Journal of King Saud University Computer and Information Sciences (2025) 37:148
https://doi.org/10.1007/s44443-025-00154-8

REVIEW PAPER

Unmasking AI-created visual content: a review of generated images and deepfake detection technologies

Yupeng Zhang[1] · Zongwei Pang[1] · Shanyuan Huang[1] · Chengyou Wang[1,2,3,4] · Xiao Zhou[1,2,3,4]

Received: 19 February 2025 / Accepted: 1 July 2025 / Published online: 31 July 2025

Abstract

In this era, digital images and videos are ubiquitous in people's lives, and generative models can easily produce high-quality images and videos. These images and videos enrich people's lives and play important roles in various fields. However, maliciously generated images and videos can mislead the public, manipulate public opinion, invade privacy, and even lead to illegal activities. Therefore, detecting AI-created visual content has become a significant research topic in the field of multimedia information security. In recent years, the rapid development of deep learning technology has greatly accelerated the progress of AI-created visual content detection. This survey introduces the detection technologies for AI-created visual content that have developed in recent years, divided into two parts: AI-generated image detection and deepfake detection. In the AI-generated image detection section, we introduce current generative models and basic detection frameworks, and overview existing detection methods from the perspectives of unimodal and multimodal. In the deepfake detection section, we provide an overview of existing deepfake generation technique classifications, commonly used datasets, followed by some common evaluation metrics within the field. We also analyze the technical characteristics of existing methods based on the different feature information they utilize, summarizing and categorizing them. Finally, we propose future research directions and conclusions, offering suggestions for the development of AI-created visual content detection technologies.

Keywords AI-created visual content · AI-generated image detection · Deepfake detection · Deep learning

1 Introduction

In recent years, the rapid development of technology has significantly improved AI-created visual content techniques in terms of visual quality, semantic complexity, and operational efficiency. People can easily obtain high-quality images and videos by simply clicking a mouse or entering a text description. However, this unprecedented technology has also raised concerns about the spread of false information. Therefore, developing effective tools for AI-created visual content detection has become increasingly important. In recent years, the rapid development of technology has significantly improved AI-created visual content techniques in terms of visual quality, semantic complexity, and operational efficiency. People can easily obtain high-quality images and videos by simply clicking a mouse or entering a text description. However, this unprecedented technology has also raised concerns about the spread of false information. Therefore, developing effective tools for AI-created visual content detection has become increasingly important. In this survey, we categorize AI-created visual content into AI-generated images and deepfake. This classification is

Yupeng Zhang and Zongwei Pang contributed equally to this work.

✉ Chengyou Wang
wangchengyou@sdu.edu.cn

Yupeng Zhang
zhangyp_sdu@mail.sdu.edu.cn

Zongwei Pang
pangzw@mail.sdu.edu.cn

Shanyuan Huang
huangshanyuan@mail.sdu.edu.cn

Xiao Zhou
zhouxiao@sdu.edu.cn

1 School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264209, China

2 Shandong Key Laboratory of Intelligent Communication and Sensing-Computing Integration, Shandong University, Jinan 250061, China

3 Shandong Key Laboratory of Intelligent Electronic Packaging Testing and Application, Weihai 264209, China

4 Shandong University–Weihai Research Institute of Industrial Technology, Weihai 264209, China

primarily based on the difference in data formats (deepfakes include both videos and images). Additionally, AI-generated images are more diverse in content, encompassing various entities, whereas deepfake technology is distinctive in that it primarily focuses on modifying or generating human faces, involving issues of identity forgery and facial authenticity. Therefore, it needs to be studied as a separate topic. Based on this judgment, we will separately discuss detection techniques for AI-generated images and deepfake detection techniques based on deep learning. AI-created visual content detection techniques are categorized based on different methods, as shown in Fig. 1.

The rapid development of generative technologies has enabled current generative models to not only produce images with high efficiency but also achieve exceptionally high-quality outputs. While these technologies enrich various aspects of human life, they have also precipitated multiple societal crises, including the spread of misinformation (Kertysova 2018), socio-ethical challenges (Ho 2023), and security vulnerabilities (Golda et al. 2024). AI-generated image detection serves as a critical tool for safeguarding information ecosystems, protecting individual rights, and ensuring social stability, playing a pivotal role in ensuring the ethical and secure application of these technologies. In response, researchers have begun systematically exploring methods for detecting AI-generated images, adhering to rigorous academic standards. The current generated images methods can be primarily divided into two categories: image classification tasks and image attribution tasks. Image classification tasks treat generated images as a binary classification problem, where the detector learns to differentiate between real and fake images by identifying distinct features, thus outputting different labels to detect fake images. Image attribution tasks, on the other hand, leverage unique fingerprints and other characteristics specific to different generative models, matching them with the input image features to identify the generating model of fake images. Some studies also explore which aspects are more beneficial for detecting AI-generated images. Currently, most generated images methods are based on image classification tasks. Starting with the simplest classifiers, the field has progressed to using deep neural networks (DNNs), convolutional neural networks (CNNs), and other neural networks for generated images by incorporating spatial, frequency, texture, and other features. Later advancements involved cross-domain feature fusion and the use of image-text methods. Generated images technology has rapidly developed in recent years. Even though high-quality generated images may be indistinguishable from real ones to the human eye, statistical characteristics of the images still exhibit differences from real images. These differences enable detectors to distinguish between real and fake images. With the development of deep learning techniques, computers are now capable of learning these differences and performing effective detection.

