Face Forgery Detection Technology Based on Deep Learning

Tingyu Liang *

Department of International College, School of Chongqing University of Posts and Telecommunications, Chongqing, 404100, China

* Corresponding Author Email: tingyuliang666@outlook.com

Abstract. The rapid evolution of generative artificial intelligence has significantly transformed the way image and video content is produced. Deepfake facial recognition, a prominent example, has become a research hotspot in computer vision and multimedia security. However, its potential for widespread application also carries the risk of potential misuse, sparking continued attention from both academia and industry. This paper systematically reviews the face forgery detection technology based on deep learning. With the development of artificial intelligence technologies such as Generative Adversarial Networks (GAN) and diffusion models, face forgery content has brought about serious security and ethical issues while bringing application possibilities. Traditional image forensics methods rely on manual features and have limited generalization ability, while deep learning significantly improves detection performance through end-to-end feature learning. This paper reviews and analyzes the principles and characteristics of mainstream detection methods from the perspectives of physiological signal features, image tampering traces, GAN generation features and image-level learning strategies, and summarizes the challenges that the current models still face in terms of generalization ability and anti-interference ability. Finally, the future research directions such as cross-domain detection framework, multi-modal fusion and active defense are proposed.

Keywords: Deep Learning, Face Forgery, Detection Technology.

1. Introduction

With the rapid development of artificial intelligence technology, especially the wide application of generative artificial intelligence technologies such as GAN and diffusion models, face forgery technologies (such as Deepfake, Face Swap, etc.) have shown great potential in the fields of film and television production, entertainment and social networking. However, such technologies also bring serious security and ethical challenges, such as false information dissemination, identity fraud, and privacy violations, which pose a serious threat to the social trust system. Therefore, the development of efficient and robust face forgery detection technology has become a research hotspot in the field of multimedia security and artificial intelligence governance.

Traditional image forensics methods mainly rely on hand-designed features. Although they perform well in specific tampering types, their generalization ability and robustness are significantly insufficient in the face of highly realistic deep forgery content. In recent years, the method based on deep learning has made significant progress in detection accuracy, generalization ability and anti-interference by automatically learning the characteristics of forged traces and has gradually become the mainstream research direction.

This paper systematically reviews the face forgery detection technology based on deep learning. From the perspectives of physiological signal feature analysis, image tampering trace capture, Generative Adversarial Networks (GAN) generation feature recognition, and image-level learning strategy, the principles, characteristics, and performance of current mainstream methods are sorted out and analyzed. The aim is to provide technical reference and development direction for subsequent research.

2. Datasets

Deep forgery detection research relies on a variety of public data sets. The main data sets and their core features are shown in Table 1.

Table 1. Summary of Major Deepfake Detection Datasets

Dataset name	type	Tampering method	scale	Characteristics and limitations
FaceForensics++ (FF++)	video frequency	Deepfakes, Face2Face, FaceSwap, Neural Textures	1,000 original videos + 4,000 forged videos	A variety of tampering types provide a compressed version (C0 / C23 / C40); however, the generated quality is low, and there are obvious boundary traces.
Celeb-DF	video frequency	Deepfakes	408 original video+795 forged video	High resolution, low flicker; however, the type of tampering is single, limited to face change.
DFDC	video frequency	Various unknown methods	119,154 videos (true-false ratio 1:5)	Large-scale, multi-scene (strong light/side face / multi-person); however, the tampering method is not disclosed, and the cross-domain generalization challenge is large.
DeeperForensics-1.0	video frequency	DeepFake Variational Auto-Encoder	50,000 original + 10,000 forged videos	Large-scale, multi-light conditions; however, the low proportion of forgery (5: 1) may exacerbate the problem of sample imbalance.
Deepfake-TIMIT	video frequency	Faceswap-GAN	GAN 640 video (high-definition / low-definition 320 each)	The first GAN generation data set, but the resolution is low (128 × 128), and the boundary marks are obvious.
UADFV	video frequency	FakeAPP	49 original + 49 forged videos	Early data set; however, the scale is small, the resolution is low, and the generated quality is rough.

As can be seen from Table 1, the existing data sets are mainly video-based and cover a variety of tampering methods, but there are still deficiencies in generation quality, real scene simulation (such as compression, noise), sample balance, and adversarial sample coverage. In the future, it is necessary to build a larger scale, closer to the real application scenario, and a more challenging data set to improve the robustness of the detection model.

