Image-level Face Forgery Detection Methods: Revolutions, Challenges and Future Look

Zikeng Chen *

School of Computer Science and Engineering, South China University of Technology, Guangzhou, 510006, China

* Corresponding Author Email: czk910293827@outlook.com

Abstract. The rapid development of deep forgery technology poses a serious challenge to social information security. This paper systematically reviews the image-level face forgery detection methods, focusing on two technical routes: the spatial domain and the frequency domain. Spatial-domain methods achieve detection by analyzing pixel-level texture, color consistency, and boundary artifacts, supplemented by deep learning models; while frequency-domain methods use spectral features (e.g., high-frequency mesh traces) to reveal anomalies left behind by the generative model to detect the authenticity of an image. The study shows that the spatial-domain approach based on deep learning is easy to understand and has excellent detection accuracy, but lacks generalization; the frequency-domain approach is robust to compression perturbations, but is difficult to adapt to new generation techniques (e.g., diffusion models). In the future, people need to break the bottleneck of generalization, integrate multimodal features and explore lightweight design to improve the evaluation system of detection accuracy.

Keywords: Deepfake detection, Spatial domain characters, Frequency characters, Diffusion model, Generalizability.

1. Introduction

It is Breakthroughs in generative artificial intelligence (AIGC) that are particularly evident in the generation of images and have driven iterations of deepfake technology. Deepfake refers to a technology that utilizes AI techniques, particularly deep learning algorithms, to create highly realistic fake video, audio, or images. It achieves the effect of falsification by replacing human facial information, voice information, etc., in the original image. From the early low-fidelity forgery based on traditional image processing, to the high-fidelity face replacement dominated by generative adversarial network (GAN), until the seamless face synthesis generated by diffusion model, the forgery technology has gone through at least three generations of evolution. However, the evolution of deep forgery technology has also brought about a number of ethical as well as legal issues, which have a large number of negative impacts on people's well-being. Therefore, it is of great significance to investigate deep forgery detection methods to counter such threats. Researchers have proposed various image level detection methods. This review will introduce the common datasets of the mentioned methods, and categorize the methods in the article according to the different intrinsic principles - spatial and frequency domains, and systematically sort out the technical principles, performance advantages and disadvantages, and future development trends of these two major categories of methods, and finally discuss the possible future research directions in this field to provide more information for the research on image-level deep forgery detection research to provide more theoretical references.

2. Datasets

High-quality face deep forgery detection datasets are the key support to improve the model performance. The real data of such datasets usually comes from web collection or professional actors shooting, while the fake data is generated by various kinds of forgery techniques. The following is a brief overview of commonly used datasets.

Face Forensics++ is one of the most authoritative datasets in the field of face forgery detection, which consists of 1,000 original video sources with more than 5,000 forged videos generated by four different methods. The dataset provides a large amount of high-quality data with diverse forgery methods and accurate labeling. However, it still has problems such as low timeliness, low generalization, and a single scene.

Celeb-DF is a high-quality video forgery dataset aimed at solving the problems of low quality and homogeneous scenes in earlier datasets. It is mainly used for cross-dataset testing with high authenticity. However, this dataset suffers from problems such as small size and needs to be used complementarily with other datasets.

Open Forensics is an industrial-scale face forgery detection dataset. It is advanced in being huge in size and containing data generated by multiple methods, but there are problems, such as a lack of temporal ordering (it contains only static images) and a high threshold of access, which makes it less friendly to ordinary users.

DeepfakeTIMIT is a dataset whose source video is derived from the existing audio-visual dataset-VidTIMIT, which contains audio tracks for each video, but which suffers from small size, outdated generation techniques, and low quality.

DFDC is the largest publicly available deep faked dataset currently. This data level is taken in a real scene and possesses a realistic sense of perturbation and higher data complexity. However, the method of generating its data is not transparent enough.

WildDeepfake is a specialized dataset for deep forgery detection in the real world. It is highly authentic, but may suffer from problems such as privacy leakage. Table 1 provides a summary overview of the datasets covered above.