In AI-created visual content technologies, deepfake technology allows the amazingly accurate modification of faces, sounds, or whole scenarios. This includes modifying facial expressions, swapping faces, adjusting lip-syncing in films, and more Yadav and Vishwakarma (2024). In 2017, a Reddit user named "Deepfake" used deep learning techniques to create and spread a pornographic video of Gal Gadot, marking the beginning of this technology's tremendous rise. In 2022, a video of Ukrainian President Zelensky urging soldiers to surrender went viral, with over 250,000 viewers. In

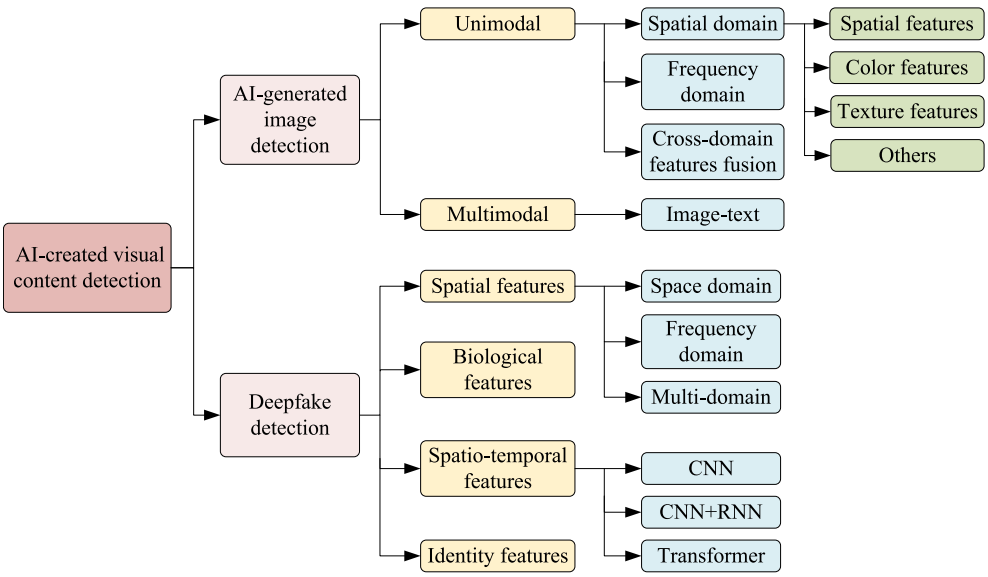

Fig. 1 The basic categories of AI-created visual content detection

2024, South Korea experienced a surge in deepfake-related sexual crimes, with potentially up to 220,000 victims, including many adolescent students and even minors. According to a security report by QAX in 2024, AI-based deepfake frauds surged 30 times in 2023, and AI-driven phishing emails increased 10 times. While deepfake technology does have some positive applications, its abuse has posed significant threats to national security, social media, and public trust. To address these challenges, researchers have focused on deepfake detection tasks, improving robustness and generalizability, and developing advanced methods. Among these, deep learning-based approaches have shown clear superiority in detection performance, so this survey primarily focuses on deep learning-based methods.

In the past few years, several surveys on AI-generated image detection technologies have been published. Hu and Wang (2020) outlined the mainstream frameworks of neural networks, briefly introduced the applications of deep learning in generative image and natural image forensics, and finally pointed out the challenges and future prospects of deep learning in this field. Deng et al. (2023) studied research on defending against AI-generated visual media attacks. They summarized existing attack methods and defense strategies, and within a unified passive and active framework, reviewed mainstream defense-related tasks, evaluating their robustness and fairness. Additionally, they summarized commonly used evaluation datasets, standards, and metrics, but noted that there is limited research on AI-generated image detection methods. Guo et al. (2023) categorized AI-generated images into active forensics and passive forensics, discussing the superiority of active forensics over passive forensics. However, most generative models do not embed watermarks in generated images, making this method highly limited. Lin et al. (2024a) conducted an extensive survey on AI-generated content detection, but the AI-generated image detection methods they reviewed were all from 2023, without analyzing or reviewing earlier methods. This survey briefly introduces generative models used for image generation, provides a comprehensive review of AI-generated image detection methods, and compares their advantages and disadvantages.