3. Face Forgery Detection Technology

3.1. Method Based On Physiological Signal Characteristics

In the research of deep forgery video detection using physiological signal features, deep learning technology is applied to the extraction and classification of key features. Yang et al. observed through experiments that although the face image generated based on GAN shows a high degree of authenticity in local details, the spatial distribution of facial key points is significantly different from that of real faces [1]. Further analysis shows that there is still a gap between such synthetic images and real data in terms of overall naturalness and structural coherence. To this end, this work establishes an effective detection framework by extracting normalized key point coordinates to form feature vectors and combining a Support Vector Machine (SVM) classification algorithm. Li et al. designed a long-term recurrent convolutional network (combined with Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN)), which is specifically used to identify the opening and closing state of the human eye in the video, and detect forged videos by analyzing the abnormal blink frequency and pattern [2]. Zhu Xintong et al. designed a tamper and forgery image detection algorithm based on image texture features [3]. As shown in Fig.1, the algorithm innovatively combines the first-order gradient edge texture image extracted by the Scharr operator in the Cb and Cr channels and the second-order gradient edge texture image extracted by the Laplacian operator in the G channel. The gray level co-occurrence matrix (GLCM) is used to fuse and extract

the features of the two images. Finally, Efficient Net is used to realize the classification and detection of tampering and deep forged images.

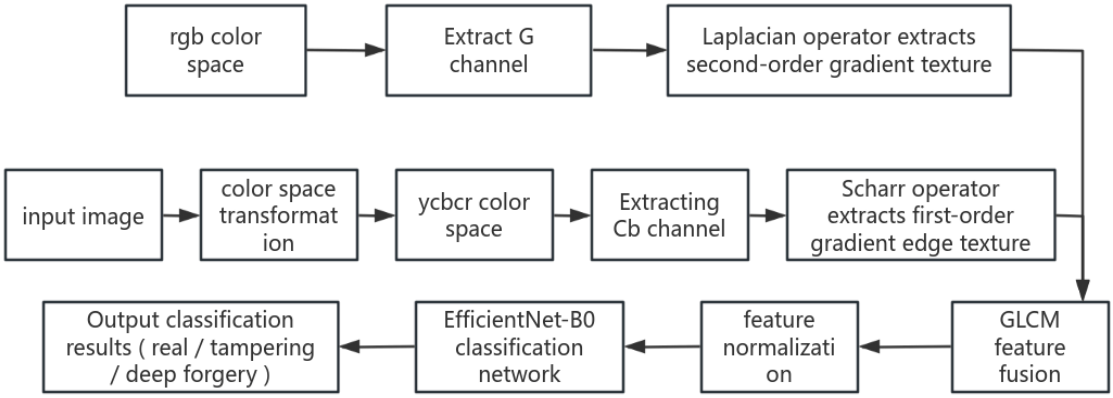

Figure 1. Tampering and forgery image detection process based on image texture features [3]

3.2. Method Based On Image Tampering Traces

Aiming at the shortcomings of the early deep forgery images in detail, the researchers use deep learning to capture the artificial traces left in the generation process for detection. Matern et al. focused on the global inconsistency of forged faces, such as color asymmetry of eyes and nose, illumination inconsistency, such as lack of reflective details of eyes, and geometric position inconsistency, such as blurred details of teeth. They extract the features of these specific regions (teeth, eyes, etc.) and train a multi-layer perceptron (MLP) for classification [4]. However, the redundant feature interference in the existing methods (for example, the position of the facial features is basically the same) and the current mainstream face forgery detection data sets generally have sample singularity (that is, by training the continuous frame sampling of a single person in the same video, it is easy to cause the model to overfit). The study [5] proposed an innovative face structure deconstruction module. This module effectively suppresses the over-reliance of the model on task-independent facial features by destroying the overall structural features of the face, thereby guiding the model to focus on more discriminative local forged features. The specific implementation process is shown in Fig.2.Firstly, the input face image is divided into n × n uniform image blocks, and then these blocks are randomly replaced and reorganized. This processing method significantly destroys the global structure information of the face while completely retaining the local forgery traces. In order to ensure the correspondence between image processing and forged region labeling, the same block replacement operation is used for the labeling mask M. In the model training phase, the system randomly selects the image or the original image processed by structural deconstruction as input to enhance the robustness of the model. This innovative design achieves two key goals. On the one hand, it forces the model to give up its dependence on the overall facial features by eliminating the prior information of the face structure. On the other hand, by retaining the local area features, it ensures that the key forged trace information is not destroyed. Experiments show that this method can effectively improve the model's ability to capture subtle forgery features.

