Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 6
ISET 2024

Key Technologies and Typical Applications in Deepfake Detection

Bonan Feng

School of Economics and Management, Beijing University of Chemical Technology, Beijing, China

2021070002@buct.edu.cn

Abstract. As deepfaking technology advances, it increasingly facilitates the seamless splicing of a person's voice, facial expressions, and body movements into source images or videos to create hyper-realistic virtual content. This growing technological sophistication, paired with the open-source nature of neural network algorithms, has unfortunately led to the exploitation of deepfake technology by individuals for illicit activities, seeking financial gain. This situation highlights the urgent necessity to thoroughly understand and vigilantly monitor the evolution of deepfake technology. This paper delves into the fundamental concepts and operational mechanisms underlying contemporary deepfake detection methods. It aims to augment the comprehension of these technologies and assesses the strengths and weaknesses of current detection approaches. This evaluation provides critical insights into their operational efficacy and reliability in differentiating genuine from manipulated content. Furthermore, the paper explores the challenges that current technologies face, such as the rapid evolution of deepfake methods that may outpace detection capabilities. By thoroughly analyzing these technologies, the paper endeavors to offer a comprehensive overview and projects future trends in deepfake detection. It anticipates developments that could significantly enhance the efficacy of countermeasures and thus bolster the fight against digital identity fraud and misinformation, safeguarding digital media integrity in an increasingly synthetic landscape.

Keywords: Information Tampering; Deepfake; Feature Extraction.

1. Introduction

The rapid advancement of deepfake technology has unlocked unprecedented possibilities for generating and editing digital media, yet it has also introduced numerous detrimental effects. Deepfake, a term for false media content created using artificial intelligence, encompasses altered images, videos, and audio. Such manipulated content is often used maliciously to disseminate misinformation on social media, engage in cyber fraud, slander individuals, or fabricate statements by significant public figures, leading to the propagation of fake news. These activities severely threaten social, political, and personal trust and security, raising critical concerns about the ethical implications and potential harms associated with this technology. As deepfakes become increasingly indistinguishable from authentic content, they present a formidable challenge in maintaining the integrity of information across various platforms.

Figure 1. Faked audio of national leaders (Photo credit: Original).

Figure 2. Deepfake changes the face image (Photo credit: Original).

As illustrated in Figures 1 and 2, the image of important leaders is tampered with through deepfake, so as to release false news and seriously influence the social order.

In order to meet the challenge of deepfake, researchers actively explore the detection technology of deepfake. The goal of deepfake detection is to differentiate authentic media from fabricated content, aiming to expose misinformation and safeguard the public from deception. The development of this technology is critical to maintaining the authenticity and credibility of digital content. Deepfake detection technology is the result of interdisciplinary research, involving computer vision, audio processing, signal processing, machine learning and artificial intelligence. Researchers have developed a series of innovative algorithms and methods to detect deepfake by deeply understanding the technical principles and potential feature patterns of deepfake generation.

This paper will introduce the detection technology based on spatial inconsistency, time-space inconsistency, frequency-domain clue mining and multi-mode deepfake detection, introduce its basic operation logic, summarize the advantages and disadvantages of each detection method, and analyze and summarize the advantages and disadvantages of the current detection methods based on the advantages and disadvantages of each method. In addition, the future development of detection methods will be discussed to facilitate the further improvement and application of deepfake detection technology. Through the research and application of deepfake detection technology, the authenticity and credibility of digital media can be better protected, and the potential risks of deepfake to society and individuals can be reduced. The development of this technology is essential to maintaining security and trust in the digital world,.Exploring the challenges posed by deepfake provides key tools and methods.

2. Analysis of the Status Quo of Deepfake Detection Technology

The traditional depth forgery detection method mainly focuses on the texture or time sequence difference in the video, and learns the difference between the forged video and the real video based on the Convolutional Neural Network (CNN), the Recurrent Neural Network (RNN) or the Vision Transformer (ViT). in that follow.

2.1. Detection of Deep Forgery Based on Spatial Inconsistency Measure

2.1.1. Basic operating logic.

Judging by extracting the details of the joint between the forged image and the real image, wherein a part of the glasses frame exists in the real image in (a) in the Figure 3, but the false image lacks the main part of the glasses during splicing, and the glasses frame is reserved; (b) There is a skin color inconsistency between the real image and the forged image. The detection method based on spatial inconsistency is to judge the difference between the real image and the fake image.

