transformer to detect local inconsistencies at different scales and uses frequency features to improve robustness. Wang et al.[86] proposed a method called LiSiam, in which the feature extractor based on the twin network takes the original image and the corresponding quality-degraded image as paired inputs and outputs two segmentation maps. A local invariance loss is further proposed to impose local consistency

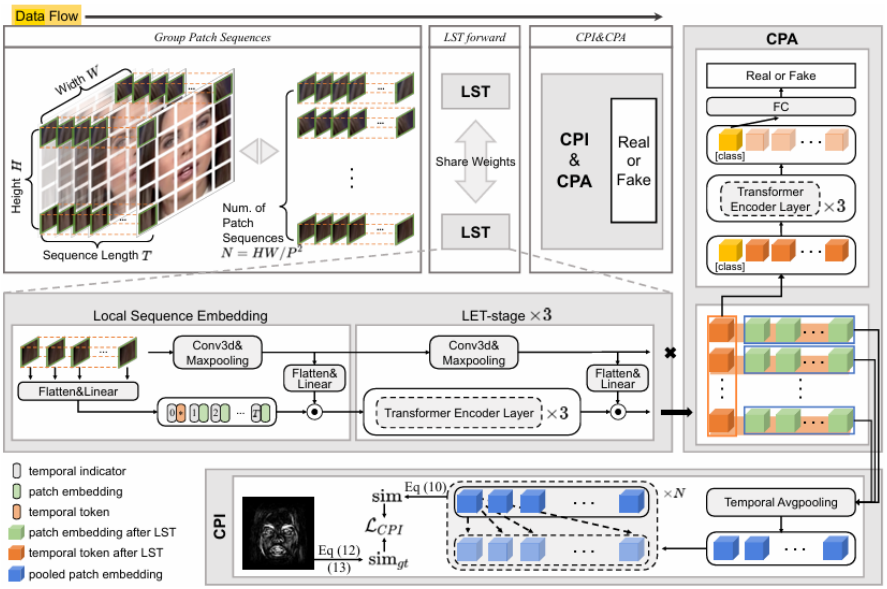

Fig. 9. Proposed method of Guan, Jiazhi, et al.[88]

between the two segmentation maps. A Mask-guided Transformer is designed to capture the co-occurrence between the forged area and its surroundings. A multi-task learning strategy is used to obtain robust and discriminative feature representations, and multiple objective functions are jointly optimized in an end-to-end manner. In other related studies, Khan et al. [87] used two CNN networks, XceptionNet and EfficientNet-B4, as feature extractors, and then input the extracted features into the Transformer for training. For video-level detection, Guan, Jiazhi, et al.[88] proposed a Transformer-based Local & Temporal-aware Transformer-based Deepfake Detection (LTTD) framework, which adopts a local-to-global learning protocol, with special emphasis on the valuable temporal information in local sequences. The common point of these methods is to use the powerful features of Transformer to improve the accuracy and interpretability of Deepfake detection.

Other approaches should not be ignored. Hu, Juan, et al.[85] proposed a frame-based inference detection framework (FInfer) designed for detecting deepfake videos with high visual quality. Meanwhile, Kingra et al.[89] developed LBPNet, a network that exploits texture inconsistencies to distinguish deepfake faces from real faces. The network uses a convolutional neural network (CNN)-based model that focuses on analyzing the local binary patterns (LBP) of deepfakes and original faces. This approach emphasizes the key role of texture features in identifying and distinguishing deepfake content, demonstrating the effective combination of traditional image processing techniques and modern machine learning methods.

2023 By 2023, Deepfake detection research has reached a new high point. Table 7 summarizes the progress of the relevant research. Ke et al.[90] proposed a detection method for degraded deepfake videos called DF-UDetector. This method improves detection efficiency by modeling degraded images and converting extracted features into high-quality features. Yu, Yang, et al.[91] developed a new enhanced multi-scale spatiotemporal inconsistency amplifier (AMSIM), which contains a global inconsistency view (GIV) and a more detailed multi-time scale local inconsistency view (MLIV). Zhao, Cairong, et al.[92] pro-

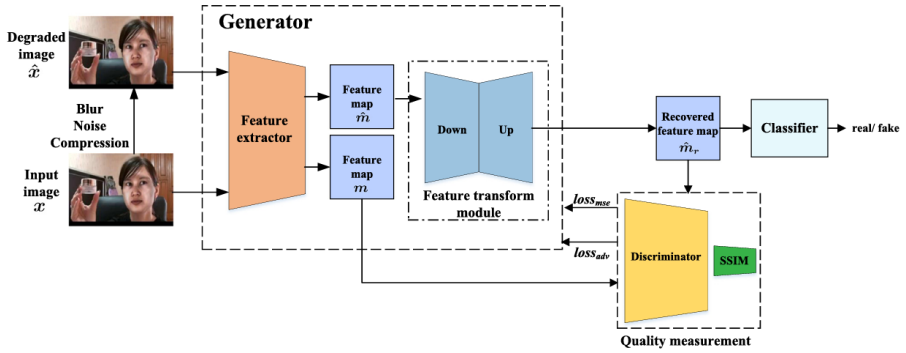

Fig. 10. Proposed method of Ke et al.[90]

posed an Interpretable Spatiotemporal Video Transformer (ISTVT), which consists of a novel decomposed spatiotemporal self-attention and self-reduction mechanism to capture spatial artifacts and temporal inconsistencies for robust deepfake detection. BR, Shobha Rani, et al.[93] adopted a different technology fusion approach, combining ResNet50 and long short-term memory network (LSTM) to form a hybrid architecture. Li, Xin, et al.[94] proposed an artifact disentanglement adversarial learning (ADAL) method to address the problem that traditional supervised binary classification methods often extract a large amount of information irrelevant to artifacts, which leads to performance degradation in deepfake detection.