Table 1. The Summary overview of the datasets

The name of datasets	Pubulishing year	Data size	Resolution
FaceForensics++	2019	1000 original videos + 5000 fake videos	Raw HD (multiple compression qualities)
Celeb-DF	2019	590 real videos + 5639 fake videos	720p/1080p
OpenForensics	2021	1,000,000+ images (1,000+ videos)	Undecided
DeepfakeTIMIT	2018	640 videos (320 real + 320 fake)	512×512
DFD (DFDC)	2020	128,154 videos (real + fake mix)	Multiple resolutions
WildDeepfake	2020	7,314 “wild” fake videos	Multiple resolutions

3. Image-level Fake Face Detection

3.1. Spatial Domain-Based Detection Methods

This kind of methods performs authenticity discrimination by directly analyzing the color, texture, and structural features in the pixel space of an image. In addition, there are researchers who extract image features through deep and automatic learning to implement effective detection. Specifically, they can be categorized into detection methods based on traditional image forensics, fusion of boundary artifacts, and deep learning.

3.1.1. Detection Methods Based on Traditional Image Forensics

Early depth forgery left statistically countable texture and color differences in the facial region, but the early methods relied on handmade features, and the extraction of handmade features was susceptible to compression and damage. Matern et al. found that the forged face had color reflection anomalies in the eye and nose regions, which were classified by logistic regression [1]. The method is easy to understand, but some regions are not visible in the picture, leading to its poor generalization ability. Wang et al. proposed the Fused Facial Region Feature Descriptor (FFR_FD), which divides

the face into 8 sub-regions, extracts the statistics of traditional feature-point descriptors such as SIFT, and finally classifies it with a random forest [2]. This method has strong feature interpretability (the mouth region has the highest sensitivity). However, this traditional forensic method is sensitive to low-resolution images with limited feature area differentiation.

3.1.2. Detection Methods Based on Fusion Boundary Artifacts

Artifacts such as chromatic aberration and distortion are often present in the fusion boundaries of the forged face and the background, but it is difficult to locate the subtle traces in the traditional models. This feature can be investigated using artifact-based methods. Li and Lyu argued that the generated face will leave artifacts when it undergoes geometric distortions [3]. These artifact features can be modeled and ResNet-50 (residual network-50) can be used to achieve classification to discriminate the authenticity, but its accuracy is still debatable. Geng et al. proposed a two-stream Xception model [4]. Its core is to use a center-obscuring mask to allow one branch to specialize in learning the artifact features at the face boundary (this region is prone to feature conflicts due to forgery fusion), thus improving the detection ability and compression resistance, but the setting of the mask ratio has a large impact on the performance, and it needs to be controlled appropriately. The structure of its network model is shown in Figure 1.

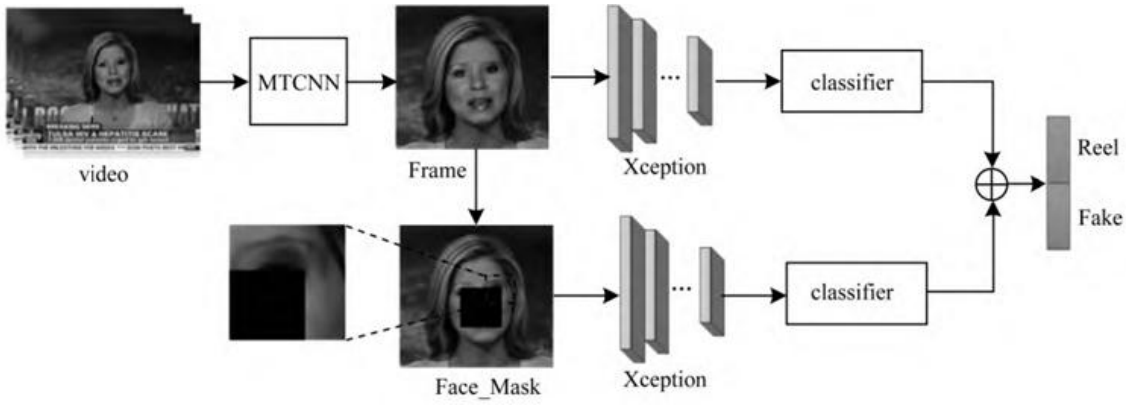

Figure 1. Structure of Xception network model [4]