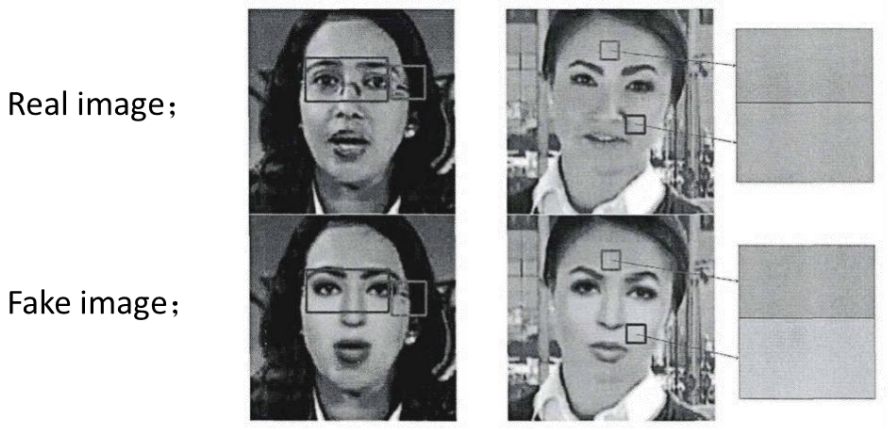

Figure 3. Principle of detection based on spatial inconsistency (Photo credit: Original).

2.1.2. Classification and characteristics of deep forgery detection based on spatial inconsistency.

In the early stage, some models with excellent performance in the field of image classification are used for detection, such as Xception, EfficiencyNet [1, 2]. Take Xception as an example, and its basic operation process is shown Table 1:

Table 1. The running logic of Xception.

Stage	Enter Size	Output Size	Operation
convolution layer	Input Image Size	Feature Map Size	convolution operation
Residual Module 1	Feature Map Size	Feature Map Size	deep separable convolution
Residual Module N	Feature Map Size	Feature Map Size	deep separable convolution
Global Average Pooling	Feature Map Size	1x1x number of channels	Pooling Operations
fully connected layer	1x1x number of channels	class probability distribution	Full Connect Operation

Convolution layer and batch normalization: the input x is convolved to obtain the characteristic graph f1. Apply the batch normalization operation to normalize the feature map f1. f1 = Conv2D(x), f1 = BatchNormalization(f1).

Depth-separable convolution block: repeatedly stacking a plurality of depth-separable convolution blocks, each block containing the following operations:

Deep convolution: Apply a deep convolution operation on feature map f1 to generate feature map f2. f2 = DepthwiseConv2D(f1).

Batch Normalization: Normalizes the feature map f2. f2 = BatchNormalization(f2).

Point-by-point convolution: Apply a pointwise convolution operation on feature map f2 to produce feature map f3. f3 = PointwiseConv2D(f2).

Batch Normalization: Normalizes the feature map f3. f3 = BatchNormalization(f3).

Residual connection: Carry out residual connection between the feature map f3 and the input x to derive the feature map f4. f4 = f3 + x f1 = f4.

Multi-level feature fusion: Use multiple depth-separable convolution blocks to extract features of different scales and fuse them together.

d.global average pooling and full connection layer: carrying out global average pooling operation on the fused feature map to obtain a vector v. The vector v is input to the fully connected layer for classification, and the prediction result is obtained. v = GlobalAveragePooling2D(f) output = Dense(v).

Although these models also show good performance in the task of forgery detection, they are not specially designed for deep forgery detection. When the models detect the false information synthesized by the forgery method which has not appeared in the training set, the detection performance of the models will drop sharply, and the robustness is insufficient when faced with common disturbance methods such as image compression.