Applications based on ViT continue to develop steadily. The model proposed by Lin, Hao, et al.[95] particularly emphasizes the combination of multi-scale convolution and visual transformers, using dilated convolution and depth-separable convolution to capture more facial details and signs of tampering at different scales. Unlike traditional classification methods, this model uses visual transformers to learn further and classify global information of facial features. At the same time, Heo et al.[103] developed an efficient visual transformation model specifically for DeepFake detection, which can extract local and global features simultaneously. Unlike the typical method of combining CNN and ViT, this new method combines vector-concatenated CNN features and block-based localization to point out the forged area. This method also introduces the concept of distilled labeling, which improves the performance and generalization ability of the model by optimizing logit in sigmoid function training using binary cross entropy. In response to the challenge of diffusion model generating clearer and more detailed images, Aghasanli et al.[104] proposed a detection scheme that combines a fine-tuned visual transformer (ViT) and a classic classifier such as a support vector machine (SVM). This method demonstrates its explana-

Work	Deepfake	Method
Jeong, Yonghyun, et al.[80]	Image	BiHPF
Jeong, Yonghyun, et al.[81]	Image	Frepgan
Khormali et al.[82]	Image	DFDT ViT-based
Chen, Liang, et al.[83]	Image	GAN TO GAN
Wang, Junke, et al.[84]	Image	M2TR ViT-based
Hu, Juan, et al.[85]	Video	Finfer algorithm
Wang et al.[86]	Image	LiSiam
Khan et al. [87]	Image	Hybrid transformer
Guan, Jiazhi, et al[88]	Video	LTTD transformer-based
Kingra et al.[89]	Image	LBPNet

Table 6. Deepfake detection research in 2022

Work	Deepfake	Method
Ke et al.[90]	Image	DF-UDetector
Yu, Yang, et al.[91]	Video	AMSIM
Zhao, Cairong, et al.[92]	Video	ISTVT
BR, Shobha Rani, et al.[93]	Video	Resnet50 + LSTM
Li, Xin, et al.[94]	Image	ADAL
Lin, Hao, et al. [95]	Image	ViT-based
Wu, Jianghao, et al.[96]	Video	Two-stream network
Salvi, Davide, et al.[97]	Audio-Video	Time-aware neural networks
Wang, Tianyi, et al.[98]	Video	Transformer-based
Feng et al[99]	Audio-Video	Anomaly detection
Tan, Lingfeng, et al.[100]	Video	FADE
Hou, Yang, et al.[101]	Image	StatAttack
Liang, Yufei, et al.[102]	Image	Two-stream network
Heo et al.[103]	Video	ViT-based
Aghasanli et al [104]	Image	ViT-based
Guo, Zhiqing, et al.[105]	Image	SFIConv
Shuai, Chao, et al.[106]	Image	Two-stream network

Table 7. Deepfake detection research in 2023

tory power by analyzing the support vector of SVM, proving the possibility of explaining DeepFake detection through prototypes. In addition, Wang, Tianyi, et al.[98] proposed a deep convolutional Transformer method to solve the problem that local features alone are not enough to provide sufficient information for effective Deepfake detection. This method combines local and global decisive image features, enriches the expression of features, and improves the efficiency and accuracy of the model by applying convolution pooling and re-attention techniques.

In Audio-Video Deepfake detection, Salvi, Davide, et al.[97] extract time-varying audiovisual features from the input video and analyze them using a time-aware neural network. The video and audio modalities exploit inconsistencies between and within them, improving final detection performance. Feng et al[99], believe that there are often subtle inconsistencies between the visual and audio signals of processed videos. A video forensics method based on anomaly detection is proposed that can identify these inconsistencies and can be trained using only real unlabeled data.

Some research focuses on developing a TWO-STREAM framework to enhance Deepfake detection. Liang, Yufei, et al.[102] proposed a method that combines conventional spatial and frequency streams, especially for low-quality images, because artifacts in the frequency domain are often more obvious in these images. Wu, Jianghao, et al.[96] also adopted the TWO-STREAM framework, relying on discrete cosine transform (DDCT) to enhance the framework's frequency analysis capability to capture better the frequency features introduced by Deepfake technology. Shuai, Chao, et al.[106] used SRM (robust image model) filters as part of the two-stream framework to further improve the model's sensitivity to subtle texture changes in Deepfake videos.

In addition, Tan, Lingfeng, et al.[100] transformed the deep fake video detection problem into a graph classification task and proposed a new paradigm for deep fake video detection called facial action dependency estimation (FADE). On the other hand, Hou, Yang, et al.[101] proposed a statistical consistency attack (StatAttack) for DeepFake detection. At the same time, Guo, Zhiqing, et al.[105] proposed a technique called space-frequency interaction convolution (SFIConv), specifically designed to simulate and detect manipulation clues in Deepfake videos effectively.

From 2020 to 2023, we can observe that in the field of Deepfake detection, the research on traditional methods is gradually decreasing, but it has not faded out of our field of

vision, while the methods based on deep learning and machine learning are increasing. In addition, it is worth noting that the application of some key technologies has also increased significantly. For example, Xception has been a common technology for deep fake detection since its launch in 2016, while the Transformer introduced in 2017 and its derivative technology Vision Transformer in 2020, these advanced technologies have been widely used in the field of Deepfake detection. These developments not only mark the speed of technological progress, but also reflect the unremitting efforts of the research community in finding more effective solutions to the increasingly complex Deepfake problem.

4.3 For Now

By 2024, deepfake detection technology has developed rapidly, forming a variety of research directions and methods. As shown in Table 8. Zou, Mian, et al. [107] A semantic-oriented deepfake detection method, but due to its training of semantic-oriented DeepFake detectors requires a large amount of manual annotation to specify the degree of manipulation parameters and semantic label hierarchy Structure, to address this challenge, in August of the same year, Zou et al. further proposed an improved method [109], which exploits the relationship between face semantics through joint embedding. With ViT as the backbone, the joint embedding of face images and their corresponding labels is used for prediction, and a two-layer optimization strategy is used to dynamically balance the fidelity weights of various tasks, making the training process fully automated.