3.1.3. Deep Learning Based Detection Methods

In the early detection methods, researchers tended to utilize manual methods to design algorithms for extracting the desired features. And with the continuous upgrading of forgery methods and the large time cost of manual extraction, more and more researchers tend to utilize deep learning models for autonomous learning of features. In traditional methods of forensics, single RGB spatial features are difficult to capture covert artifacts and are easily hidden, and large models have a high computational cost and poor generalization. Amin et al. proposed to enhance the saliency of forgery traces through multi-color spatial transformations (e.g., YCrCb, HSV, LAB, etc.) to solve the problem of insufficient single RGB spatial features [5]. It proposes a two-stage framework: the first stage is enhanced forgery trace enhancement; the second stage is featuring fusion, which fuses multi-space features such as HSV/YCrCb through the CvT encoder and extracts semantic information through self-attention. This framework has excellent cross-color space detection, but the real-time performance is insufficient.

Meanwhile, traditional deep forgery detection methods rely on manually designing the network architecture, which suffers from insufficient generalization capability, high computational cost and difficulty in adapting to new forgery techniques. Jin et al. proposed DFD-NAS, an end-to-end framework based on Neural Architecture Search (NAS), with the core of optimizing the network structure through automatic search [6]. It constructs a search space oriented to forgery detection in experiments and utilizes CDC convolution to capture subtle traces of forgery. Ultimately, the unit cascade pyramid network is utilized to aggregate multi-scale features for detection. This method is highly accurate in detection with less manual intervention.

Aiming at the problems of existing traditional deep forgery detection methods in terms of poor generalization across datasets and insufficient robustness across compression rates, Chen et al. proposed an efficient forgery detection method based on dual-stream ViT, whose network architecture is shown in Fig. 2 [7]. The method extracts RGB spatial features and high-frequency noise features of the image after SRM high-pass filtering through two branches to capture both low-level texture and high-level forgery traces. In order to improve the generalization and reduce the computational overhead, the authors use low-rank adaptive (LoRA) to fine-tune the parameters of ViT efficiently, and design a bidirectional adapter (BCA) and a cross-attention adapter (DCA) to achieve a deep fusion and complementarity of the dual-stream features, and finally complete the classification by a multilayer perceptual machine (MA) to identify the authenticity of the image. The method significantly reduces the training cost while ensuring the performance.

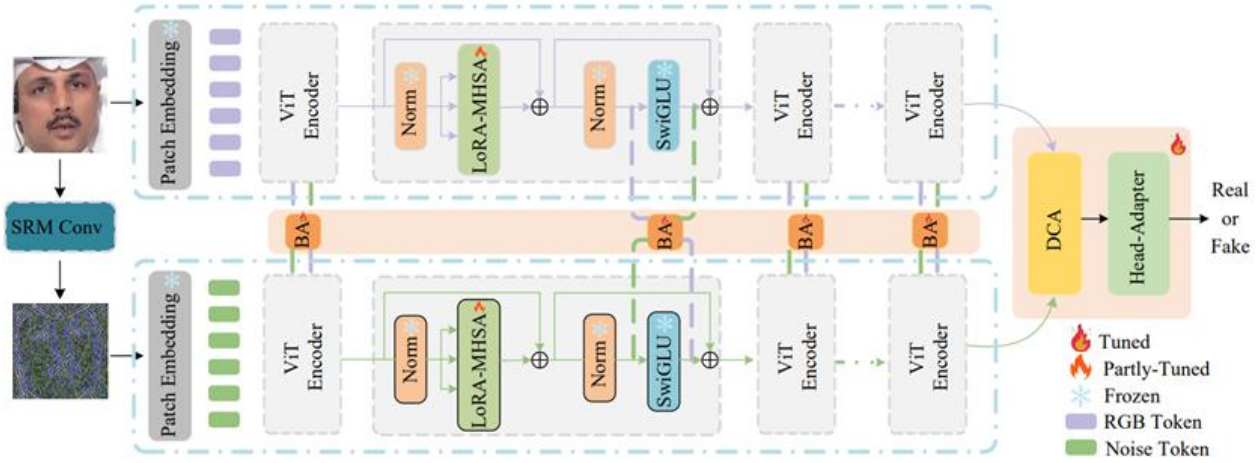

Figure 2. Structure of the ViT two-stream network model with self-supervised pre-training [7]