As false information becomes more and more realistic, scholars believe that the difference between true information and false information exists in subtle local details. Therefore, a texture enhancement module, an attention generation module and a bilinear attention pooling module are proposed to guide the model to pay more attention to the texture details of the face. and the texture enhancement module is used for enhancing the texture information extracted by the model in the shallow network. The attention generation module can generate multiple attention maps, so that the princess network is focused on various positions of the image, and the subtle texture artifacts are fully mined. The bilinear attention pool module ensures the non coincidence of interest regions between attention maps, thus ensuring more comprehensive extraction of image details and textures. Although it has been improved, it tends to overfit the forgery artifacts of the feature forgery method, and the model detection performance decreases rapidly when faced with the forgery method not seen during training.

In order to improve the generalization of detection technology, scholars began to use some methods to impose some operations on real faces to generate self-counterfeiting information to avoid overfitting to specific forgery methods. For example, Face X-Ray model proposed by Li begins to focus on improving the generalization of the model, and there are also pair-wise self-consistency learnin g, PCL proposed by Zhao and self-blended images, SB proposed by Shiohara [3, 4]. The next section focuses on the Face X-Ray model. Face X-Ray model actively generates a false face during training, and acquires the gray image of the mixed boundary of the face. During training, in addition to the true-false binary classification prediction for the human face image, the binary classification is also performed pixel by pixel for the gray image of the fusion boundary of the generated image. The operation logic is as shown in the Figure 4.

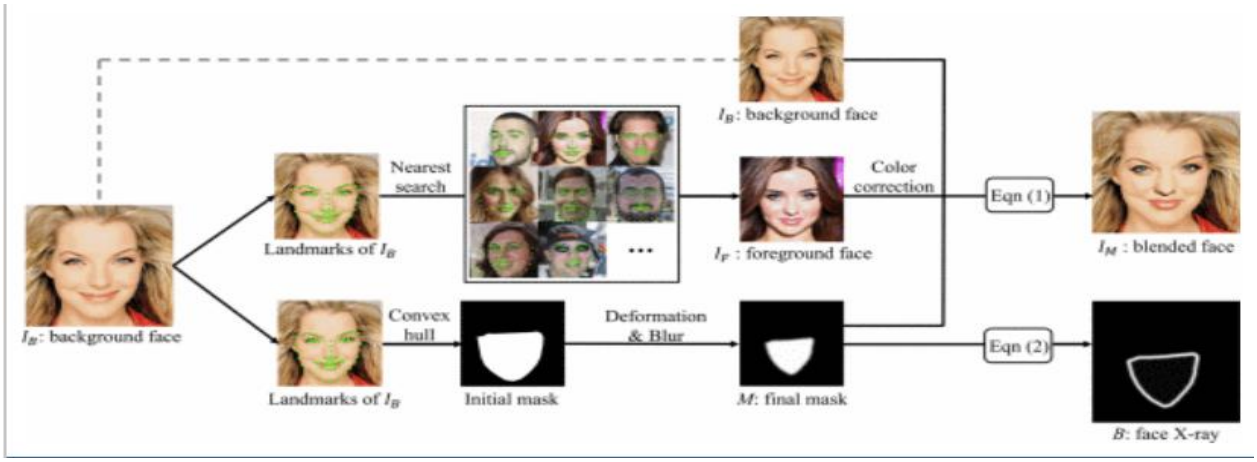

Figure 4. Operation diagram of Face X-Ray model (Photo credit: Original).

The basic operation formula of Face X-Ray is [5];

$$Lb=-\sum\{I,B\}\in D1N\sum i,j(Bi,jlogB\hat{}i,j+(1-Bi,j)log(1-B\hat{}i,j))$$
$$Lc=-\sum\{I,c\}\in D(clog(c\hat{})+(1-c)log(1-c\hat{})) \quad (1)$$
$$L=\lambda Lb+Lc$$

This method does not depend on the detection of artifact traces of a specific forgery method, so the method shows good generalization when detecting an unknown forgery method. However, due to the self-forgery process, such methods do not perform well in the detection of false images synthesized by the whole face synthesis method.

in recent research in China, some scholar make full use of that spatial relations in the image,

The characteristics of the counterfeit area and the real area are obtained.

Calculation process:

$$Fman,Fori=fpr(fe(x)) \quad (2)$$

$$s=frr(Fman,Fori) \quad (3)$$

FPR stands for PR module, fe stands for feature extractor, Fman and Fori stand for fake area feature and real area feature respectively.