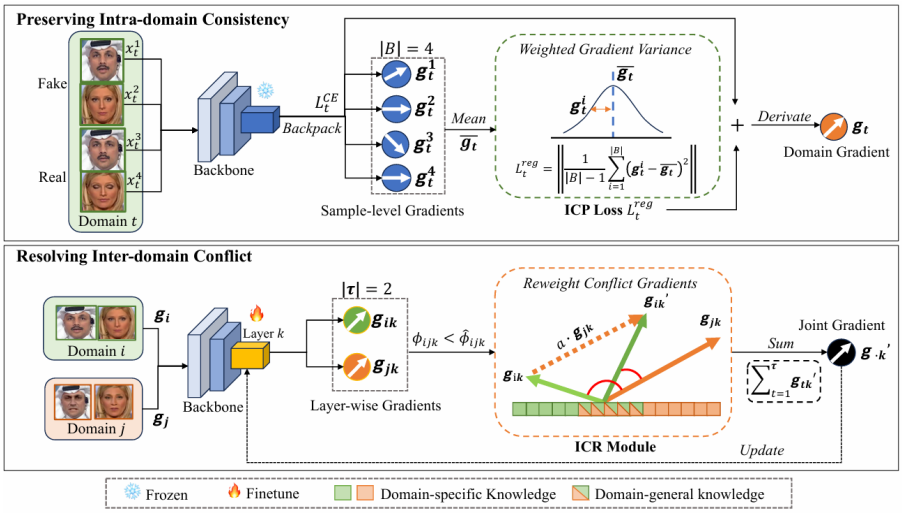

Fig. 11. Proposed method of hen, Jin, et al.[108]

She, Huimin, et al. [110] proposed a dual-branch network to extract node features from RGB images and their color difference images (CDI) through a Transformer-based trainable node encoder module (TNEM). The two features are linked and input into the graph classifier and node classifier for forgery detection and forgery localization respectively. At the same time, Liu, Baoping, et al.[112] introduced a method called Motion-enhanced Spatiotemporal Transformer (MeST-Former), which specifically introduced the identity decoupled attention (IDC-Att) module to separate components related to and unrelated to personal identity. By focusing on components unrelated to identity, this method constructs more generalized spatiotemporal features, enabling the model to effectively adapt

to unknown identities and improving the wide applicability of the model in practical applications. Zakkam, John, et al.[122] developed a Deepfakes detection framework called CoDeiT, which combines the hierarchical attention mechanism and contrastive learning in the HiLo Transformer architecture. By using HiLo Attention technology, CoDeiT is able to distinguish and process high-frequency (Hi-Fi) and low-frequency (Lo-Fi) information, thereby improving computational efficiency and detection accuracy.

In the application of supervised and unsupervised learning methods, Zhang, Rui, et al.[111] innovatively proposed the Self-supervised Face Geometry Information Analysis Network (SF-GAN), which uses graph convolutional networks (GCN) to establish explicit and implicit geometric relationships to exploit facial geometry. By analyzing the geometric relationship between facial landmark maps and information region maps, effective abnormal regions can be identified, thereby minimizing uncertainty. At the same time, Zheng, JunShuai, et al.[116] developed a method that combines unsupervised and supervised contrastive learning frameworks. This is the first attempt to apply unsupervised contrastive learning and supervised contrastive learning to deep fake detection at the same time. Experiments show that this method can improve the model's generalization ability.

Wang, Fei, et al.[113] proposed a novel two-stream framework approach, which is different from the traditional framework in which the binary and multi-classification models work independently. In their model, the binary and multi-classification models work together to enhance the accuracy and efficiency of the model through an innovative fusion and freezing mechanism. On the other hand, Zhang, Dengyong, et al.[117] introduced a two-stream framework called Double-Frequency Transformer Module (DFTM). This framework relies on SRM convolution, which is different from the technology relied on by other two-stream frameworks such as [96], [102] and [106].

In addition, various innovative methods have been proposed. Chen, Jin, et al.[108] proposed applying conflict resolution (ConfR) to minimize conflicts and learn features that generalize across forgeries. Fahad, Muhammad, et al. [114] used a deep learning-based enhanced Resnet-18 and convolutional neural network (CNN) multi-layer max pooling to classify processed videos. Zhang, Kuiyuan, et al. [115] proposed using a well-trained teacher model to train their extended model, and then transferred the extended model to the target domain. At the same time, they proposed a frequency extraction module to extract frequency features as a supplement to spatial features, and introduced spatial frequency contrast loss to enhance feature learning capabilities. Lu, Lin, et al. [118] proposed a method called Deep Forgery Detection by Separable Self-Consistency Learning (SSCLDFD). Lin, Yuzhen, et al. [120] proposed a method called Curricular Dynamic Forgery Augmentation (CDFA). Alazwari, Sana, et al. [121] proposes an Artificial Rabbits Optimization with Transfer Learning Deepfake Detection for Biometric Applications (AROTL-DFDBA);

Finally, Ain, Qurat Ul, et al. [119] studied the vulnerability of deepfake detectors to adversarial black-box attacks from a penetration testing perspective, revealing security holes in existing deepfake detection techniques. They proposed a facial mole-aware black-box adversarial attack against deepfake detectors, where the attacker has limited knowledge of the detector's architecture and settings. The attack showed that subtle perturbations that are visually natural on the face can severely interfere with and degrade the detector's accuracy by up to 40.3%, with a maximum success rate of 48.7%. This provides a clear research direction for future optimization and enhancement of deepfake detectors.

Work	Deepfake	Method
Zou, Mian, et al. [107]	Image	ViT-based
Chen, Jin, et al.[108]	Image	ConfR
Zou, Mian, et al.[109]	Image	ViT-based
She, Huimin, et al.[110]	Video	Transformer-based
Zhang, Rui, et al.[111]	Image	SF-GAN
Liu, Baoping, et al.[112]	Video	Transformer-based
Wang, Fei, et al.[113]	Image	Two-stream based
Fahad, Muhammad, et al.[114]	Video	Resnet-18+ CNN
Zhang, Kuiyuan, et al.[115]	Video	Transfer learning
Zheng, JunShuai, et al.[116]	Video	Contrastive learning
Zhang, Dengyong, et al.[117]	Video	Two-steam based
Lu, Lin, et al.[118]	Video	SSCLDFD
Ain, Qurat Ul, et al.[119]	Video	Attack
Lin, Yuzhen, et al.[120]	Video	CDFA
Alazwari, Sana, et al.[121]	Image	AROTL-DFDBA
Zakkam, John, et al.[122]	Video	Transformer-based

Table 8. Deepfake detection research in 2024

5 Legal aspects

In addition to exploring the technical methods to deal with the challenge of Deepfake, this work also discusses the countermeasures at the legal level. As of 2024, many countries have formulated specific legal provisions for Deepfake.