3.1.4. Comparison, Analysis and Discussion

Table 2 summarizes the methods reviewed in this section. The data listed are the performance of the detection methods on their respective datasets, from which it can be intuitively seen that: spatial-domain detection methods rely on pixels and shallow features early on and have limited generalization ability; boundary artifact-based methods are more general but sensitive to low-quality images and are difficult to test across datasets. The current mainstream deep learning-based methods can automatically learn features and perform well in subtle artifact recognition, but they rely on a large amount of data and computational resources and are less friendly to ordinary users.

Table 2. Summary of advantages and disadvantages of spatial domain-based detection methods

Methods	Advantages	Disadvantages	Datasets	Experimental results (%)	Cross-dataset testing	Experimental results (%)
[1]	Highly explainable	Poor generalizability	FaceForensics	AUC=86.60	-	-
[2]	Regional features can be interpreted with high precision	Regional features can be interpreted with high precision	FF++	AUC=92.30	Celeb-DF	AUC=68.80
[4]	High compression resistance	Precision degradation across compression scenarios	FF++	AUC=99.60	-	-
[3]	Efficient training	Accuracy needs to be improved	DeepfakeTIMIT	AUC=93.20	-	-
[5]	Excellent cross-space detection	Poor real-time performance and high computational complexity	DFDC	AUC=94.54	Celeb-DF	AUC=86.56
[6]	lightweighting	Lack of compatibility and loss of high-resolution detail	FF++	AUC=98.15	Celeb-DF	AUC=72.48
[7]	Dual-stream features are complementary and generalizable	Pre-training relies on big data	Celeb-DF	AUC=100.00	DF	AUC=90.00

3.2. Frequency Domain Based Detection Methods

3.2.1. Method Overview

The core idea of frequency-domain based detection methods is to capture the anomalous high-frequency signals or irregular distributions of an image by transforming it from the spatial domain to the frequency domain. Spatial-domain methods are insensitive to high-frequency artifacts (e.g., GAN mesh traces), while frequency features can enhance the compression resistance. Based on this idea, Uddin et al. designed a multiscale frequency feature extraction module (MSF3-Conv) [8]. This module can effectively capture the frequency domain artifacts left by different AI models (e.g., grid traces of GAN) by separating the high, medium and low frequency information of the image, and is robust to compression, but the generalization ability needs to be improved.

Unlike GAN-generated images, diffusion model-generated images usually do not present obvious grid-like traces in the frequency domain, so detectors trained on GAN-generated images cannot effectively detect diffusion model-generated images. Liu et al. detected AI-generated images by analyzing the response of diffusion model-generated images to Gaussian noise (noise residuals), which is a simple and effective approach but suffers from the overfitting problem [9].

Dai et al. improved the conventional steganalysis filter (SRM) into a learnable multi-scale convolutional pyramid (SRMCP) that adaptively extracts high-frequency noise features of an image for detection [10]. The method enhances robustness to compression artifacts, but its effectiveness is limited under recompression with severe loss of high-frequency information.

Existing deep forgery detection methods mostly focus on high-quality data and ignore low-quality content due to high compression, which usually results in loss of high-frequency information during compression, concealment of forgery traces, and degradation of detection performance. Gao et al. proposed a high-frequency enhancement (HiFE) framework (Figure 3), which extracts features through three branches, namely, Basic (RGB), Local (DCT-enhanced local high-frequency), and

Global (DWT enhanced global high frequency) branches to extract features and integrate the information via a cross-stage fusion module for final classification [11]. This scheme can effectively enhance the high-frequency forgery traces in compressed videos, and has the advantages of lightweight and strong compression resistance, but it lacks the ability of time-series analysis, and is insufficiently generalizable to cross-datasets.