Then you need to use (2) Rr module to compare Fman and Fori to capture the difference between the two regions, so as to measure their spatial inconsistency, so as to realize the detection.

According to the above, since the video face depth forgery method does not directly generate a complete image, but generates a partial image and fuses the partial image into the real image, the counterfeit part is different from the real part. such as a strong generalization depth forgery face detection technology based on local anomalies, which inputs a face image to be detected into a trained local anomaly detection network to carry out true-false classification. In the method, the local anomalies are mined from the depth feature map of the face image by performing second-order local anomaly learning in the spatial domain so as to realize the detection. Although the detection performance of this type of method is good under certain conditions, there are also some problems such as low generalization and poor detection performance for the whole-face synthetic image. Therefore, the advantages and disadvantages of the spatial inconsistency detection method are summarized as follows:

Advantages: High detection accuracy for false images with specific trained data sets.

Disadvantages: The detection performance is poor and lack of generalization when the false image synthesized by the forgery method which does not appear in the training set and the whole-face forged image is detected.

2.2. Deep Forgery Detection Based on Spatiotemporal Inconsistency Modeling

Basic operating logic shown in Figure 5:

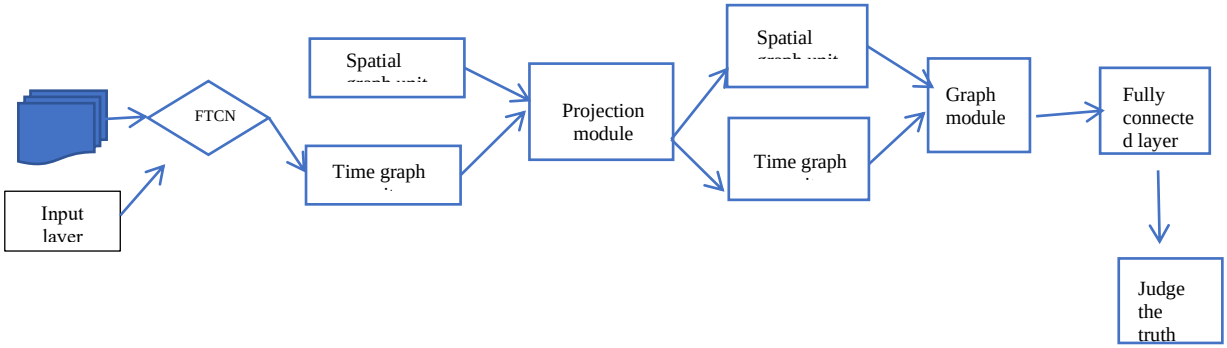

Figure 5. Based on spatiotemporal inconsistency detection method flow (Photo credit: Original).

Most of the counterfeiting techniques are to generate a fake image frame by frame and then splice it into the real video, so it is inevitable that there will be a discontinuity in timing. For example, one method uses a forgery technique to create temporal inconsistencies between frames of video, combining a bidirectional long-short-term memory network and a conditional random field algorithm to achieve detection. As shown in the Figure 6, there are significant inconsistencies in the eyebrow area at frames 1 and 2, 3 and 2, 3 and 4.

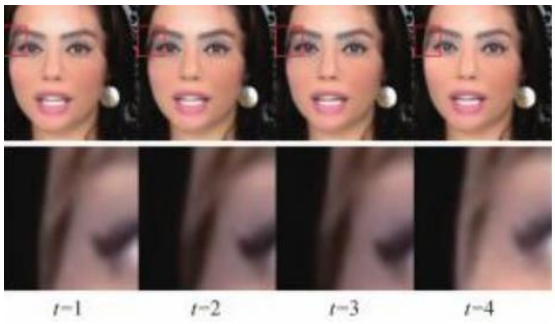

Figure 6. Based on frequency domain clue mining features (Photo credit: Original).