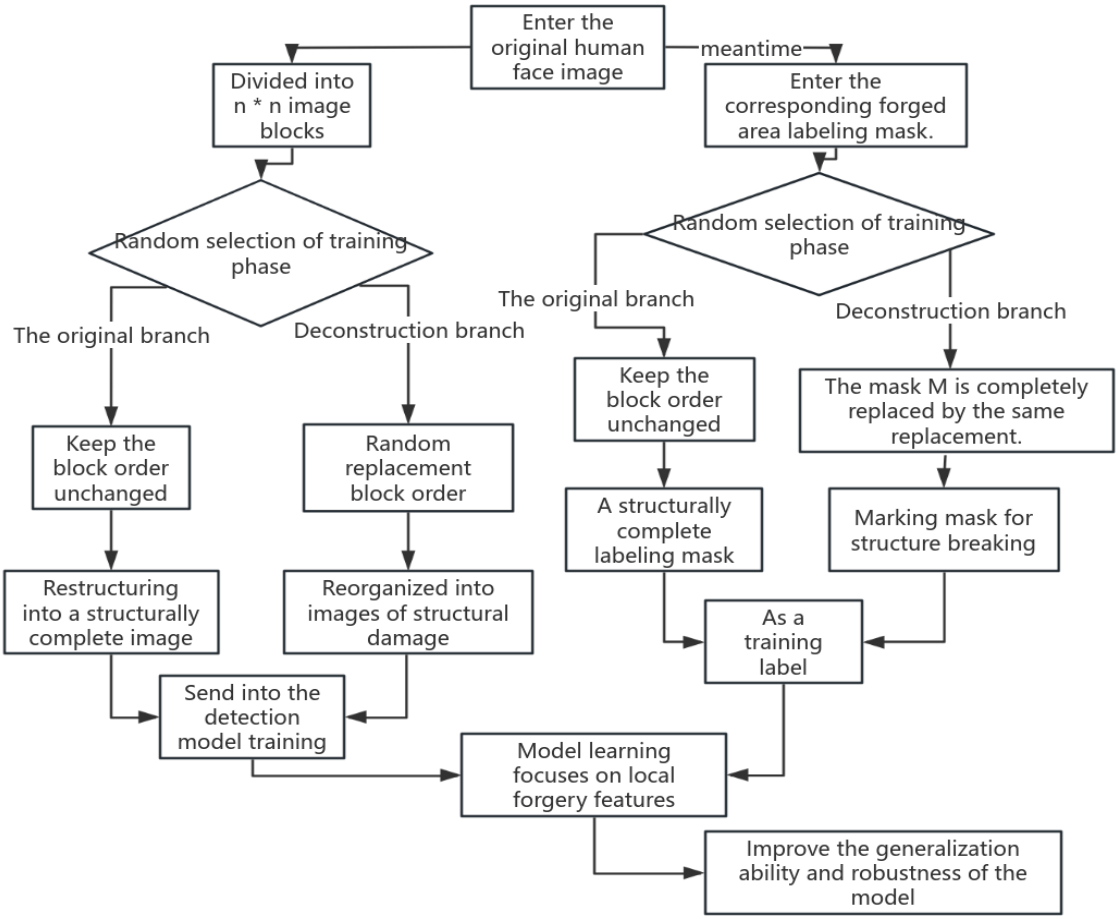

Figure 2. Face deconstruction discrimination forged feature flow [5]

3.3. Methods Based On GAN Image Features

In view of the fact that the current deep forgery video generally relies on GAN technology, researchers have developed a detection method based on deep learning, which mainly focuses on capturing the intrinsic statistical features and model fingerprints of the GAN-generated images. Yu et al. proposed a deep forgery active forensics scheme based on image steganography, which realized forgery detection by introducing artificial GAN fingerprints (AFP) [6]. As shown in Fig. 3, the core idea of this method is to embed specific fingerprint information in the original face image in advance. When these labeled images are used to train the deep forgery model, the forged image generated by the trained model will retain these fingerprint features. In the actual detection process, it is only necessary to verify whether the image to be tested contains pre-set fingerprint features to determine its authenticity, so as to initially realize the effective identification of deep forgery content.