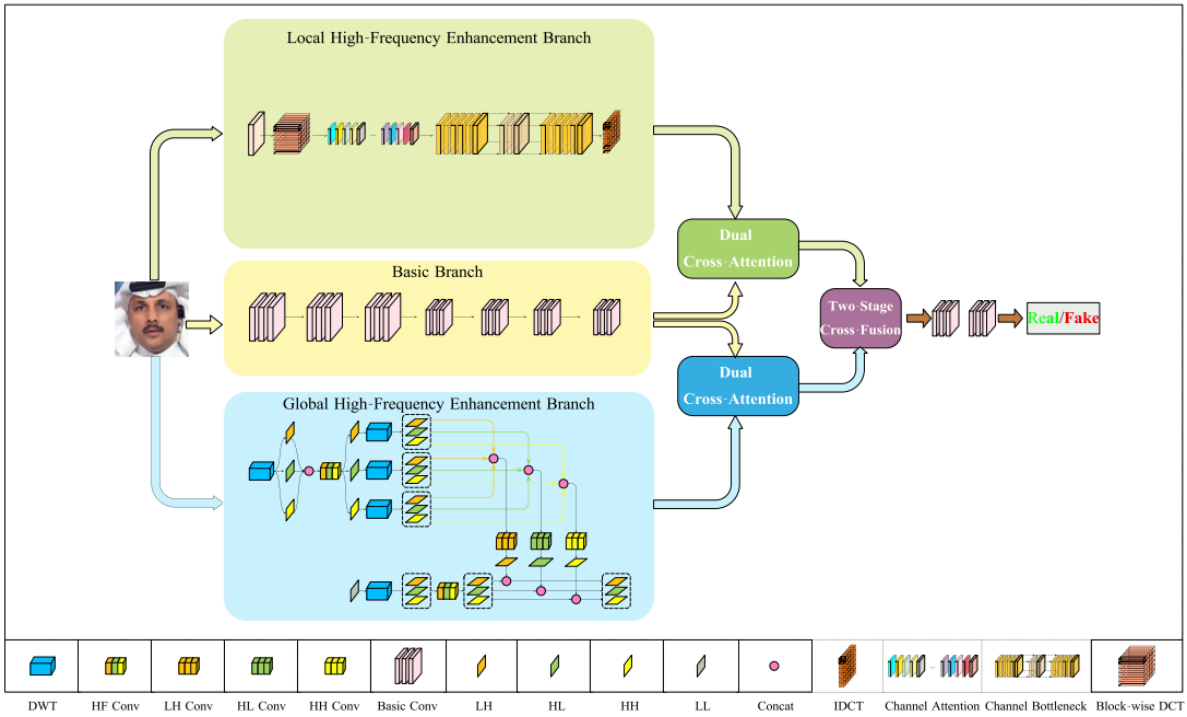

Figure 3. Three-branch conformation of the high-frequency-enhanced (HiFE) framework [11]

3.2.2. Comparison, Analysis and Discussion

Table 3 summarizes the frequency domain-based detection methods, from which it can be intuitively seen that: the core advantage of the frequency domain-based detection methods lies in the ability to reveal forged traces (e.g., GAN grids) in the spatial domain that are difficult to detect and extract, as well as the overall performance is much better, and has stronger robustness to image compression and noise interference. However, it still faces challenges such as insufficient generalization ability, high computational complexity, and difficulty in fusion with multimodal information.

Table 3. Summary of advantages and disadvantages of frequency domain-based detection methods

Methods	Advantages	Disadvantages	Datasets	Experimental results (%)	Cross-dataset testing	Experimental results (%)
[8]	High compression resistance	Poor generalization and unexplored multimodal integration	FF++	AUC=98.91	Celeb-DF	AUC=68.77
[9]	simple and effective	Overfitting exists	DF	AUC=100.00	DiffusionDB	AUC=73.21
[10]	Insufficient generalization and unexplored multimodal integration	Heavy compression (C40) performance degradation and computational overhead increase	FF++	ACC=96.26	DF	ACC=91.40
[11]	Lightweight design with excellent compression resistance	Ignore timing information	FF++	ACC=99.86	-	-