Therefore, scholars began to explore the timing inconsistency. For example, Masi and others proposed a two-branch network detection model, in which one branch is used to extract the dynamic timing inconsistency of consecutive frames, and the other branch uses the Laplacian of Gaussian operator to amplify the artifact details. At the same time, due to the strong correlation between forgery detection and anomaly detection, Masi et al. also introduced the deep support vector data description to improve the intra-class compactness of the real face and the discrimination between the forgery information and the real information, thus enhancing the generalization of the model [6]. Because of its dual-branch structure, and the detection mechanism is timing inconsistency, its generalization performance is good. However, this kind of timing inconsistency is easily interfered by noise, compression and other factors, which leads to the detection clues being destroyed, so the robustness of this kind of method is poor.

Advantages: Functions in anomaly detection tasks can be introduced, with double branches: extracting dynamic timing inconsistencies of consecutive frames, magnifying artifact details. So the generalization ability is good.

Disadvantages: Timing inconsistency is easily disturbed by noise, compression and other factors, resulting in the detection clues being affected, so the robustness is poor.

2.3. Detection of Deep Forgery Based on Falsification Cue Mining in Frequency Domain

Basic operation logic;

Take Filterbase, a filter based on artificially set parameters, as an example; For the input face video sequence X= {x0,x1.... xN-1}, where N is the number of frames of the video, each set of pixel sequences in the same position and different frames are respectively transformed by discrete cosine

transform to obtain the frequency domain representation Fx={f0, f1.... fN-1}. And the artificial cut-off frequency fbase is used to enhance the temporal frequency domain forgery clues, so that the model can fully learn the frequency domain forgery features.

$$Filter_{base}(f)=\begin{cases}0 & f < f_\theta \\ 1 & f \geq f_\theta\end{cases} \quad (4)$$

It can be seen from the above example that the video transmitted over the network many times is usually compressed many times, and the artifacts are hardly visible, but the frequency domain can be mined to find the details of the artifacts very well. For example, one method uses a frequency domain image as a unit to train a convolutional neural network detection model, so that no matter how many redundant features are generated by video compression in the time domain, the detection can be carried out smoothly. In order to more comprehensively capture the counterfeiting artifacts in the frequency domain, Li and others propose a module for adaptive frequency feature generation that extracts differential features from various frequency bands in a learnable manner. At the same time, considering the different feature distribution of different forgery methods, a single center loss (SCL) is proposed to improve the intra-class compactness of real face and increase the intra-class difference between real face and fake face. In addition, in order to capture artifact traces with different scales, Wang et al. proposed a multi-modal and multi-scale Transformer model to detect local artifact details of different levels in the image [7]. The above-mentioned frequency domain-based method has strong forgery detection capability in the face of highly compressed forgery images, but the forgery detection capability in the face of unknown forgery methods is still drastically reduced.

Advantages: Redundancy caused by multiple propagation compression can be reduced, and high detection accuracy can still be maintained in the face of highly compressed forged images.

Disadvantages; In the face of unknown forgery methods, the detection capability drops dramatically.

2.4. Multimodality Detection Method

Basic operation logic; Take RealForensics as an example, as shown in the Figure 7.

