Deepfake Face Detection from GAN to Diffusion Model

Hanyao Xu

{hayxu15@gmail.com}

College of Computer and Cyber Security, Fujian Normal University, Fuzhou, Fujian 350117, China

Abstract. As Generative Adversarial Networks (GANs) evolve towards diffusion models, deeply forged face images exhibit increasingly high similarity to real images in terms of pixel-level statistical characteristics. This causes detection methods based on low-level texture artifacts or local statistical distortions to fail in cross-model scenarios. This paper explores image-level face forgery detection from three perspectives: datasets, spatial domain methods, and frequency domain methods. Spatial domain methods have evolved from microscopic trace extraction and attention-based modeling to physical logic verification based on visual language models, while frequency domain methods have progressed from amplitude-based statistical analysis to phase spectrum consistency modeling. Analysis shows that cross-domain performance is closely related to the physical universality of the detection cues. Methods based on physical constraints exhibit more stable generalization capabilities, while those methods that heavily rely on the statistical characteristics of the training set experience significant performance degradation under distribution shifts.

Keywords: Deepfake face detection; Data domain detection; Frequency domain detection.

1 Introduction

Deepfake detection is one of the core research topics in the field of digital forensics. Since around 2018, forgery methods based on Generative Adversarial Networks (GANs), such as DeepFakes, Face2Face, FaceSwap, and NeuralTextures, have emerged. GAN-generated images suffer from two inherent defects: first, periodic high-frequency artifacts (checkerboard artifacts) introduced by the transposed convolution operation; and second, inconsistent noise statistical characteristics across different image regions. This provided early detection methods with usable criteria for discrimination.

Since 2022, generative techniques based on the Denoising Diffusion Probability Model (DDPM) have gradually become widespread in the field of image synthesis. The diffusion model reconstructs the data distribution through an iterative denoising Markov chain process, reducing the statistical distance between the generated image and the real image in terms of texture transitions, lighting, and anatomical details. The two defects of GANs mentioned above are avoided in the diffusion model. When faced with content generated by the diffusion model, the area under the receiver operating characteristic (AUC) of the detection model

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365359

trained on FaceForensics++ (FF++) generally drops to around 50%, indicating that the discriminative features of existing methods have become largely ineffective under the new generation paradigm [1]. The aforementioned performance failure can be explained from two levels. The first level is the data level. Existing benchmarks (FF++, Celeb-DF) are built on GAN-generated content and cannot reflect the statistical characteristics of content generated by the diffusion model. The discriminative features established by detection methods on such data lack coverage of the new generation paradigm. The second level is the method level. The discriminative criteria of existing spatial domain and frequency domain methods are coupled with the specific implementation of the generation model: spatial domain methods rely on pixel statistical bias, and frequency domain methods rely on amplitude statistical fingerprints. Both lose their effective discriminative criteria in the face of content generated by the diffusion model. Based on the above analysis, this paper reviews the data from three perspectives: dataset construction, spatial domain methods, and frequency domain methods. The dataset section addresses the time lag between benchmark construction and generation technology iteration; the spatial domain section tracks the evolution of detection criteria from pixel statistical deviation to physical consistency verification; and the frequency domain section analyzes the technical path of phase spectrum analysis as an alternative discrimination dimension after the failure of amplitude statistical fingerprint. By comparing the mainstream methods of each path, this paper analyzes their applicable boundaries in cross-domain detection scenarios.

2 Face deepfake detection dataset

2.1 GAN-based datasets

FF++ was released by Rossler et al. in 2019. It uses 1000 real videos from YouTube as raw materials and generates corresponding tampered versions through four types of GAN forgery algorithms: DeepFakes, Face2Face, FaceSwap, and NeuralTextures, for a total of 5000 video sequences [1]. The dataset provides three compression levels: original quality (RAW), high quality (HQ), and low quality (LQ), and supports evaluation of the robustness of the algorithm under different compression scenarios. After its release, it became the most widely used training and evaluation benchmark in this field. Celeb-DF, released by Li et al. in 2020, contains 590 real celebrity videos and 5639 face-swapped videos processed by improved GAN algorithms [2]. Its design goal is to improve the quality of fake samples in terms of the naturalness of face-swapping boundaries and skin color consistency, to fill the gap of FF++ on high-difficulty samples, and is used as a benchmark for evaluating the generalization performance of the method across datasets. The common limitation of the two datasets is that their fake content comes from GAN-type methods, which is fundamentally different from the content generated by the diffusion model in terms of statistical characteristics. This is the direct cause of the data-level failure mentioned in Chapter 1.