The United States is the first country to legislate on Deepfake, and has successively introduced a number of bills, including the Malicious Deep Fake Prohibition Act of 2018, the Deepfakes Accountability Act, the Damon Paul Nelson and Matthew Young Pollard Intelligence Authorization Act for Fiscal Year 2020, and the Deepfakes Report Act of 2019, etc., aiming to regulate the use scenarios of Deepfake and promote the research and commercialization of counterfeit detection technology.

The EU officially implemented the General Data Protection Regulation (GDPR) in May 2018, which aims to protect personal data, including data such as citizen images that may be used to create deep fake content. It is applicable to personal privacy leaks that may be caused by face-changing software products released by social media platforms and software companies.

In May 2019, the Singapore Parliament passed the Protection from Online Falsehoods and Manipulation Act, which gives the government the power to require individuals or online platforms to correct or remove false content that may have a negative impact on the public interest. The bill applies to false audio and video produced using deep fake technology.

The "Deep Synthesis Management Regulations" and "Interim Measures" promulgated by China clearly stipulate that when providing and using generative artificial intelligence services, it is not allowed to infringe on the portrait rights, reputation rights, honor rights, privacy rights and personal information rights of others.

In addition, Germany, the United Kingdom, the United Arab Emirates, South Korea, Russia, Vietnam and many other countries have also issued corresponding laws and regulations to deal with the challenges of Deepfake technology.

From a legal perspective, the regulation of Deepfake is not unified globally, and the legal systems of many countries have not yet clarified the use boundaries and legal responsibilities of such technology. Some countries have not yet formulated a response policy. The production and dissemination of Deepfake content requires international cooperation

to formulate a cross-border, multi-field legal framework to regulate it. However, from an ethical perspective, individuals and institutions that develop and use Deepfake technology should consciously abide by ethical standards, avoid improper use of technology, and protect the rights of affected individuals and the public interest.

6 Q & A

In this section, we will set up a series of Q & A sessions to help readers understand deeper about Deepfake detection technology, Deepfake-related databases, and our views on the field. We hope that these questions and answers can provide a comprehensive perspective, allowing readers to have a clearer understanding of Deepfake detection methods and technological evolution and explore the potential impact challenges of these technologies in actual applications.

Q1: What is the main goal of deepfake detection?

The main goal of deepfake detection is to identify and confirm whether the content in videos, images, and audio is manipulated or forged by artificial intelligence technology. As mentioned in Chapter 2, the leading forgery technologies involved include content generated using generative adversarial networks (GANs) and other deep learning methods. Effective deepfake detection can help prevent the spread of misleading information, protect personal privacy, prevent fraud and forgery, and thus maintain the authenticity and credibility of digital media content.

Q2: What are the traditional methods mentioned in the Literature review section? How do they compare to methods such as deep learning?

The Traditional method mentioned in this work mainly refers to the Deepfake detection method based on traditional image processing technology, such as the technology mentioned in the literature[49][50][51][52][56][68] (except SVM), which usually focuses on analyzing the statistical characteristics or frequency components of the image without involving complex learning algorithms.

Compared with methods such as deep learning, traditional methods are usually simple to calculate, less dependent on computer resources, and easy to understand and implement. In this respect, they are lightweight. However, they may not be as flexible and powerful as deep learning-based methods dealing with complex changes and highly realistic Deepfake content. Deep learning methods, especially models based on convolutional neural networks (CNNs) and visual transformers (ViTs), can learn and extract more advanced feature representations, enabling them to detect and combat carefully crafted Deepfake content more effectively. These models can identify subtle patterns and differences by training large amounts of data, providing higher accuracy and robustness.

Q3: What is the difference between Deepfake image and Deepfake video detection?

Deepfake image and Deepfake video detection target different forms of media, each with unique challenges.

Image detection mainly focuses on a single static image. For example, in the Deepfake image detection method shown in Fig.10 and Fig.11, it can be seen that the input of its model is a single picture. The detection work analyzes anomalies in these static images, such as unnatural correction of facial features, mismatch between background and foreground, etc.

However, in video detection, video sequences are processed. Deepfake video detection methods such as those shown in Fig.6, Fig.7, and Fig.9 (also Fig.8 and Fig.5) focus on extracting information from video clips or continuous frames. The challenges include

identifying inconsistencies and abnormal dynamic features between continuous frames in the video, etc.

Q4: Why are FF++, DFDC and Celeb-DF so widely used?

These three databases are widely used mainly because they provide diverse and large-scale forged data samples, which are helpful for training and evaluating the performance of Deepfake detection algorithms. FF++ (FaceForensics++), DFDC (DeepFake Detection Challenge), and Celeb-DF all contain many videos and images. These data have been processed with special technologies, covering different forgery scenarios and technologies, providing rich resources for research.

Although these databases are widely recognized and used in academia and industry, we also encourage the creation of more datasets that reflect real-world application scenarios. This can further stimulate the development of Deepfake detection technology, improve detector accuracy and generalization ability, and ensure their effectiveness and reliability in the real world.

Q5: What is an artifact?

Artifacts in the field of Deepfake detection refer to unnatural features in images due to operations such as editing, compression, and generation. Early Deepfake technology often leaves some easily recognizable traces during the synthesis process. These traces appear as artifacts such as unnatural correction of facial features, blurred edges, inconsistent textures, or color distortion. Some studies, such as [36][43][44], rely on this feature as an important clue to identify deepfakes. In addition, there is also [35] that relies on color clue. However, with the advancement of technology, artifacts have gradually faded.

Q6: What does the evolution of technology mean?

We simply divide deepfake detection into 2018-2020, 2021-2023, and 2024. From the time the concept of Deepfake was first proposed in 2017 to the time it began to attract widespread attention in 2018, the evolution of technology reflected the gradual maturity from early simple applications to later complex technologies.

Between 2018 and 2020, Deepfake technology was mainly identified through visible visual clues. At the time, research focused on verifying the effectiveness of these identifiable clues and whether traditional methods were applicable to Deepfake detection. Much of the research at this stage is based on hypothesis verification and traditional image processing techniques.

As we move into 2021-2023, the advancement of generation technology has significantly improved the quality of Deepfake content produced. It has become increasingly difficult to distinguish authenticity with the naked eye. This has driven the development of detection technology, especially the widespread application of deep learning methods, while traditional methods have gradually declined due to their limitations.