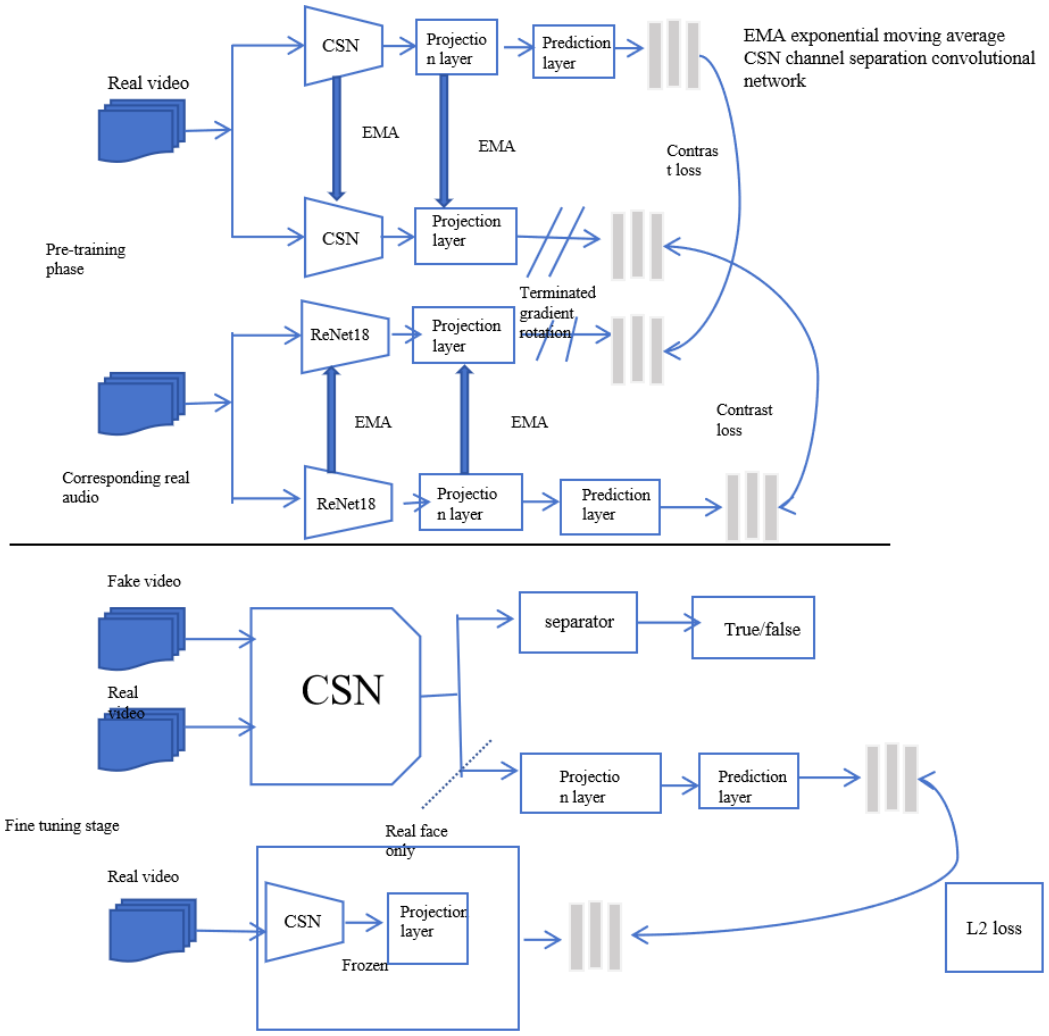

Figure 7. RealForensics running diagram (Photo credit: Original).

Haliassos et al. proposed the LipForensics model, which was first lip-read pre-trained on the LRW dataset and then fine-tuned on the forgery detection dataset in order to ensure that the model learns the natural lip motion features [8, 9]. The model shows excellent performance when tested across counterfeiting methods, and the method is far superior to other methods in robustness testing due to the fact that the lip semantics are not easily disturbed. However, the model only focuses on the mouth region of the human face and ignores the information from other facial regions, which limits the performance of the detection model. Haliassos et al. also proposed Real Forensics model, which extracts audio features through residual network (resnet) and video features through channel-separated convolutional networks (CSN) to acquire the inherent relationship between facial movements and audio in authentic videos through self-supervised learning, for example, facial expression changes with audio, to make up for the shortcomings of previous models [10, 11].

This kind of method shows more excellent performance in robustness and generalization, but because it needs pre-training task on large-scale data set first, it needs a lot of computational power support, which sets a great threshold for follow-up work [12].

Advantages: Because lip semantics are not susceptible to common interference, its robustness and detection accuracy are high.

Disadvantages: A lot of computing power is required, and the threshold for follow-up work is high.

2.5. Summarize The Strengths and Weaknesses of Current Test Methods

Through the respective analysis and research of the above mainstream methods, at present, no matter the spatial inconsistency, spatiotemporal inconsistency, frequency domain clue mining and multi-

mode detection have their own advantages and disadvantages, and most of them can achieve ideal results when they are suitable for different situations [13]. However, due to the difference between the actual situation and the training data set and the different methods of deep forgery, the generalization of the current mainstream detection technology is poor. This is summarized in Table 2 [14].

Table 2. Detection method summary.

