A Comprehensive Evaluation of Deepfake Detection Methods: Approaches, Challenges and Future Prospects

Xixi Hu*

Software Engineering, Yunnan University, 650504 Kunming, China

Abstract. Advances in technology have made deepfake forgeries easier, posing serious ethical and security risks that highlight the urgent need for better detection methods. This paper provides a comprehensive discussion of various Deepfake detection approaches, including methods based on physical attributes and visual inconsistencies, data-driven techniques (such as spatial and frequency domain detection methods), and those using generative models. Based on the classification and introduction of representative methods, the paper further compares their performance across different datasets, revealing that while current methods can detect deepfake to some extent, they generally suffer from poor generalization and accuracy when dealing with different types of forgeries or low-quality data. In conclusion, this study offers insights into the development of future deepfake detection technologies, emphasizing the importance of combining multiple approaches and improving model generalization to address increasingly complex forgery scenarios. It can serve as a valuable reference for researchers looking to understand the advancements in this field.

1 Introduction

In today's digital era, images serve as a crucial medium for information transmission across various sectors of society. However, advances in digital image processing have made manipulation and falsification easier and more widespread than ever before. Deepfake technology, which combines deep learning with content generation techniques like autoencoders, generates highly realistic fake images and videos by altering facial features and expressions [1]. Initially popular in the entertainment industry, its use has since spread to areas like politics, media, and personal privacy [1]. This expansion has led to significant concerns such as fake news, public opinion manipulation, and pornography, all of which pose considerable challenges to social trust and information authenticity.

As deepfake technology evolves, traditional detection methods struggle to counter its highly precise fabrications, necessitating more advanced detection solutions. Such solutions are essential not only for safeguarding the authenticity of information and maintaining social trust but also play an indispensable role in combating cyber fraud, protecting privacy, and safeguarding intellectual property. Consequently, research on deepfake detection holds

* Corresponding author: neutrino2306@mail.ynu.edu.cn

significant academic value by pushing the boundaries of machine learning and computer vision, while also carrying vital societal implications by protecting information authenticity.

In pursuit of better performance, deepfake detection methods have advanced in tandem with advancements in deep learning, transitioning from simple neural network-based techniques to more sophisticated ones. Initially, foundational architectures such as ResNet50 [2], VGG-19 [3], and DenseNet121 [4] were used, relying on operations like convolution and pooling within the network to extract deeper-level features from images. However, as the field advanced, these initial methods gave way to more nuanced detection strategies, which further enhance the accuracy of detecting deepfake images through deeper feature extraction and refined detection techniques. Despite their effectiveness, these methods tend to drop significantly in performance when encountering unfamiliar deepfake types, as they are not designed to generalize well to previously unseen forgery techniques.

To address these limitations, researchers have explored various advanced approaches that go beyond basic feature extraction, targeting specific artifacts and inconsistencies in facial identity, expression, and biological signals present in manipulated content. For instance, Chen et al. [5] designed a bi-granularity artifact detector aimed at identifying intrinsic artifacts caused by model generation and extrinsic artifacts resulting from blending operations. Similarly, Dong et al. [6] proposed an identity-unaware model that seeks to remove the influence of facial identity information, relying solely on artifact cues for authenticity identification. Additionally, several methods have focused on detecting inconsistencies in manipulated content based on mid-level manipulation traces (Gao et al., 2023 [7]; Liu et al., 2023 [8]; Yu et al., 2023 [9]).

Overall, this paper classifies mainstream deepfake detection methods and introduces selected representative approaches, followed by a performance comparison across various datasets to provide a comprehensive evaluation of their effectiveness. The remainder of this paper is structured as follows: Section 2 introduces the fundamental concepts and terminologies related to deepfake technology and its detection. Section 3 discusses the technical aspects of deepfake detection, presenting a detailed classification and review of selected methods. Building on this, Section 4 provides a detailed comparison of various methods, offering a comprehensive and balanced evaluation of their performance. Finally, Section 5 concludes the paper by summarizing the key findings and briefly discussing future directions for research and development.