As technology continues to advance and diversify in 2024 and beyond, more efficient and sophisticated methods will emerge in Deepfake detection. These methods include Transformer-based models and detection systems that combine multimodal information, which all mark the evolution of technology. The new stage not only improves detection accuracy but also broadens the scope of the application.

At the end of Section 4.2, we also mentioned the development of Deepfake detection technology, from traditional methods to deep learning methods, from Xception to ViT.

Q7: This work mentions cutting-edge technologies such as Vision Transformers and multimodal detection frameworks. What are the nuances of these technologies, and how do their performance compare?

First, Vision Transformers (ViT) is a deep learning model mainly used for image recognition tasks. It processes information in images by utilizing the self-attention mechanism.

In Deepfake detection, ViT identifies artificial synthesis traces by analyzing the global features of the image. Related works include[69][79][82][84][95][103][104], etc. These methods demonstrate ViT's powerful ability in capturing image details.

Multimodal detection frameworks improve the accuracy and robustness of detection by combining different types of data, such as video and audio. For example,[97] and[99] show how to capture inconsistencies by analyzing visual and audio features in videos, thereby more effectively identifying Deepfake content. This approach is particularly suitable for dealing with forged content in complex scenes because it can verify the authenticity of information from multiple dimensions.

Two-stream frameworks, such as [96][102][106], usually process two different types of image information, such as spatial stream and frequency stream. This framework can analyze the spatial and frequency features of the image separately and then fuse the information of the two to improve the ability to identify subtle forgery traces. This approach can provide a more comprehensive analysis, especially when dealing with high-quality Deepfake generated content.

In general, ViT performs well when processing single-image data, while multimodal and two-stream frameworks are more effective when dealing with complex tasks that require the integration of multiple sources of information. The choice of which technology depends on the requirements of the specific task and the type of data available. In the field of Deepfake detection, combining the advantages of these technologies can design more powerful and flexible detection systems.

Q8: What is robustness, and what steps have researchers taken to improve this property?

When discussing robustness, we must first clarify its definition: the ability of a system to maintain stable performance and functionality despite changes in its internal structure or external environment. For Deepfake detection systems, robustness means that the system can maintain normal operation and good performance even in the face of adverse conditions such as interference, noise, and failures.

To improve robustness, researchers have used a variety of data augmentation techniques, such as resizing, random rotation, horizontal flips, random resized cropping, color jittering, JPEG compression, etc. These methods help the model to be more robust when dealing with input data of different variations, as shown in the literature [108], [110], [114], [115], [117], [119], [122] etc.

In addition, there are robustness enhancement methods for specific situations. For example, [112] uses "aggregate bounding boxes" to crop images and obtain a series of face images with static backgrounds, thereby improving the robustness of the model under background changes. [97] uses multimodal data and independent modality datasets for training and uses multimodal fusion technology during testing to verify the effectiveness of multimodal methods and enhance the adaptability and accuracy of the model. These measures aim to make the detection system more reliable and effective in practical applications.

Q9: What are the challenges of deepfake detection?

This is a commonplace problem. We are still faced with problems such as model generalization ability, computer resource requirements, adversarial attacks, insufficient training data, and highly realistic forged content. Just as no one is perfect, the same is true for models. There is no perfect model. Solving current problems may not necessarily solve future problems. We can only keep seeking, exploring, and pursuing.

7 CONCLUSION

In the past few years, the rapid development of Deepfake technology has posed a serious challenge to the authenticity of information in society. In order to deal with this problem, many studies related to deepfake detection have been proposed one after another. This paper reviews some Deepfake detection methods from 2018 to 2024, and systematically discusses and summarizes them.

Through the comparison and analysis of the development of each year, we observe that Deepfake detection technology has gradually shifted from the initial simple artifact evidence and traditional image processing technology to the application of deep learning models, including but not limited to CNN, Xception, Transformer, Vision Transformer. These methods have shown higher efficiency and accuracy in processing highly realistic Deepfake content. In addition, with the advancement of technology, new challenges continue to emerge, such as the generalization ability of the model, the defense against adversarial attacks, robustness, and the computational resource requirements for processing larger and larger data sets.

In the future, research on deepfake detection will need to consider how to reduce the consumption of computer resources and improve the real-time performance of the algorithm while maintaining high accuracy to meet the needs of daily life. At the same time, we encourage the emergence of more Deepfake datasets to make the detector more powerful and improve the credibility of images and videos.

At the same time, with the rapid advancement of technology, countries need to continuously improve their legal norms on Deepfake. The cost of generating fake videos, pictures or audio is becoming lower and lower, and the use of Deepfake technology to spread false information should also be strictly restricted by law. This is not only related to the protection of personal privacy and rights, but also to the authenticity of public information and the maintenance of social stability. Therefore, it is crucial to build a comprehensive and up-to-date legal framework to effectively respond to this challenge and ensure that social order is not affected by the abuse of such technology.

References

1. McCulloch, W.S. and Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5, pp.115-133.
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. Advances in neural information processing systems, 27.
3. Hauser, Andrea, and Marc Ruef. "Deepfake-An Introduction." skip Labs (2018).
4. Mirza, Mehdi. "Conditional generative adversarial nets." arXiv preprint arXiv:1411.1784 (2014).
5. Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." Proceedings of the IEEE international conference on computer vision. 2017.
6. Karras, Tero. "Progressive Growing of GANs for Improved Quality, Stability, and Variation." arXiv preprint arXiv:1710.10196 (2017).
7. Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
8. Brock, Andrew. "Large Scale GAN Training for High Fidelity Natural Image Synthesis." arXiv preprint arXiv:1809.11096 (2018).
9. Park, Taesung, et al. "Semantic image synthesis with spatially-adaptive normalization." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
10. Kingma, D.P., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.
11. Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in neural information processing systems 33 (2020): 6840-6851.
12. Rössler, Andreas, et al. "Faceforensics: A large-scale video dataset for forgery detection in human faces." arXiv preprint arXiv:1803.09179 (2018).