Test Method	Advantages	Disadvantages
Detection of Deep Forgery Based on Spatial Inconsistency Measure	For false images of a specific trained data set, the detection accuracy rate is high	The false image synthesized by the forgery method and the whole face which have not appeared in the training set are detected Time measurement, poor detection performance, lack of generalization
Detection of Deep Forgery Based on Spatio temporal Inconsistency Modeling	It has two branches to magnify artifact detail. The generalization ability is better than other methods.	It is easy to be disturbed by noise, compression and other factors, resulting in the detection clue being affected and the robustness is poor
Depth Falsification Based on Falsification Cue Mining in Frequency Domain manufacturing inspection	In the face of highly compressed counterfeit images, high detection accuracy is kept.	In the face of unknown forgery methods, the detection capability drops dramatically
Multi-mode detection method	Lip semantics is not susceptible to common interference, and its robustness and detection accuracy are high	It requires a lot of computing power support, and the threshold of follow-up work is high

Through the above, the specific advantages and disadvantages of today's detection methods can be summarized as follows:

Pros: High detection accuracy on a single compression rate for a single dataset, increasing as the dataset is updated. And now the detection angles are more and more abundant, and the extraction of forged details is more and more diversified.

Disadvantages: It is difficult to learn how to learn the common features of face images under different forgery methods and different resolutions. It is difficult to achieve the unity of generalization and detection accuracy. Even if the unity is achieved, a lot of extra work is needed to complete, and the cost performance is not very high.

3. Development Trend of Speculative Detection Technology

3.1. How to Extract More Generalized and Distinguished Features in The Future

How to accurately capture the counterfeit trace in video is one of the key problems of deep forgery detection technology. Today's detection technology optimizes the extraction of counterfeit features from the directions of spatial inconsistency, temporal-spatial inconsistency, frequency-domain cue, multi-mode and so on. However, at present, these mainstream deep forgery detection technologies

still do not reach the ideal target, so how to extract more distinctive and generalized forgery features is one of the development directions of the future detection technology.

How to accurately capture the counterfeit trace in video is one of the key problems of deep forgery detection technology. Today's detection technology optimizes the extraction of counterfeit features from the directions of spatial inconsistency, temporal-spatial inconsistency, frequency-domain cue, multi-mode and so on. However, at present, these mainstream deep forgery detection technologies still do not reach the ideal target, so how to extract more distinctive and generalized forgery features is one of the development directions of the future detection technology.

3.2. Deep Forgery Detection Technology Aiming at Specific Person

In real life, the in-depth forged fake news aiming at key figures such as national important leaders, military officials, large-scale enterprise executives and the like has high social harm and safety risks, so the performance requirements of the deep-forgery detection technology aiming at specific characters are more strict. As shown in the Figure 8, the faces of important leaders and actors in the United States are replaced. In this way, there are not a few people who tamper with the faces of important characters, and the damage is huge. Therefore, the deep forgery technology aiming at some specific characters becomes one of the development directions of detection technology in the future.

Figure 8. Deep forgery of important characters (Photo credit: Original).

In the general deep forgery detection process, the identity of the detected person cannot be known in advance, and the detection of the specific person is carried out on the premise that the identity is known, so that the important facial detail characteristics of the specific person can be input in advance, and the detection accuracy of the model for the person can be trained so as to improve the detection performance. Therefore, in order to better protect the information security of specific characters, it is worth further development to explore the deep forgery detection technology of specific characters.

3.3. The Training Data Set of Deep Forgery Detection Technology Needs to Be Oriented to The Actual Needs

The deep forgery detection technology is a technology which needs to be applied in practice to solve practical problems. However, in the training data set of the current mainstream deep-forgery detection technology, the real data are often selected or prepared in advance, and the deep-forgery data are generally generated through a few mature deep forgery methods. this results in a training data set with fewer character elements, fewer scene changes, consistent video content, counterfeiters, Due to the lack of diversity, such training data sets are difficult to simulate real application scenarios effectively, resulting in poor performance in detecting deep forgery techniques in real scenarios.

Therefore, it is of great significance to design and collect the training data set of real data and forged data with pertinence and diversity in combination with the actual requirements for promoting the application of deep forgery detection technology.