The research shows that the image synthesized by a generative adversarial network (GAN) is significantly different from the real image in pixel-level statistical characteristics. Based on this finding, researchers have proposed a variety of detection methods using image residual analysis and frequency domain filtering. Li et al. proposed a feature extraction method based on high-pass filter residual and co-occurrence matrix, and constructed discriminant features by analyzing the filtered residual image [7]. In contrast, Nataraj et al. developed a GAN-generated false image detection method that combines the pixel co-occurrence matrix and deep learning technology. This method does not depend on the image residual calculation, but first constructs the pixel co-occurrence matrix in the three-color channels of the image, and then uses the deep convolutional neural network for model training. Experiments show that after training and testing on different data sets, the model can achieve higher detection accuracy, but the robustness to noise interference needs to be improved [8].

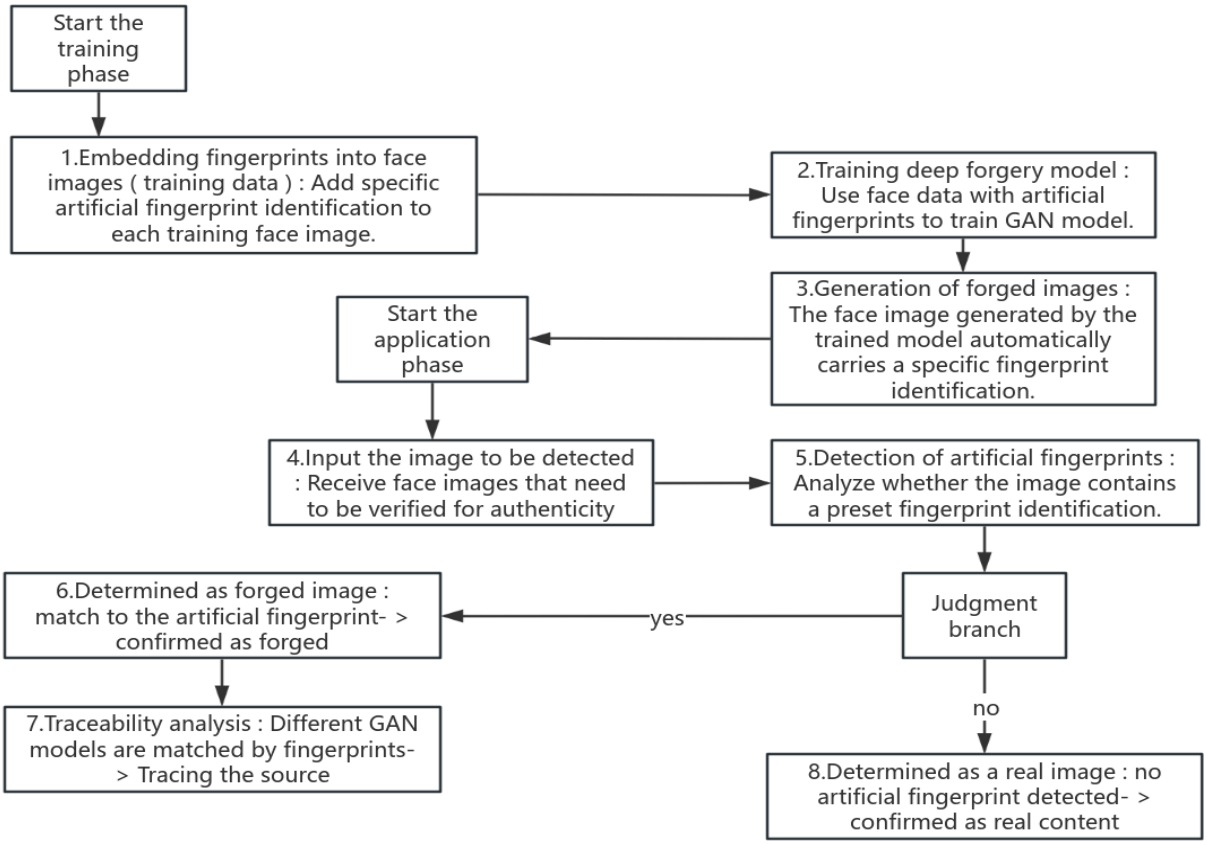

Figure 3. Using AFP to realize the forgery detection process [6]