2 Preliminaries related to Deepfake detection

2.1 Problem definition

Deepfake detection can be regarded as an image-level or pixel-level classification problem, aiming to identify forged content in images or videos. Specifically, the detection process can be represented as:

$$S_o = \varphi_{\boldsymbol{D}(I_o)} \tag{1}$$

Where φ_D represents the abstract detection network, and S_o denotes the fake score for the generated content I_o.

Different forgery techniques possess unique generation characteristics that directly impact the authenticity of the forged content. These characteristics not only determine the style and quality of the generated content but also pose specific challenges for detection methods. The following are several primary forgery techniques along with their definitions:

Face Swap: Face swapping is a technique that replaces the face from a source image onto a target image to create a forged image. This technique first extracts facial landmarks from

the source, and then utilizes them to match a 3D template model, subsequently copying the texture from the source image to the facial features of the target character [10].

Face2Face: This method captures facial movements and expressions from a video and applies them to the face in another video, enabling real-time facial expression transfer. It employs 3D models to track facial expressions, allowing for the dynamics of the source face to be copied while preserving the features of the target face [11].

Neural Textures: Neural Textures introduce a facial reenactment technique that modifies only the mouth region of the target video. It encodes the texture information of the target into a renderable neural representation, optimized using adversarial loss and photometric reconstruction loss to ensure the realism and consistency of the generated images [12].

Deepfakes: Deepfakes primarily employs Generative Adversarial Networks (GANs) to produce highly realistic forged images and videos. It first extracts facial features from the source image using autoencoders and reconstructs the face. These features are then replaced onto the target face, followed by post-processing to synthesize realistic images further [10].

2.2 Datasets

The selection of datasets plays a pivotal role in deepfake detection research, as each dataset may capture different forgery techniques or variations in data quality. The following Table 1 offers a concise overview of the most commonly used datasets, including data volume, forgery techniques, and their respective advantages and disadvantages. Each dataset has its unique strengths and limitations, providing a valuable reference for the training and evaluation of forgery detection models.

Table 1. Detailed information of different datasets.

Dataset Name	Real/Forged Data Volume	Real Data Source	Forgery Techniq ue	Advantages	Disadvanta ges	Year of Publicati on
FaceForensics ++ (FF++) [13]	1,000/5,000	YouTu be	Deepfak es, Face2Fa ce, FaceSwa p, Neural Textures	Multiple forgery techniques; includes multiple resolutions and compressio n levels	Overall low dataset quality; obvious forgery artifacts	2019
Celeb-DF [14]	590/5,639	YouTu be	Deepfak es	High resolution; natural video colors; high-quality dataset	Limited forgery methods; small dataset volume	2019
DFDC [15]	23,654/104,500	Real-world videos	DFAE, FaceSwa p, FSGAN, StyleGA N, etc.	Entirely real-world scenes; large dataset volume; close to reality	Overall low dataset quality; obvious forgery artifacts	2020

DeeperForensics-1.0 [16]	50,000/10,000	Real-world videos	Deepfakes	Large data volume; high diversity and comprehensive consideration of real-world scenarios	Single forgery generation algorithm using an encoder method	2020
FFIW [17]	12,000/10,000	YouTube	FSGAN, FaceSwap	Designed for multi-scene scenarios; high-quality forgery	Uses face-swapping as the forgery generation method	2021
KoDF [18]	62,166/175,776	YouTube	FaceSwap, FSGAN, etc.	Large dataset volume; balanced representation of Asians in the dataset	Inconsistent dataset quality	2021
Vox-DeepFake[19]	1,125,429/1,045,786	VoxCeleb	Deepfakes, FSGAN, FaceShifter	Large dataset volume; each person in the videos has corresponding reference information and real videos	Uses face manipulation as the forgery technique, specifically focusing on identity forgery detection research	2020
WildDeepfake [20]	3,805/3,509	Internet	Not publicly disclosed	Includes multiple real-world scenes; highly realistic	Small dataset volume	2019
FFPMS [21]	0/14,000 (frames)	FF++	Deepfakes, Face2Face, FaceSwap, etc.	Provides both frame-level and video-level labels	Small dataset volume; overall low forgery quality	2020