Several recent reviews have systematically summarized the existing deepfake detection techniques. Rana et al. (2022) analyzed 112 relevant papers and categorized their methods into four types: deep learning-based techniques, classical machine learning-based methods, statistical techniques, and blockchain-based techniques. However, this paper does not delve into future trends. Seow et al. (2022) provided a detailed introduction to deepfake generation, including the types of deepfakes and some available forgery tools. They reviewed existing deepfake detection work from two perspectives: conventional methods and deep learning-based methods. Gong and Li (2024) grouped the surveyed methods into four categories: conventional CNN-based detection, CNN backbone with semi-supervised detection, transformer-based detection, and biological signal detection, according to their feature extraction methods and network architectures. Heidari et al. (2024) focused on deep learning-based detection methods, providing a detailed study of four applications: video detection, image detection, audio detection, and hybrid multimedia detection. They also highlighted several unresolved issues that require further attention. Sandotra and Arora (2024) focused on the generation of deepfakes, covering topics such as face manipulation methods, open-source tools, and so on. It classified forgery detection methods from the perspectives of space, time, and frequency features. Kaur et al. (2024) provided a detailed classification of detection methods while discussing some challenges in the field, which are summarized into three categories: data challenges, training challenges, and reliability challenges. They also highlighted some of the main differences between deepfake image detection and video detection. Finally, it offered an outlook on future opportunities. The above reviews have conducted an in-depth analysis of past work, but none of them summarize detection methods from the perspective of the features used. Therefore, this review will start with feature selection and provide a discussion and analysis of existing deepfake detection algorithms.

The remainder of this survey is organized as follows: Section 2 provides the foundational knowledge of AI-generated image detection and deepfake detection, including datasets, basic detection frameworks, evaluation metrics, and more. Section 3 presents the existing methods in AI-generated image detection. Section 4 discusses the existing deepfake detection methods, with a focus on the differences in feature selection approaches. Section 5 offers future research directions of AI-created visual content detection and conclusions.

2 Foundational knowledge on AI-created visual content detection

This section introduces common generative models, forgery methods, basic frameworks for AI-generated image detection and deepfake techniques, along with commonly used datasets and evaluation metrics.

2.1 Technical foundations for AI-generated image detection

This section will be elaborated in the following three parts: generative models for images, basic framework of AI-generated image detection and datasets for generated images.

2.1.1 Generative models for images

The rapid development of technology has led to the emergence of many generative models capable of producing high-quality images, such as generative adversarial networks (GANs) and diffusion models (DMs). We provide a brief introduction to GANs and DMs used for image generation.

GANs for images In 2014, Goodfellow et al. (2014) introduced generative adversarial network, which became a very effective method for image generation. It consists of two components: the discriminator and the generator. The goal is for the discriminator and generator to compete against each other, where the generator tries to produce images that can deceive the discriminator, and the discriminator aims to identify images generated by the generator. The overall framework is shown in Fig. 2. The loss function L for the model can be written as (1):

$$\min_G \max L = \mathbb{E}_{x-p_{data}}\left[\log D(z)\right] + \mathbb{E}_{x-p_x}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \tag{1}$$

where the probabilities $D(z)$ and $G(z)$ represent the outputs of the discriminator and the generator, respectively. The discriminator aims to minimize the loss, while the generator seeks to maximize it. Due to the superiority of GANs in image generation, numerous variants of GANs have emerged in recent years, particularly for image-related applications. In 2017, Chollet (2017) introduced deepfake, which replaces the Inception module with depthwise separable convolutions, achieving better results with the same number of parameters. Following that, Zhu et al. (2017) introduced CycleGAN, an unsupervised learning-based GAN that enables style transfer without requiring extensive data preparation. In 2018, Bellemare et al. (2018) designed a unique loss function and introduced CramerGAN, a model capable of generating high-quality images. Karras et al. (2018) introduced ProGAN, which learns to generate high-resolution images progressively, starting from low-resolution ones. To address the single-domain transfer issue seen in models like CycleGAN, Choi et al. (2018) introduced StarGAN, a model capable of performing multi-domain transfer with a single network.