13. Khodabakhsh, Ali, et al. "Fake face detection methods: Can they be generalized?." 2018 international conference of the biometrics special interest group (BIOSIG). IEEE, 2018.
14. Korshunov, Pavel, and Sébastien Marcel. "Deepfakes: a new threat to face recognition? assessment and detection." arXiv preprint arXiv:1812.08685 (2018).
15. Rossler, Andreas, et al. "Faceforensics++: Learning to detect manipulated facial images." Proceedings of the IEEE/CVF international conference on computer vision. 2019.
16. Dolhansky, B. "The dee pfake detection challenge (DFDC) pre view dataset." arXiv preprint arXiv:1910.08854 (2019).
17. Li, Yuezun, et al. "Celeb-df: A large-scale challenging dataset for deepfake forensics." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
18. 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.
19. He, Yinan, et al. "Forgerynet: A versatile benchmark for comprehensive forgery analysis." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
20. Huang, Jiajun, et al. "Deepfake mnist+: a deepfake facial animation dataset." Proceedings of the IEEE/CVF international conference on computer vision. 2021.
21. Kwon, Patrick, et al. "Kodf: A large-scale korean deepfake detection dataset." Proceedings of the IEEE/CVF international conference on computer vision. 2021.
22. Frank, Joel, and Lea Schönherr. "Wavefake: A data set to facilitate audio deepfake detection." arXiv preprint arXiv:2111.02813 (2021).
23. Zi, Bojia, et al. "Wilddeepfake: A challenging real-world dataset for deepfake detection." Proceedings of the 28th ACM international conference on multimedia. 2020.
24. Le, Trung-Nghia, et al. "Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild." Proceedings of the IEEE/CVF international conference on computer vision. 2021.
25. Zhou, Tianfei, et al. "Face forensics in the wild." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
26. Narayan, Kartik, et al. "Deephy: On deepfake phylogeny." 2022 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2022.
27. Cai, Zhixi, et al. "Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization." 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA). IEEE, 2022.
28. Khalid, Hasam, et al. "FakeAVCeleb: A novel audio-video multimodal deepfake dataset." arXiv preprint arXiv:2108.05080 (2021).
29. Mittal, Govind et al. "Gotcha: Real-Time Video Deepfake Detection via Challenge-Response." 2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P) (2022): 1-20.
30. Jia, Shan, Xin Li, and Siwei Lyu. "Model attribution of face-swap deepfake videos." 2022 IEEE International Conference on Image Processing (ICIP). IEEE, 2022.
31. Cai, Zhixi, et al. "AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset." Proceedings of the 32nd ACM International Conference on Multimedia. 2024.
32. Bird, Jordan J., and Ahmad Lotfi. "Cifake: Image classification and explainable identification of ai-generated synthetic images." IEEE Access (2024).
33. Yan, Zhiyuan, et al. "Df40: Toward next-generation deepfake detection." arXiv preprint arXiv:2406.13495 (2024).
34. Tariq, Shahroz, et al. "Detecting both machine and human created fake face images in the wild." Proceedings of the 2nd international workshop on multimedia privacy and security. 2018.
35. McCloskey, Scott, and Michael Albright. "Detecting gan-generated imagery using color cues." arXiv preprint arXiv:1812.08247 (2018).
36. Li, Y. "Exposing deepfake videos by detecting face warping artif acts." arXiv preprint arXiv:1811.00656 (2018).
37. Li, Yuezun, Ming-Ching Chang, and Siwei Lyu. "In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking." arXiv preprint arXiv:1806.02877 (2018).
38. Güera, David, and Edward J. Delp. "Deepfake video detection using recurrent neural networks." 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, 2018.
39. Afchar, Darius, et al. "Mesonet: a compact facial video forgery detection network." 2018 IEEE international workshop on information forensics and security (WIFS). IEEE, 2018.
40. Koopman, Marissa, Andrea Macarulla Rodriguez, and Zeno Geradts. "Detection of deepfake video manipulation." The 20th Irish machine vision and image processing conference (IMVIP). 2018.
41. Do, Nhu-Tai, In-Seop Na, and Soo-Hyung Kim. "Forensics face detection from GANs using convolutional neural network." ISITC 2018 (2018): 376-379.

42. Badale, Anuj, et al. "Deep fake detection using neural networks." 15th IEEE international conference on advanced video and signal based surveillance (AVSS). Vol. 2. 2018.
43. Matern, Falko, Christian Riess, and Marc Stamminger. "Exploiting visual artifacts to expose deepfakes and face manipulations." 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). IEEE, 2019.
44. Yang, Xin, Yuezun Li, and Siwei Lyu. "Exposing deep fakes using inconsistent head poses." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.
45. Nguyen, Huy H., Junichi Yamagishi, and Isao Echizen. "Capsule-forensics: Using capsule networks to detect forged images and videos." ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2019.
46. Nguyen, Huy H., Junichi Yamagishi, and Isao Echizen. "Use of a capsule network to detect fake images and videos." arXiv preprint arXiv:1910.12467 (2019).
47. Sabir, Ekraam, et al. "Recurrent convolutional strategies for face manipulation detection in videos." Interfaces (GUI) 3.1 (2019): 80-87.
48. D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, 'ForensicTransfer: Weakly-supervised Domain Adaptation for Forgery Detection', arXiv [cs.CV]. 2019.
49. Akhtar, Zahid, and Dipankar Dasgupta. "A comparative evaluation of local feature descriptors for deepfakes detection." 2019 IEEE International Symposium on Technologies for Homeland Security (HST). IEEE, 2019.
50. Kharbat, Faten F., et al. "Image feature detectors for deepfake video detection." 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA). IEEE, 2019.
51. Dorević, Miljan, Milan Milivojević, and Ana Gavrovska. "Deepfake video analysis using SIFT features." 2019 27th telecommunications forum (TELFOR). IEEE, 2019.
52. Zhang, Weiguo, and Chenggang Zhao. "Exposing face-swap images based on deep learning and ELA detection." Proceedings. Vol. 46. No. 1. MDPI, 2019.
53. Kumar, Prabhat, Mayank Vatsa, and Richa Singh. "Detecting face2face facial reenactment in videos." Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2020.
54. Ranjan, Pranjal, Sarvesh Patil, and Faruk Kazi. "Improved generalizability of deep-fakes detection using transfer learning based CNN framework." 2020 3rd international conference on information and computer technologies (ICICT). IEEE, 2020.
55. Guarnera, Luca, Oliver Giudice, and Sebastiano Battiato. "Deepfake detection by analyzing convolutional traces." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020.
56. Younus, Mohammed Akram, and Taha Mohammed Hasan. "Effective and fast deepfake detection method based on haar wavelet transform." 2020 International Conference on Computer Science and Software Engineering (CSASE). IEEE, 2020.
57. Kawa, Piotr, and Piotr Syga. "A note on deepfake detection with low-resources." arXiv preprint arXiv:2006.05183 (2020).
58. De Lima, Oscar, et al. "Deepfake detection using spatiotemporal convolutional networks." arXiv preprint arXiv:2006.14749 (2020).
59. Wang, R., Ma, L., Xie, X., Huang, Y., Wang, J., & Liu, Y. (2019). FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces. ArXiv. https://arxiv.org/abs/1909.06122
60. Rana, Md Shohel, and Andrew H. Sung. "Deepfakestack: A deep ensemble-based learning technique for deepfake detection." 2020 7th IEEE international conference on cyber security and cloud computing (CSCloud)/2020 6th IEEE international conference on edge computing and scalable cloud (EdgeCom). IEEE, 2020.
61. Nirkin, Y., Wolf, L., Keller, Y., & Hassner, T. (2020). DeepFake Detection Based on the Discrepancy Between the Face and its Context. ArXiv. https://arxiv.org/abs/2008.12262
62. Chintha, Akash, et al. "Leveraging edges and optical flow on faces for deepfake detection." 2020 IEEE international joint conference on biometrics (IJCB). IEEE, 2020.
63. Du, Mengnan, et al. "Towards generalizable deepfake detection with locality-aware autoencoder." Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020.
64. Pan, Deng, et al. "Deepfake detection through deep learning." 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT). IEEE, 2020.
65. Dong, Xiaoyi, et al. "Identity-driven deepfake detection." arXiv preprint arXiv:2012.03930 (2020).
66. Xie, Daniel, et al. "Deepfake detection on publicly available datasets using modified AlexNet." 2020 IEEE symposium series on computational intelligence (SSCI). IEEE, 2020.
67. Trinh, Loc, et al. "Interpretable and trustworthy deepfake detection via dynamic prototypes." Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021.