2.2 Diffusion mode-based dataset

In response to the above limitations, Diffusion Facial Forgery (DiFF), released by Cheng et al. in 2024, is the first large-scale face forgery detection benchmark specifically built for diffusion models [3]. This dataset integrates 13 mainstream diffusion model frameworks,

including Stable Diffusion, and generates more than 500,000 images through large-scale text prompts, and labels the generation model category and the corresponding prompt semantics. Compared to FF++ and Celeb-DF, DiFF directly covers the content generated by the diffusion model, and its multimodal association annotation method supports both cross-model detection research and forensic experiments combining semantic information. Table 1 summarizes the key attributes of the dataset.

Table 1. Deepfake Detection Datasets.

Dataset	Release Year	Forgery Type	Size	Main Features
FF++ [1]	2019	GAN (4 types)	5000 videos	Multiple compression levels, video-level benchmark
Celeb-DF [2]	2020	GAN face swap	5639 videos	High-quality face swap, high-difficulty evaluation
DiFF [3]	2024	Diffusion model (13 types)	500,000 images	Multimodal annotation, image-level benchmark

From the overall evolution of the datasets, there is a continuous time lag between benchmark construction and generation technology. FF++ and Celeb-DF provide a unified evaluation protocol for detection research in the GAN era, but neither can cover the statistical characteristics of the content generated by the diffusion model. The emergence of DiFF fills this gap, but with the continuous evolution of generation models based on new architectures such as Diffusion Transformer (DiT), the construction of datasets still faces the pressure of continuous updates.

3 Spatial domain-based deepfake face detection

3.1 Detection based on microscopic traces and fusion boundaries

Spatial domain methods directly analyze the pixel representation of images, which is the earliest established research path in deepfake detection. Early methods relied on low-level distortions introduced at the pixel level by image editing operations. MesoNet, proposed by Afchar et al. in 2018, captures the mid-level feature deviations remaining during video compression by constructing a shallow convolutional neural network [4]. The effectiveness of this method depends on the inherent spectral anomalies introduced by GAN upsampling operations. These anomalies are still partially retained after lossy compression of the image, thus maintaining a certain detection capability in LQ video scenarios. However, its discriminative features are closely tied to the implementation of specific generation algorithms.

Face X-ray, proposed by Li et al. in 2020, shifts the detection basis from the spectral fingerprint of specific algorithms to the general physical traces of image stitching behavior [5]. When two images from different sources are stitched together, no matter how the stitching algorithm processes the boundary region, the inherent differences in color balance, luminance distribution and noise statistics between the two images cannot be completely eliminated. This difference will form a learnable discriminative signal at the stitching boundary. Face X-ray enables the network to explicitly model the general representation of the fusion boundary

through supervised learning, and can locate the tampered area in the form of grayscale image. It also maintains relatively stable detection performance when facing GAN generation algorithms that are not present in the training set. Compared with MesoNet, Face X-ray has improved cross-algorithm generalization ability, but the discriminative premise of this method - that is, the existence of stitching boundary in the image - no longer holds in the full image synthesis scenario of diffusion model.

3.2 Detection based on attention mechanisms and data augmentation

As the quality of GAN generated images continues to improve, pixel-level fusion boundary artifacts tend to be invisible. Detection research has turned to introducing attention mechanism to achieve focus on key local regions. Multi-attentional (MAT) was proposed by Zhao et al. in 2021 [6]. Its design logic is that the traces of facial forgery are concentrated in biologically sensitive locations such as the eyes, bridge of the nose, and corners of the mouth, rather than being evenly distributed across all regions of the image. MAT uses multiple parallel attention branches to enable the network to extract features independently from these local regions while extracting global features. The outputs of different branches are weighted and fused for final discrimination. Multi-scale Multi-modal Transformer (M2TR) was proposed by Wang et al. in 2022. It introduces the self-attention mechanism of Transformer into forgery detection, extracts multi-scale features in parallel at different spatial resolutions, and combines multi-modal input information for joint discrimination [7]. However, the discrimination features of the above two methods are still rooted in the pixel statistical distribution of the training data. Experimental data show that MAT's AUC drops to 49.0% and M2TR's AUC drops to 56.3% in the CelebA-HQ cross-domain detection task, both close to the level of random guessing, indicating that the discrimination criteria based on pixel statistics have structural limitations in the face of content generated by the diffusion model. Self-Blended Images (SBI) was proposed by Shiohara et al. in 2022. Its approach is different from the above methods [8]. SBI does not rely on statistical artifacts of specific GAN models. Instead, it constructs forged samples during the training phase by performing complex blending operations on real images themselves. This allows the network to learn to recognize the image merging behavior itself, rather than memorizing a specific type of forgery algorithm. This strategy enables SBI to achieve an AUC of 93.6% and an equal error rate (EER) of 14.0% in CelebA-HQ cross-domain detection, making it the most stable cross-domain performance among similar spatial domain methods. Its advantage lies in the decoupling of the training objective from the specific generation algorithm. However, when the forged image is completely synthesized from scratch by the generation model rather than through a merging operation, the discriminative premise of SBI no longer holds, and its applicable scope has inherent limitations.