4. Future Prospects

Among the two major classes of methods reviewed in the previous section, the spatial domain methods are easy to understand and operate, but the compression and perturbation resistance are still a shortcoming; whereas the frequency domain methods are robust to perturbation, but have poor generalization to new forgery models. Based on the current research bottleneck of image-level face forgery detection, future breakthroughs can be made in the following directions: firstly, to improve lightweight and real-time performance. Design hardware-friendly model architectures to replace computationally intensive cross-domain adapters (e.g., BCA/DCA in [7]), and realize millisecond response on the mobile side to achieve real-time feedback. Equally important is the generation of model adaptive detection to improve robustness. The detection strategy can be adjusted to integrate different scale frequency features (e.g., improve the SRM filter in [9]) with adversarial training, reduce the dependence on a large amount of labeled data, and allow the model to self-learn when extracting and detecting features, improve adaptive capability, and enhance the generalization ability and resistance to emerging generative techniques. Enhanced multimodal synergy should not be neglected. For example, fusing physiological signals (e.g., heart rate fluctuations) with frequency-domain features (e.g., HiFE of [10]), modeling cross-modal correlations using graph neural networks, and deciphering high-quality forgeries. And the universal problem faced by all models is the lack of generalization of detection models, which should be worked on to improve the generalization of detection models. In future research, along with the rapid development of AIGC technology and the increasing complexity of deep forgery detection methods, it is also necessary to construct a more comprehensive evaluation system from multiple dimensions, such as data demand, model complexity, detection accuracy, and generalization ability.

5. Conclusion

Face forgery technology has objective application potential in the fields of film and television production, teaching simulation, etc., but its unlawful abuse challenges the ethical bottom line of society, and the traditional moral concepts have been greatly impacted. In order to cope with the continuous upgrading and changing of in-depth forgery methods, face forgery detection technology is also continuously developing and improving. First, this paper summarizes the commonly used rating metrics and datasets in the field of image-level face forgery detection methods. Next, compared with other reviews, this paper categorizes and summarizes the two major technical routes in this field according to the different feature extraction methods: spatial domain methods and frequency domain methods, and shows the latest research results in this field. Finally, the bottlenecks in the development of this field are elucidated, and the future research directions are envisioned.

References

[1] Matern F, Riess C, Stamminger M. Exploiting visual artifacts to expose deepfakes and face manipulations. Proceedings of 2019 IEEE Winter Applications of Computer Vision Workshops, 2019, Waikoloa, USA: IEEE: 83 - 92.

[2] Wang Gaojian, Jiang Qian, Jin Xin, Cui Xiaohui. FFR_FD: Effective and fast detection of DeepFakes via feature point defects. Information Sciences, 2022, 596: 472 - 488.

[3] Li Yuezun, Lyu Siwei. Exposing deepfake videos by detecting face warping artifacts. Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018: 46 - 52.

[4] Geng Pengzhi, Fan Hongxing, Zhang Yiyang, Tang Yunqi. Deepfake detection method based on tampering artifacts. Computer Engineering, 2021, 47 (12): 156 - 162.

[5] Amin M A, Hu Y, Guan Y, et al. Exploring varying color spaces through representative forgery learning to improve deepfake detection. Digital Signal Processing, 2024, 147: 104426.

[6] Jin X, Yu W, Chen D W, et al. DFD-NAS: General deepfake detection via efficient neural architecture search. Neurocomputing, 2025, 619: 129129.

[7] Chen Yonghao, Cai Manchun, Zhang Yiwen, et al. Face Forgery Detection Based on Parameter-Efficient Fine-Tuning and Dual-Stream Network [J]. Journal of Computer Engineering & Applications, 2025, 61 (10): 288 - 298.

[8] Uddin M, Fu Z, Zhang X, et al. Spatial and frequency feature fusion using multi-scale cross attention for enhancing deepfake face detection. Multimedia Systems, 2025, 31(4): 270 - 285.

[9] Yao Wenda, Li Panchi, Zhao Ya, Wu Hongchao. Review of research on face deepfake detection methods. Journal of Image and Graphics, 30 (7): 2343 - 2363.

[10] Dai Yunshu, Fai Jianwei, Xia Zhihua, Liu Jianan, Weng Jian. Local similarity anomaly for general face forgery detection. Journal of Image and Graphics. 2023, 28 (11): 3453 - 3470.

[11] Gao J, Xia Z, Marcialis G L, et al. DeepFake detection based on high-frequency enhancement network for highly compressed content. Expert Systems with Applications, 2024, 249: 123732.