2.3 Evaluation metrics

The performance of deepfake detection models is typically evaluated using metrics common to traditional binary classification tasks, where the objective is to classify images as either 'real' or 'forged.' By comparing the labels output by the detection model with the actual

categories of the images, four result categories can be derived: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

Based on these foundational categories, several performance ratios can be computed:

True Positive Rate (TPR) or Recall: The proportion of correctly identified forged images out of all actual forged images.

$$TPR = Recall = \frac{TP}{TP + FN} \tag{2}$$

False Positive Rate (FPR): The proportion of real images incorrectly classified as forged out of all real images.

$$FPR = \frac{FP}{FP + TN} \tag{3}$$

True Negative Rate (TNR) or Specificity: The proportion of correctly identified real images out of all real images.

$$TNR = Specificity = \frac{TN}{TN + FP} \tag{4}$$

False Negative Rate (FNR): The proportion of forged images incorrectly classified as real out of all forged images.

$$FNR = \frac{FN}{TP + FN} \tag{5}$$

In addition to the four primary outcomes and ratios listed above, several important derived evaluation metrics can be used to assess the model's performance:

Accuracy: The proportion of correctly classified images (both real and forged).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Precision: The proportion of predicted forged images that are truly forged.

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

F1 Score: The harmonic mean of precision and recall, balancing false positives and negatives

$$F1\ Score = \frac{2 \cdot Precision \times Recall}{Precision + Recall} \tag{8}$$

Area Under the Curve (AUC-ROC): AUC measures the area under the ROC curve, which plots the false positive rate against the true positive rate (recall) across different decision thresholds.

Equal Error Rate (EER): The point where false acceptance rate (FAR) equals false rejection rate (FRR). A lower EER reflects better performance, indicating fewer errors when balancing false acceptances and rejections.

3 Deepfake detection: A review of methods and techniques

This chapter presents an overview of mainstream deepfake detection methods, including those based on physical attributes and visual inconsistencies, data-driven methods, and generation-based detection techniques. Through a systematic classification and analysis, it lays a foundation for advancing future deepfake detection research and improving detection capabilities.

3.1 Detection methods based on physical attributes and visual inconsistencies

Facial X-ray: One effective approach to detecting deepfake images is to focus on the inconsistencies and physical anomalies that arise during image manipulation. Among these methods, Facial X-ray is considered a classic approach [22]. It detects the blending boundary

between the foreground and background in manipulated images. Instead of directly analyzing specific forgery artifacts, it focuses on the inconsistencies introduced at the boundary during image composition. The technique enhances the detection of the blending region by applying a soft mask with Gaussian blur, while also using color correction to match the foreground and background colors, thereby generating more realistic forged images for learning and detection.

Self-blended images (SBIs): Building on this idea of boundary inconsistencies, Shiohara et al. proposed the SBIs method, which encourages classifiers to learn feature representations by using more generalized deepfake images [23]. Unlike Face X-ray, SBIs applies various operations to the original image, generating manipulated "source images" with distinct features and unaltered "target images." Then, a grayscale mask is generated to define the blending region between the two images, creating boundary inconsistencies that mimic typical artifacts found in forged images. By leveraging these more generalized manipulations, the method enhances robustness and performs better when detecting novel deepfake techniques.

3.2 Data-driven detection methods

3.2.1 Spatial domain detection

Convolutional Neural Network (CNN): Convolutional Neural Networks (CNNs) play a crucial role in deepfake detection [1], with many existing methods being based on CNNs or their variants, such as ResNet, VGGNet, XceptionNet [24], and EfficientNet [25]. These deep learning models are capable of extracting deep-level features from images, which helps identify subtle anomalies in manipulated images. XceptionNet and EfficientNet, in particular, are frequently utilized as foundational models or backbone architectures in many deepfake detection approaches due to their strong feature extraction capabilities.