3.3 Physical logic verification based on visual language models

The performance failure of the above two types of methods in the face of content generated by diffusion models stems fundamentally from the limitations of the discriminative criteria—low-level pixel statistics and mid-level local textures cannot effectively distinguish between images generated by diffusion models and real images. In this context, using large-scale pre-trained visual language models (VLMs) for semantic-level physical logic verification has become a new research direction. The basic premise is that the generative model's modeling

ability at the pixel level is close to that of real images, but its training process does not introduce explicit constraints on the laws of the physical world, thus resulting in detectable intrinsic defects in the physical consistency of image content.

Pirogov's research systematically evaluated the feasibility of using VLM for zero-shot depth forgery detection [9]. This method uses VLM to make physical inferences about image content from three dimensions: geometric consistency, illumination transmission logic, and biometric authenticity. In terms of geometric consistency, the authenticity of the image is determined by checking whether the reflection direction of the pupil highlight in the image is consistent with the geometric vector of the main light source in the environment; in terms of illumination transmission logic, potential forgeries are identified by checking whether the projection direction of the shadows of facial features conforms to the linear propagation law of point light sources; in terms of biometric authenticity, the judgment is made by checking whether the skin subsurface scattering presents an unnatural uniform distribution in the generated image, and whether there is a physical contradiction in the environmental mapping in the corneal reflection. The basis for the above three dimensions of inspection comes from the objective laws of the physical world and is unrelated to the specific implementation of the generative model. In addition, at the output modeling level, this method addresses the limitation of traditional visual question answering (VQA) systems that only output discrete word units by proving that a normalization reconstruction mechanism based on the probability of the Logits layer is proposed. This mechanism transforms the original log probability of the VLM for the corresponding word units of "real" and "fake" into continuous detection confidence. This continuous confidence score enables the detection system to adapt to mainstream evaluation metrics such as receiver operating characteristic (ROC) and EER, and supports flexible trade-offs between false alarm rate (FAR) and false negative rate (FRR) by adjusting the classification threshold. Experimental data show that zero-shot detection based on InstructBLIP achieves an AUC of 81.3% in the CelebA-HQ cross-domain task, which is improved to 92.1% after targeted fine-tuning, and the EER is reduced to 12.2%, both of which are better than all traditional spatial domain methods except SBI[9].

4 Frequency domain-based deep face depth spoofing detection

Frequency domain detection methods map images from pixel space to the transform domain and observe statistical anomalies left by the generative model in the frequency space, providing a complementary discriminative dimension for spatial domain detection. The basic basis is that generative models such as GANs often fail to strictly follow the sensing noise, imaging blur and physical texture formation mechanisms in the real optical imaging link during upsampling, interpolation and texture reconstruction, thus leaving analyzable abnormal patterns in the spectral distribution. Compared with spatial domain features, these anomalies can often still be retained in statistical form under degradation conditions such as compression and smoothing, thus making frequency domain methods more stable in low-quality media forensics.

4.1 Dual-Stream frequency-aware detection based on amplitude statistics

Frequency-aware Face Forgery Network (F^3-Net) is the first work to systematically use frequency-aware cues for deep forgery detection [10]. Its design goal is to solve the problem of significant decline in detection accuracy in LQ compressed video scenarios. The core architecture of F^3-Net consists of two complementary branches: frequency-aware decomposition (FAD) and local frequency statistics (LFS). The FAD branch performs multi-band decomposition of the image spectrum using a pre-defined frequency filter, mapping the spectrum to three independent energy bands: low-frequency, mid-frequency, and high-frequency. Each energy band is then characterized independently, enabling the model to distinguish the distribution of forgery traces across different frequency ranges. The LFS branch uses a sliding window mechanism to statistically analyze the distribution patterns of local frequency responses, providing detailed modeling of local spectral fluctuations at different locations in the image, thus compensating for the FAD branch's shortcomings in perceiving local details through global analysis. The output features of the two branches are aligned and fused across dimensions using matrix multiplication in the MixBlock module. Ablation experiments show that the combination of FAD and LFS branches has a significant complementary effect; removing either branch alone leads to a significant performance degradation. F^3-Net establishes the feasibility of using frequency domain features as a supplementary dimension for forensics, but its discrimination criteria are still limited to the statistical distribution of the amplitude spectrum, facing similar limitations in application to spatial domain methods when generating content for diffusion models.