68. Xu, Bozhi, et al. "DeepFake Videos Detection Based on Texture Features." Computers, Materials & Continua 68.1 (2021).
69. Wodajo, Deressa, and Solomon Atnafu. "Deepfake video detection using convolutional vision transformer." arXiv preprint arXiv:2102.11126 (2021).
70. Fung, Sheldon, et al. "Deepfakeucl: Deepfake detection via unsupervised contrastive learning." 2021 international joint conference on neural networks (IJCNN). IEEE, 2021.
71. Chen, Hong-Shuo, et al. "Defakehop: A light-weight high-performance deepfake detector." 2021 IEEE International conference on Multimedia and Expo (ICME). IEEE, 2021.
72. Kim, Minha, Shahroz Tariq, and Simon S. Woo. "Fretal: Generalizing deepfake detection using knowledge distillation and representation learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
73. Zhao, Hanqing, et al. "Multi-attentional deepfake detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
74. Ismail, Aya, et al. "A new deep learning-based methodology for video deepfake detection using XGBoost." Sensors 21.16 (2021): 5413.
75. Gu, Zhihao, et al. "Spatiotemporal inconsistency learning for deepfake video detection." Proceedings of the 29th ACM international conference on multimedia. 2021.
76. Zhao, Tianchen, et al. "Learning self-consistency for deepfake detection." Proceedings of the IEEE/CVF international conference on computer vision. 2021.
77. Das, Sowmen, et al. "Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
78. Zhao, Lei, et al. "MFF-Net: Deepfake detection network based on multi-feature fusion." Entropy 23.12 (2021): 1692.
79. Kaddar, Bachir, et al. "HCiT: Deepfake video detection using a hybrid model of CNN features and vision transformer." 2021 International Conference on Visual Communications and Image Processing (VCIP). IEEE, 2021.
80. Jeong, Yonghyun, et al. "Bihpf: Bilateral high-pass filters for robust deepfake detection." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022.
81. Jeong, Yonghyun, et al. "Frepgan: robust deepfake detection using frequency-level perturbations." Proceedings of the AAAI conference on artificial intelligence. Vol. 36. No. 1. 2022.
82. Khormali, Aminollah, and Jiann-Shiun Yuan. "DFDT: an end-to-end deepfake detection framework using vision transformer." Applied Sciences 12.6 (2022): 2953.
83. Chen, Liang, et al. "Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
84. Wang, Junke, et al. "M2tr: Multi-modal multi-scale transformers for deepfake detection." Proceedings of the 2022 international conference on multimedia retrieval. 2022.
85. Hu, Juan, et al. "Finfer: Frame inference-based deepfake detection for high-visual-quality videos." Proceedings of the AAAI conference on artificial intelligence. Vol. 36. No. 1. 2022.
86. Wang, Jian, Yunlian Sun, and Jinhui Tang. "LiSiam: Localization invariance Siamese network for deepfake detection." IEEE Transactions on Information Forensics and Security 17 (2022): 2425-2436.
87. Khan, Sohail Ahmed, and Duc-Tien Dang-Nguyen. "Hybrid transformer network for deepfake detection." Proceedings of the 19th international conference on content-based multimedia indexing. 2022.
88. Guan, Jiazhi, et al. "Delving into sequential patches for deepfake detection." Advances in Neural Information Processing Systems 35 (2022): 4517-4530.
89. Kingra, Staffy, Naveen Aggarwal, and Nirmal Kaur. "LBPNet: Exploiting texture descriptor for deepfake detection." Forensic Science International: Digital Investigation 42 (2022): 301452.
90. Ke, Jianpeng, and Lina Wang. "DF-UDetector: An effective method towards robust deepfake detection via feature restoration." Neural Networks 160 (2023): 216-226.
91. Yu, Yang, et al. "Augmented multi-scale spatiotemporal inconsistency magnifier for generalized deepfake detection." IEEE Transactions on Multimedia 25 (2023): 8487-8498.
92. Zhao, Cairong, et al. "ISTVT: interpretable spatial-temporal video transformer for deepfake detection." IEEE Transactions on Information Forensics and Security 18 (2023): 1335-1348.
93. BR, Shobha Rani, et al. "Deepfake video detection system using deep neural networks." 2023 IEEE international conference on integrated circuits and communication systems (ICICACS). IEEE, 2023.
94. Li, Xin, et al. "Artifacts-Disentangled Adversarial Learning for Deepfake Detection." IEEE Trans. Cir. and Sys. for Video Technol., vol. 33, no. 4, IEEE Press, Apr. 2023, pp. 1658–70, doi:10.1109/TCSVT.2022.3217950.
95. Lin, Hao, et al. "DeepFake detection with multi-scale convolution and vision transformer." Digital Signal Processing 134 (2023): 103895.