Capsule Network: Compared to traditional CNNs, capsule networks are better at capturing the relationships between objects, making them particularly suitable for detecting deepfake images. As GAN-generated images increasingly reduce visible artifacts and distinctive patterns, detecting these forgeries becomes more challenging. To address the limitations of neural networks in handling adversarial examples, Xue et al. designed a model combining Capsule Networks (CapsNet) with Generative Adversarial Networks (GANs) [26], where the generator creates forged images, and the Capsule Network serves as the discriminator, classifying the input data and distinguishing between real and fake samples. By jointly training the generated forged samples alongside real samples, the model enhances its classification performance, especially improving its generalization capability in scenarios where labeled samples are limited.

Following this line of research, Nguyen et al. proposed the Capsule-Forensics method, which also leverages capsule networks to detect deepfake images [27]. The method first uses a face detection algorithm to crop the facial region and then applies VGG-19 for image preprocessing. After that, the extracted image features are fed into the capsule network, which is used to detect deepfake images. This approach requires fewer parameters compared to traditional CNNs, thereby reducing computational costs.

HiFi-Net: Rather than relying on specific features like capsule networks or frequency domain techniques, HiFi-Net adopts a feature-based strategy for deepfake detection [28]. It extracts image features using a multi-branch structure, where different resolution branches are responsible for forgery detection at varying levels of granularity. The highest-resolution branch employs a self-attention mechanism to localize forgery regions, while metric learning is used to distinguish between real and manipulated pixels. Then, partial convolution is applied to focus on features in the forged areas, generating a binary mask for classification.

Finally, hierarchical classification refines the detection from coarse-grained forgery identification to specific forgery method recognition, enhancing the model's generalization and accuracy.

3.2.2 Frequency domain detection

F3-Net: F3-Net is a method for detecting facial forgeries by leveraging frequency domain features [29]. It utilizes both the Discrete Cosine Transform (DCT) and a Frequency-aware Decomposition (FAD) module to extract multi-band components from the image, allowing it to identify subtle artifacts introduced during the forgery process. Additionally, the Local Frequency Statistics (LFS) module is employed to capture frequency statistics from local spatial regions, highlighting the differences between authentic and forged images in the frequency domain. F3-Net integrates the outputs from the FAD and LFS modules using a cross-attention mechanism, enhancing its detection performance, particularly when dealing with low-quality or heavily compressed forged images.

HiEF: The High-Frequency Enhancement Network (HIEF) aims to improve the performance of Deepfake detection on highly compressed images [30]. It begins by extracting high-frequency features using block-based Discrete Cosine Transform (DCT) at the local level, with channel attention mechanisms and bottleneck modules to adaptively enhance these features, followed by inverse DCT to reconstruct the image. At the global level, it applies multi-level Discrete Wavelet Transform (DWT) to capture multi-scale high-frequency information, utilizing cascaded residuals for feature fusion. Finally, a dual-stage cross-fusion module integrates both local and global high-frequency information to boost detection accuracy, particularly for low-quality and highly compressed Deepfake images.

3.2.3 Other data-driven methods

Attention Mechanism: Attention mechanisms have been increasingly applied in deepfake detection due to their ability to focus on manipulated regions and improve detection accuracy. By localizing forgery artifacts and integrating deep features, they significantly enhance the effectiveness of detection models.

In line with this trend, Zhao et al. presented a multi-attention mechanism for deepfake detection [31]. First, a convolutional neural network (CNN) is used to extract spatial and texture features from the image. Subsequently, these features are enhanced by a Bilinear Attention Pooling (BAP) module. Through this process, the spatial attention mechanism focuses on identifying specific regions in the image that may have been manipulated, while the texture feature enhancement module targets finer details, particularly texture anomalies that may arise during the forgery process. Finally, the integrated features are passed through a classifier to determine whether the image is fake.

SOLA Module: Fei et al. developed the SOLA method, which enhances the generalization ability of face forgery detection by identifying local anomalies [32]. The method decomposes local features of the image and computes first-order and second-order local anomaly maps to capture forgery traces. Additionally, it introduces the Local Enhancement Module (LEM) and Adaptive Spatial Refinement Model (ASRM) to extract subtle forgery features. SOLA operates without requiring pixel-level annotations and demonstrates excellent performance across multiple datasets, particularly in handling unseen forgeries, where it exhibits strong generalization capabilities.