4.2 Detection based on phase spectrum consistency

As the quality of generated models continues to improve, frequency domain detection methods that solely rely on statistical anomalies in the amplitude spectrum face greater limitations in cross-model scenarios. To address this issue, research has begun to extend the detection perspective from amplitude statistics to phase information in the complex frequency domain, in order to uncover anomalous features in the frequency structure of generated images.

The Spatial-Phase Shallow Learning (SPSL) study proposed by Liu et al. shows that face forgery images are usually accompanied by cumulative upsampling operations during the generation process, which introduces observable changes in the frequency domain, especially in the phase spectrum. Compared with the amplitude spectrum, the phase spectrum can retain richer frequency composition information, providing supplementary evidence for forgery trace analysis [11]. Based on this, SPSL enhances the model's ability to represent internal generation traces in forgery images by jointly modeling spatial domain image and phase spectrum features, and improves transfer performance in cross-dataset detection tasks. This indicates that frequency domain detection should not be limited to amplitude energy distribution, but should also pay more attention to the consistency and integrity of the image's frequency structure. For real optical imaging images, their frequency structure is usually constrained by scene geometry, texture boundaries and imaging links, thus having a more stable organization; while fake images are more likely to produce unnatural mismatches in local frequency structure during generation, which also provides theoretical support for phase spectrum analysis to detect deep face fakes [11].

5 Conclusion

This paper systematically reviews the research on deepfake face detection over the past five years from three perspectives: dataset construction, spatial domain methods, and frequency domain methods. At the dataset level, benchmark construction is evolving from FF++ and Celeb-DF, which cover GAN forgery types, to DiFF, which is specifically designed for diffusion model construction. However, the time lag between benchmark construction and the iteration of generation techniques remains a structural problem restricting research progress. At the spatial domain level, method evolution has gone through three stages: from pixel-based statistical microscopic trace recognition such as MesoNet and Face X-ray, to local feature aggregation incorporating attention mechanisms such as MAT and M2TR, and then to physical logic verification based on VLM. The performance ceiling of each stage significantly improves with the increase in the level of discrimination criteria. At the frequency domain level, the amplitude statistical fingerprint detection framework established by F^3-Net has a clear advantage in compression robustness, while the phase spectrum consistency modeling introduced by F2C extends frequency domain analysis to the diffusion model era.

In future research, phase spectrum consistency and VLM semantic verification have independent discriminative capabilities for diffuse-generated images at the signal and content layers, respectively. How to design an effective cross-modal fusion mechanism to combine these two types of heterogeneous evidence remains an unexplored issue in current research. With the continuous evolution of novel architectures such as DiT, the coverage and update frequency of existing benchmarks cannot meet the needs of detection research. Constructing a large-scale evaluation benchmark with broader coverage and more refined annotation is of urgent practical significance. The robustness of existing methods under adversarial processing such as strong compression and geometric transformations needs systematic evaluation. Exploring the defense mechanisms of detection systems in proactive adversarial scenarios is a necessary step in translating research conclusions into practical applications. Furthermore, the ability to generate interpretable forensic reports is of great value in applications such as legal evidence collection. VLM methods have a natural advantage in this direction, but how to ensure the accuracy and verifiability of the report content still requires further research.

References

[1] Rossler, A., Cozzolino, D., Verdoliva, L. et al.: Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1-11 (2019)

[2] Li, Y., Yang, X., Sun, P. et al.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3207-3216 (2020)

[3] Cheng, H., Guo, Y., Wang, T. et al.: Diffusion facial forgery detection. In Proceedings of the 32nd ACM International Conference on Multimedia. pp. 5939-5948 (2024)

[4] Afchar, D., Nozick, V., Yamagishi, J. et al.: Mesonet: A compact facial video forgery detection network. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1-7 (2018)

[5] Li, L., Bao, J., Zhang, T. et al.: Face x-ray for more general face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5001-5010 (2020)
[6] Zhao, H., Zhou, W., Chen, D. et al.: Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2185-2194 (2021)
[7] Wang, J., Wu, Z., Ouyang, W. et al.: M2tr: Multi-modal multi-scale transformers for deepfake detection. In Proceedings of the 2022 International Conference on Multimedia Retrieval. pp. 615-623 (2022)
[8] Shiohara, K., Yamasaki, T.: Detecting deepfakes with self-blended images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18720-18729 (2022)
[9] Pirogov, V.: Visual language models as zero-shot deepfake detectors. arXiv preprint arXiv:2507.22469 (2025)
[10] Qian, Y., Yin, G., Sheng, L. et al.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In European Conference on Computer Vision. pp. 86-103. Springer International Publishing, Cham (2020)
[11] Liu, H., Li, X., Zhou, W. et al.: Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 772-781 (2021)