3.4. Methods Based On Picture-level Learning

With the rapid evolution of deep forgery generation algorithms and the continuous expansion of data scale, researchers adopt an image-level analysis strategy. First, the input video is deconstructed into a single frame sequence, and then the authenticity of each frame is determined by a specially designed deep neural network. Finally, the video-level integration decision is made based on the frame-level detection results. Afchar et al. designed a small convolution module to capture microscopic tampering features [9], while Rossler et al. [10] used the Xception [11] architecture to find that even if the image is deeply compressed, the training effect for the face region is much better than the full-frame model (training is simpler and detection is more accurate). Songsri-in et al. used experiments to prove that the use of face region information can improve the detection efficiency and accuracy [12]. Nguyen et al. [13] introduced a capsule network and used VGG-19 [14] to extract facial feature coding and then classify. In terms of input, Mo et al. improved the detection effect by adding high-pass filtering and background information [15]. It is found that although the existing deep neural networks can quickly fit specific tampering features and extract highly discriminative feature representations, these features often have the problem of insufficient generalization ability and are difficult to effectively migrate to new detection scenarios. Cozzolino et al. designed a structure based on an autoencoder to enhance cross-domain generalization ability, and achieved good results with only a small amount of fine-tuning of target domain data [16]. On this basis, Nguyen et al. proposed a Y-type decoder, which combined segmentation and reconstruction loss to assist classification [17].

4. Conclusion

This paper systematically reviews the research progress of face forgery detection technology based on deep learning, including the extraction and analysis of physiological signal features, deep learning capture of image tampering traces, intrinsic feature recognition of GAN-generated images, and

image-level learning strategies. Research shows that deep learning technology significantly improves the accuracy and robustness of forgery detection through end-to-end feature learning and pattern recognition, especially when dealing with highly simulated and multi-source forgery content. However, the current research still faces many challenges. For example, the generalization ability is insufficient, which is reflected in the significant performance degradation of most detection models in cross-dataset and cross-generation method tests, and it is difficult to cope with the evolving forgery technology. There is also the vulnerability of adversarial samples. For example, forgery generation technology can avoid detection through adversarial training, and the robustness of existing models against disturbances still needs to be improved.

Future research directions can include: developing a cross-domain detection framework with more generalization ability, combining meta-learning, self-supervised learning and other technologies; a multi-modal fusion detection system is constructed to improve the detection reliability by combining context information such as audio and text. Strengthen the research of active defense technology, such as digital watermarking, GAN fingerprint embedding, etc., to curb the generation and dissemination of forged content from the source.

In a word, with the continuous game between generation technology and detection technology, face forgery detection will gradually develop from single image analysis to multimodal, interpretable and active defense, providing more solid technical support for social governance and digital security.

References

[1] Yang X, Li Y, Qi H, Lyu S. Exposing GAN-synthesized faces using landmark locations. In: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, 2019: 113 - 118.

[2] Li Y, Chang MC, Lyu S. In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018: 1 - 7.

[3] Zhu Xintong, Tang Yunqi, Geng Pengzhi. Detection algorithm of tamper and deepfake image based on feature fusion. Netinfo Security, 2021, 21 (8): 70 - 81

[4] Matern F, Riess C, Stamminger M. Exploiting visual artifacts to expose deepfakes and face manipulations. In: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). IEEE, 2019. DOI: 10.1109/WACVW.2019.00020.

[5] Yang Shaocong, Wang Jian, Sun Yunlian, et al. multi-level feature global consistency for forged face detection. Journal of Image and Graphics, 2022, 27 (9): 13.

[6] Qu Zuomin, Yin Qilin, Sheng Ziqi, et al. Review of active defense techniques for deepfake faces. Journal of Image and Graphics, 2024, 29 (2): 318 - 342.

[7] Li H, Li B, Tan S, et al. Identification of deep network generated images using disparities in color components. Signal Processing, 2020, 174: 107616.

[8] Nataraj L, Mohammed TM, Chandrasekaran S, et al. Detecting GAN generated fake images using co-occurrence matrices. arXiv preprint arXiv: 1903.06836, 2019.

[9] Afchar D, Nozick V, Yamagishi J, Echizen I. MesoNet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018: 1 - 7.

[10] Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. Face Forensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 1 - 11.

[11] Chollet F. Xception: deep learning with depth wise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1251 - 1258.

[12] Songsri-in K, Zafeiriou S. Complement face forensic detection and localization with facial landmarks. arXiv preprint arXiv: 1910.05455, 2019.

[13] Nguyen HH, Yamagishi J, Echizen I. Capsule-Forensics: using capsule networks to detect forged images and videos. In: ICASSP 2019 — 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019: 2307 - 2311.

[14] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014.

[15] Mo H, Chen B, Luo W. Fake faces identification via convolutional neural network. In: Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security, 2018: 43 - 47.

[16] Cozzolino D, Thies J, Rössler A, Riess C, Nießner M, Verdoliva L. Forensic Transfer: weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510, 2018.

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