3.3 Generation-based detection methods

As deepfake technology continues to advance, the images generated have become increasingly realistic, making it more challenging for detection methods to distinguish between authentic and fake images. However, GAN-generated images often leave behind subtle, unique artifacts during the generation process. Detection methods based on GAN fingerprints leverage these high-level features to enhance detection capabilities.

For instance, Guarnera et al. observed that the local pixel correlations in forged faces are influenced by the operations performed across all layers of a GAN, particularly in the transposed convolution layers [33]. They proposed a mathematical model using the Expectation-Maximization algorithm to capture these pixel correlations and extract a set of local feature vectors specifically designed to model the convolutional generation process.

Similarly, Wang et al. proposed an effective pipeline based on a generative network, referred to as the Deepfake Destroyer, which trains a perturbation generator to safeguard source images from deepfake manipulation algorithms while ensuring that the detector's perspective remains distinct from that of the human eye [34]. Jeong et al. introduced a novel framework consisting of a fingerprint generator and a GAN-generated image detector [35]. The detector reduces dependence on real image data by using mixed samples of real and generated images, enhancing the model's robustness against high-quality GAN-generated images.

In another approach, Junyi Cao et al. developed RECCE, a framework for deepfake detection that integrates reconstruction learning, multi-scale graph reasoning, and reconstruction-guided attention [36]. The reconstruction network is trained solely on real facial images, allowing it to capture anomalies in fake images. The multi-scale graph reasoning module aggregates feature discrepancies across different scales, improving the detection of forgery traces. Moreover, the reconstruction-guided attention mechanism directs the classification network to focus more accurately on manipulated regions, enhancing overall detection accuracy.

4 The comparison of performance based on various Deepfake methods

This chapter presents a performance evaluation of the majority of deepfake detection methods discussed earlier, utilizing commonly used datasets with data sourced directly from the original papers. It includes an examination of each method's performance in both in-domain and cross-domain testing, offering a comprehensive assessment of their generalization capabilities and robustness in handling unseen forgery scenarios.

The following sections will present the experimental results across in-domain, cross-domain, and various forgery types. The performance of each method is analyzed using metrics such as ACC, AUC, and EER, providing an in-depth comparison that serves as a reference for future research and improvements.

Table 2. Performance Comparison of Deepfake Detection Methods on Various Datasets (In-Domain Training)

Methods	FF++ (c23, HQ)		FF++ (c40, LQ)		Celeb-DF		WildDeepfake	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Xception	95.73 %	96.30 %	86.86 %	89.30 %	97.90 %	99.73 %	77.25 %	86.76 %

Face X-ray	-	87.40 %	-	-	-	-	-	-
SBIs	-	99.64 %	-	-	-	93.74 %	-	-
Capsule-Forensics	93.11 %	-	-	-	-	-	-	-
F3-Net	98.95 %	99.30 %	93.02 %	95.80 %	95.95 %	98.93 %	80.66 %	87.53 %
HiEF	-	92.83 %	-	71.84 %	-	-	-	-
MultiAtt	97.60 %	99.29 %	88.69 %	90.40 %	97.92 %	99.94 %	82.86 %	90.71 %
Guarnera	-	-	-	-	90.20 %	-	-	-
SOLA	-	99.25 %	-	-	-	-	-	-
EfficientNet-B4	-	95.59 %	-	-	-	96.30 %	-	-
RECCE	97.06 %	99.32 %	91.03 %	95.02 %	98.59 %	99.04 %	83.25 %	92.02 %

Table 2 presents the results of in-domain testing, where each method was trained and evaluated on the same dataset. On the high-quality FF++ dataset, most methods exhibited excellent detection capabilities. SBIs stood out with the highest performance, while F3-Net, MultiAtt, and RECCE also showed strong detection results. Although Xception performed well, it slightly lagged behind the leading methods. EfficientNet-B4 demonstrated stable but slightly lower performance, and Capsule-Forensics showed reasonable accuracy despite missing AUC data.

When tested on the low-quality FF++ dataset, where video compression posed a challenge, methods like F3-Net, RECCE, and MultiAtt continued to perform robustly, handling the lower-quality data effectively. However, HiEF struggled the most under these conditions, reflecting its limited ability to cope with compressed videos.