3.4. Deep-Forgery Detection Technology Needs to Develop to Active Defense

Currently, mainstream deep forgery detection technologies operate by identifying altered data post-manipulation, constituting a reactive form of defense. This approach often leaves systems vulnerable until after forgeries are committed. To shift towards a more proactive stance against forgery, innovative strategies need to be developed. One such strategy could involve integrating advanced watermarking technology into authentic images before they are compromised. These watermarks

would be designed to be indelible, resisting removal or alteration even by sophisticated forgery techniques. Additionally, incorporating disturbance noise into real video footage, based on methods that fortify neural networks against attacks, could impair the functionality of forgery technologies, thereby preempting the manipulation process. Such measures would not only disrupt the capabilities of forgery technologies but also ensure the proactive protection of genuine visual content.

4. Conclusion

Deep forgery technology is currently at the forefront of social and technological research due to its profound impact on society. While it offers numerous benefits, such as enhancing creative processes and simulating realistic scenarios for training and education, it also poses significant risks by enabling the creation of misleading or fraudulent content. The dual-edged nature of this technology places a critical emphasis on the development and refinement of deep forgery detection methods. To mitigate the potential harms of deep forgery, the effectiveness of detection technologies must continually evolve to keep pace with advancements in forgery methods. This entails not only improving the accuracy of detection algorithms but also enhancing their ability to generalize across different scenarios and media types. Future research must focus on developing robust, adaptive systems that can quickly and accurately identify deep fakes, ensuring they are used ethically and responsibly. As the technology advances, the refinement of detection methods will play a pivotal role in safeguarding information integrity and maintaining public trust in digital media.

References

[1] H.Q. Zhao, T.Y. Wei, W.B. Zhou, et al., Multi-attentional deepfake detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Nashville, USA, 2021, 2185-2194.

[2] L.Z. Li, J.M. Bao, T. Zhang, et al., Face X-ray for more general face forgery detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Seattle, USA, 2020, 5000-5009.

[3] T.C. Zhao, X. Xu, M.Z. Xu, et al., Learning self-consistency for deepfake detection, Proceedings of the IEEE/CVF International Conference on Computer Vision, IEEE, Montreal, Canada, 2021, 15003-15013.

[4] K. Shiohara, T. Yamasaki, Detecting deepfakes with self-blended images, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, New Orleans, USA, 2022, 18699-18708.

[5] I.Masi, A. Killekar, R.M. Mascarenhas, et al., Two-branch current network for isolating deepfakes in videos, Proceedings of the 16th European Conference on Computer Vision, Springer, Glasgow, UK, 2020, 667-684.

[6] L. Ruff, N. G-Érnit, L. Deecke, et al., Deep one-class classification, Proceedings of the 35th International Conference on Machine Learning, PMLR, Stockholm, Sweden, 2018, 4390-4399.

[7] J.K. Wang, Z.X. Wu, W.H. Ouyang, et al., M2TR: Multi-modal multi-scale transformers for deepfake detection, Proceedings of the International Conference on Multimedia Retrieval, ACM, Newark, USA, 2022, 615-623.

[8] A.Haliassos, K. Vougioukas, S. Petridis, et al., Lips don't lie: A generalizable and robust approach to face forgery detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Nashville, USA, 2021, 5037-5047.

[9] J.S. Chung, A. Zisserman, Lip reading in the wild, Proceedings of the 13th Asian Conference on Computer Vision, Springer, Taipei, China, 2017, 87-103.

[10] K.M. He, X.Y. Zhang, S.Q. Ren, et al., Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Las Vegas, USA, 2016, 770-778.

[11] D. Tran, H. Wang, M. Feiszli, et al., Video classification with channel-separated volatile networks, Proceedings of the IEEE/CVF International Conference on Computer Vision, IEEE, Seoul, Korea (South), 2019, 5551-5560.

[12] L. Li, J. Bao, T. Zhang, et al., Face X-Ray for More General Face Forgery Detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Seattle, WA, USA, 2020, 5000-5009, https://doi.org/10.1109/CVPR42600.2020.00505.

[13] Z. Shang, Research on Key Technologies of Video Face Depth Forgery Detection, Doctoral dissertation, University of Science and Technology of China, 2023, https://doi.org/10.27517.

[14] L.Yu, Summary of Deep Face Forgery and Detection Technology, Journal of Tsinghua University (Natural Science Edition), 63(09), 2023, 1350-1365.