96. Wu, Jianghao, et al. "Interactive two-stream network across modalities for deepfake detection." IEEE Transactions on Circuits and Systems for Video Technology 33.11 (2023): 6418-6430.
97. Salvi, Davide, et al. "A robust approach to multimodal deepfake detection." Journal of Imaging 9.6 (2023): 122.
98. Wang, Tianyi, et al. "Deep convolutional pooling transformer for deepfake detection." ACM transactions on multimedia computing, communications and applications 19.6 (2023): 1-20.
99. Feng, Chao, Ziyang Chen, and Andrew Owens. "Self-supervised video forensics by audio-visual anomaly detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
100. Tan, Lingfeng, et al. "Deepfake video detection via facial action dependencies estimation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. No. 4. 2023.
101. Hou, Yang, et al. "Evading deepfake detectors via adversarial statistical consistency." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
102. Liang, Yufei, et al. "Hierarchical supervisions with two-stream network for Deepfake detection." Pattern Recognition Letters 172 (2023): 121-127.
103. Heo, Young-Jin, Woon-Ha Yeo, and Byung-Gyu Kim. "Deepfake detection algorithm based on improved vision transformer." Applied Intelligence 53.7 (2023): 7512-7527.
104. Aghasanli, Agil, Dmitry Kangin, and Plamen Angelov. "Interpretable-through-prototypes deepfake detection for diffusion models." Proceedings of the IEEE/CVF international conference on computer vision. 2023.
105. Guo, Zhiqing, et al. "Constructing new backbone networks via space-frequency interactive convolution for deepfake detection." IEEE Transactions on Information Forensics and Security (2023).
106. Shuai, Chao, et al. "Locate and verify: A two-stream network for improved deepfake detection." Proceedings of the 31st ACM International Conference on Multimedia. 2023.
107. Zou, Mian, et al. "Semantic Contextualization of Face Forgery: A New Definition, Dataset, and Detection Method." arXiv preprint arXiv:2405.08487 (2024).
108. Chen, Jin, et al. "ConfR: Conflict Resolving for Generalizable Deepfake Detection." 2024 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2024.
109. Zou, Mian, et al. "Semantics-Oriented Multitask Learning for DeepFake Detection: A Joint Embedding Approach." arXiv preprint arXiv:2408.16305 (2024).
110. She, Huimin, et al. "Using Graph Neural Networks to Improve Generalization Capability of the Models for Deepfake Detection." IEEE Transactions on Information Forensics and Security (2024).
111. Zhang, Rui, et al. "Generalized face forgery detection with self-supervised face geometry information analysis network." Applied Soft Computing 166 (2024): 112143.
112. Liu, Baoping, et al. "MeST-Former: Motion-enhanced Spatiotemporal Transformer for generalizable Deepfake detection." Neurocomputing 610 (2024): 128588.
113. Wang, Fei, et al. "Multi-to-Binary: A Generalizable Deepfake Detection Approach with Multi-Classification Guidance." Available at SSRN 4954695.
114. Fahad, Muhammad, et al. "Advanced deepfake detection with enhanced Resnet-18 and multilayer CNN max pooling." The Visual Computer (2024): 1-14.
115. Zhang, Kuiyuan, et al. "Boosting Deepfake Detection Generalizability via Expansive Learning and Confidence Judgement." IEEE Transactions on Circuits and Systems for Video Technology (2024).
116. Zheng, JunShuai, et al. "Deepfake Detection With Combined Unsupervised-Supervised Contrastive Learning." 2024 IEEE International Conference on Image Processing (ICIP). IEEE, 2024.
117. Zhang, Dengyong, et al. "Face Forgery Detection Based on Fine-grained Clues and Noise Inconsistency." IEEE Transactions on Artificial Intelligence (2024).
118. Lu, Lin, et al. "Deepfake Detection Via Separable Self-Consistency Learning." 2024 IEEE International Conference on Image Processing (ICIP). IEEE, 2024.
119. Ain, Qurat Ul, et al. "Exposing the Limits of Deepfake Detection using novel Facial mole attack: A Perceptual Black-Box Adversarial Attack Study." 2024 IEEE International Conference on Image Processing (ICIP). IEEE, 2024.
120. Lin, Yuzhen, et al. "Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection." European Conference on Computer Vision. Springer, Cham, 2025.
121. Alazwari, Sana, et al. "Artificial rabbits optimization with transfer learning based deepfake detection model for biometric applications." Ain Shams Engineering Journal (2024): 103057.
122. Zakkam, John, et al. "CoDeiT: Contrastive Data-Efficient Transformers for Deepfake Detection." International Conference on Pattern Recognition. Springer, Cham, 2025.

Authors

JINGJING RAO received the B.E. degree in software engineering from the Dalian Neusoft University of Information, in 2016, and the M.E. Eng. degree in information science and engineering from Ritsumeikan University, in 2022, where she is currently pursuing the Ph.D. degree in information science and engineering. Her research interests include digital forensics and computer vision.

Tetsutaro Uehara received the B.E., M.E., and D.Eng. degrees from Kyoto University, in 1990, 1992, and 1996, respectively. He was an Assistant Professor at the Faculty of Systems Engineering, Wakayama University, from 1996 to 2003. From 2003 to 2005, he was an Associate Professor at the Center for Information Technology, Graduate School of Engineering, Kyoto University. From 2006 to 2011, he was an Associate Professor at the Academic Center for Computing and Media Studies, Kyoto University. From 2011 to 2013, he was the Deputy Director of the Standardization Division in the Ministry of Internal Affairs and Communication, Japan. He has been a Professor with the College of Information Science and Engineering, Ritsumeikan University, since 2013. He has also been the Vice-President of the Institute of Digital Forensics, since 2017. His research interests include systems security, digital forensics, privacy, education in information ethics, and information system management in local government.