The Celeb-DF dataset, with its more realistic forgeries, further tested the methods' capabilities. RECCE and F3-Net continued to excel, with Xception achieving the best performance. SBIs also showed good results, and Guarnera demonstrated strong accuracy, even though AUC data was not provided.

On the WildDeepfake dataset, the differences between methods became clearer under more realistic conditions. RECCE and MultiAtt led in performance, while F3-Net and Xception followed closely behind. SBIs, however, showed weaker generalizability, struggling with the more complex forgeries in this dataset.

Table 3. Performance Evaluation of Deepfake Detection Methods on Cross-Dataset Testing (Out-of-Domain Testing)-1

Methods	Train	FF++ (c23，HQ)			FF++ (c40，LQ)			Celeb-DF		
		ACC	AUC	EER	ACC	AUC	EER	ACC	AUC	EER
Xception	FF++	95.73%	96.30%	-	86.86%	89.30%	-	-	-	-

Face X-ray	FF++&BI	-	87.40%	-	-	61.60%	-	-	80.58%	-
SBIs	FF++	-	99.64%	-	-	-	-	-	93.18%	-
F3-Net	FF++	97.52%	98.10%	-	90.43%	93.30%	-	-	65.17%	42.03%
HiEF	FF++	-	92.83%	-	-	71.84%	-	-	95.69%	-
MultiAtt	FF++	97.60%	99.29%	-	88.69%	90.40%	-	-	67.02%	37.90%
SOLA	FF++	-	-	-	-	-	-	-	76.02%	-
RECCE	FF++	97.06%	99.32%	-	91.03%	95.02%	-	-	68.71%	35.73%

Table 4. Performance Evaluation of Deepfake Detection Methods on Cross-Dataset Testing (Out-of-Domain Testing)-2

WildDeepfake			DFD			DFDC		
ACC	AUC	EER	ACC	AUC	EER	ACC	AUC	EER
-	-	-	-	-	-	-	-	-
-	-	-	-	95.40%	8.37%	-	80.92%	27.54%
-	-	-	-	97.56%		-	72.42%	-
-	57.10%	45.12%	-	-	-	-	64.60%	39.84%
-	-	-	-	-	-	-	-	-
-	59.74%	43.73%	-	-	-	-	68.01%	37.17%
-	-	-	-	-	-	-	-	-
-	64.31%	40.53%	-	-	-	-	69.06%	36.08%

Table 3 and Table 4 present the cross-domain testing results, showing varying performance across methods on unseen datasets. On the Celeb-DF dataset, HiEF and SBIs demonstrate strong detection capabilities, while Face X-ray shows moderate results. In contrast, F3-Net and MultiAtt perform weaker, with both methods exhibiting higher error rates, indicating limitations on this dataset.

On the WildDeepfake dataset, overall performance drops significantly, with RECCE leading, but still showing a high error rate. MultiAtt and F3-Net also struggle, reflecting the difficulty in generalizing to more realistic forgery scenarios.

In the DFD dataset, SBIs excels with the best detection performance, followed by Face X-ray, which also adapts well. However, the DFDC dataset reveals a marked performance gap across methods, with Face X-ray performing better than most, while SBIs and RECCE show moderate capabilities but higher false detection rates. Overall, these results indicate

that most methods face challenges in handling domain shifts, particularly on more complex datasets like WildDeepfake and DFDC.

Table 5. Fine-grained Testing Results of Various Deepfake Detection Methods

Methods	Train	DF	F2F	FS	NT
		AUC	AUC	AUC	AUC
Xception	DF	99.38%	75.05%	49.13%	80.39%
Face X-ray		99.17%	94.14%	75.34%	93.85%
MultiAtt		99.51%	66.41%	67.33%	66.01%
SOLA		100%	96.95%	69.72%	98.48%
RECCE		99.65%	70.66%	74.29%	67.34%
Xception	F2F	87.56%	99.53%	65.23%	65.90%
Face X-ray		98.52%	99.06%	72.69%	91.49%
MultiAtt		73.04%	93.06%	55.35%	66.66%
SOLA		99.73%	99.56%	93.50%	96.02%
RECCE		75.99%	98.06%	64.53%	73.32%
Xception	FS	70.12%	61.70%	99.36%	68.71%
Face X-ray		93.77%	92.29%	99.20%	86.63%
MultiAtt		75.90%	54.64%	98.37%	49.72%
SOLA		99.11%	98.13%	99.98%	92.07%
RECCE		82.39%	64.44%	98.82%	56.70%
Xception	NT	93.09%	84.82%	47.98%	99.50%
Face X-ray		99.14%	98.43%	70.56%	98.83%
MultiAtt		79.09%	74.21%	53.99%	88.54%
SOLA		99.64%	97.69%	90.20%	99.76%
RECCE		78.83%	80.89%	63.70%	93.63%

Fine-grained testing is designed to evaluate a model's generalization capabilities by training it on one specific forgery category and testing it on multiple other forgery types. This approach helps reveal how well a model performs when encountering similar but slightly different forgeries, providing valuable insights into its robustness and adaptability. In the FF++ dataset, there are four main forgery types: Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), and Neural Texture (NT). During each test, methods are trained in one category and tested on all four, providing a comprehensive assessment of cross-category performance.

Table 5 presents the performance of five deepfake detection methods (Xception, Face X-ray, MultiAtt, SOLA, RECCE) trained on different datasets and evaluated on the same test set. It is evident that the choice of training dataset significantly impacts the detection performance of each method across different types of forgeries. Overall, the SOLA method consistently delivers stable and effective performance across all forgery types, regardless of the training dataset, demonstrating its strong generalization ability. In comparison, Face X-ray also shows considerable stability, particularly when detecting NT and F2F forgeries, as it achieves high AUC values across different training sets. On the other hand, the performance of Xception, MultiAtt, and RECCE appears more dependent on the training set, especially

when detecting FS and NT, where their AUC and accuracy may significantly drop with certain training datasets, indicating weaker generalization. Thus, SOLA and Face X-ray exhibit more consistent performance under varying conditions, while the effectiveness of the other methods is more contingent on the specific combination of training and forgery types.

5 Conclusion

This paper begins by reviewing the rapid advancements in deepfake generation technologies and the resulting challenges to image authenticity, highlighting the necessity for effective detection methods. It then systematically summarizes various detection approaches, including those based on physical attributes and visual inconsistencies, data-driven techniques, and generative model detection. Each approach has its strengths and weaknesses; for example, methods based on physical and visual features are often effective in detecting specific forgery artifacts, while data-driven techniques leveraging deep learning models demonstrate better feature extraction capabilities. However, when tested on different datasets, particularly those with low compression quality or highly realistic fake content, significant disparities in detection accuracy and generalization performance among different models emerge. This review provides valuable insights for the future development of deepfake detection technologies. Future work could focus on integrating existing methods while continuing to explore solutions that enhance model generalization and address increasingly complex forgery scenarios.

References

1. L. A. Passos, D. Jodas, K. A. Costa, L. A. Souza Júnior, D. Rodrigues, J. Del Ser, & J. P. Papa. A review of deep learning-based approaches for deepfake content detection. Expert Systems, 41(8), e13570 (2024).
2. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, (2016).
3. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, (2014).
4. G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700-4708, (2017).
5. H. Chen, Y. Li, D. Lin, B. Li, & J. Wu, Watching the big artifacts: Exposing deepfake videos via bi-granularity artifacts. Pattern Recognition, 135, 109179 (2023).
6. S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, & Z. Ge, Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3994-4004) (2023).
7. J. Gao, S. Concas, G. Orrù, X. Feng, G. L. Marcialis, & F. Roli, Generalized deepfake detection algorithm based on inconsistency between inner and outer faces. In International Conference on Image Analysis and Processing (pp. 343-355). Cham: Springer Nature Switzerland (2023).
8. B. Liu, B. Liu, M. Ding, T. Zhu, & X. Yu, TI2Net: temporal identity inconsistency network for deepfake detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 4691-4700) (2023).

9. Y. Yu, X. Zhao, R. Ni, S. Yang, Y. Zhao, & A. C. Kot, Augmented multi-scale spatiotemporal inconsistency magnifier for generalized deepfake detection. IEEE Transactions on Multimedia, 25, 8487-8498 (2023).
10. L. Zhang, T. Lu, & Y. Du. A survey of deepfake detection methods for facial videos. Journal of Computer Science and Exploration (in Chinese), 17(1), 1. (2023).
11. J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, & M. Nießner, Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2387-2395) (2016).
12. J. Thies, M. Zollhöfer, & M. Nießner, Deferred neural rendering: Image synthesis using neural textures. Acm Transactions on Graphics (TOG), 38(4), 1-12. (2019).
13. A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, & M. Nießner. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 1-11, (2019).
14. Y. Li, X. Yang, P. Sun, H. Qi, & S. Lyu, 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).
15. B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, & C. C. Ferrer, The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397 (2020).
16. L. Jiang, R. Li, W. Wu, C. Qian, & C. C. Loy, Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2889-2898) (2020).
17. T. Zhou, W. Wang, Z. Liang, & J. Shen, Face forensics in the wild. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5778-5788) (2021).
18. P. Kwon, J. You, G. Nam, S. Park, & G. Chae, Kodf: A large-scale korean deepfake detection dataset. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10744-10753) (2021).
19. X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, D. Chen, ... & B. Guo. Identity-driven deepfake detection. arXiv preprint arXiv:2012.03930 (2020).
20. B. Zi, M. Chang, J. Chen, X. Ma, & Y. G. Jiang, Wilddeepfake: A challenging real-world dataset for deepfake detection. In Proceedings of the 28th ACM international conference on multimedia (pp. 2382-2390) (2020).
21. X. Li, Y. Lang, Y. Chen, X. Mao, Y. He, S. Wang, ... & Q. Lu, Sharp multiple instance learning for deepfake video detection. In Proceedings of the 28th ACM international conference on multimedia (pp. 1864-1872) (2020).
22. L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, & B. Guo, 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).
23. K. Shiohara, & T. Yamasaki. Detecting deepfakes with self-blended images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18720-18729) (2022).
24. A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, & M. Nießner. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1-11) (2019).
25. Q. Liu, Z. Xue, H. Liu, & J. Liu. Enhancing deepfake detection with diversified self-blending images and residuals. IEEE Access (2024).

26. Z. Xue. A general generative adversarial capsule network for hyperspectral image spectral-spatial classification. Remote Sensing Letters, 11(1), 19-28 (2020).
27. H. H. Nguyen, J. Yamagishi, & I. Echizen. Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467 (2019).
28. X. Guo, X. Liu, Z. Ren, S. Grosz, I. Masi, & X. Liu. Hierarchical fine-grained image forgery detection and localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3155-3165) (2023).
29. Y. Qian, G. Yin, L. Sheng, Z. Chen, & J. Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In European Conference on Computer Vision (pp. 86-103). Cham: Springer International Publishing (2020).
30. J. Gao, Z. Xia, G. L. Marcialis, C. Dang, J. Dai, & X. Feng. DeepFake detection based on high-frequency enhancement network for highly compressed content. Expert Systems with Applications, 249, 123732 (2024).
31. H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, & N. Yu. Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2185-2194) (2021).
32. J. Fei, Y. Dai, P. Yu, T. Shen, Z. Xia, & J. Weng. Learning second order local anomaly for general face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20270-20280) (2022).
33. L. Guarnera, O. Giudice, & S. Battiato. Deepfake detection by analyzing convolutional traces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 666-667) (2020).
34. X. Wang, J. Huang, S. Ma, S. Nepal, & C. Xu. Deepfake disrupter: The detector of deepfake is my friend. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14920-14929) (2022).
35. Y. Jeong, D. Kim, Y. Ro, P. Kim, & J. Choi. Fingerprintnet: Synthesized fingerprints for generated image detection. In European Conference on Computer Vision (pp. 76-94). Cham: Springer Nature Switzerland (2022, October).
36. J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, & X. Yang. End-to-end reconstruction-classification learning for face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4113-4122) (2022).