In 2019, Brock et al. (2019) incorporated the idea of orthogonal regularization into GANs, leading to the introduction of BigGAN, which significantly improves the generative performance of GANs through timely truncation of the input prior distribution. He et al. (2019b) introduced AttGAN by incorporating an attribute classification constraint to enable more precise manipulation of image attributes. Wu et al. (2019) proposed RelGAN, a GAN based on relative attributes, which allows for the modification of images by continuously altering specific attributes of interest. Park et al. (2019) introduced GauGAN, a network capable of generating images from textual descriptions, marking a new era in image generation. Karras et al. (2019) re-examined the limitations of ProGAN and, drawing inspiration from style transfer, proposed StyleGAN. Due to occasional artifacts in images generated by StyleGAN, Karras et al. (2020) improved upon it, resulting in StyleGAN2, which generates higher-quality images. In 2021, Lee et al. (2021) enhanced GANs by incorporating contrastive learning and mutual information maximization techniques, presenting InfoMaxGAN.

DMs for images In 2020, Ho et al. (2020) introduced the diffusion model, which involves both forward and reverse propagation processes. The quality of images generated by this model surpassed that of GANs, offering advantages such as greater diversity and more stable training. This marked the beginning of the growing popularity of diffusion models. In 2021, Ramesh et al. (2021) proposed DALL·E, a model capable of generating surrealistic images directly from textual descriptions. Dhariwal and Nichol (2021) introduced ADM, which incorporated classifier guidance to improve the quality of generated images. In 2022, Nichol et al. (2022) introduced Glide, which employed classifier-guided generation by training an additional classifier to continually refine the generated image at each timestep, ultimately generating images belonging to specific categories. Saharia et al. (2022) proposed Imagen, which featured two super-resolution diffusion models for generating high-resolution images. Rombach et al. (2022) introduced stable diffusion (SD), which transformed the diffusion process into a low-dimensional latent space, addressing the issue of large sampling spaces. Subsequently, a series of variants of SD were released. Gu et al. (2022) introduced VQDM, which segmented images into patches and used VQ-VAE to model the relationship between patches and token indices, significantly improving computational efficiency. Then, commercial models such as Midjourney (2022)

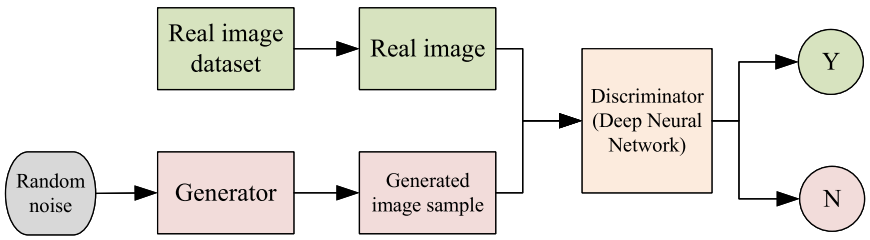

Fig. 2 Basic framework of generative adversarial network

and Wukong (2023) emerged, greatly enhancing the quality and speed of image generation. The basic process of diffusion models is shown in Fig. 3.

In 2024, Shirakawa and Uchida (2024) proposed a novel layout-aware text-to-image diffusion model, NoiseCollage, which independently estimates the noise for each object and then crops and merges them into a single noise, helping to avoid conditional mismatches. Shiohara and Yamasaki (2024) introduced the Face2Diffusion method for highly editable facial personalization. They remove identity-irrelevant information from the training process to prevent overfitting and improve the editability of facial encoding. Cao et al. (2024) proposed LeftRefill, a model that effectively learns the structural and texture correspondence between the reference and target without the need for additional image encoders or adapters. Zhou et al. (2024a) introduced MIGC, a method that ensures the generated instances are accurately placed at specified locations based on a set of predefined coordinates and their corresponding descriptions. Hoe et al. (2024) proposed an interactive control model called InteractDiffusion, which extends existing pre-trained T2I diffusion models to better condition on interactions. Höllein et al. (2024) introduced ViewDiff, which leverages a pre-trained text-to-image model as a prior and learns to generate multi-view images through a single denoising process from real-world data.

2.1.2 Basic framework of AI-generated image detection

The rapid development of deep learning technology has led to its increasingly widespread application in various fields, achieving significant advantages in many areas, with AI-generated image detection being one of them. High-quality fake images can deceive the human eye and even mislead the public through media dissemination, causing panic. Therefore, there is an urgent need for detectors capable of identifying fake images. The basic framework for AI-generated image detection based on deep learning is shown in Fig. 4.

Specifically, the AI-generated image detection task aims to develop algorithms or models trained on synthetic images produced by specific generative model, enabling them to identify universal artifacts (e.g., texture anomalies, frequency-domain features, or semantic inconsistencies) that distinguish generated images from real ones. These artifacts typically stem from architectural limitations of generative models or biases in training data. During the detection phase, the model must exhibit strong generalization capabilities to recognize diverse artifact patterns introduced by previously unseen generative approaches.

2.1.3 Datasets for AI-generated image detection

To validate the performance of AI-generated image detection technology, many studies have published relevant datasets. Table 1 describes eight public datasets used for AI-generated image detection technology, showing the types of generative models included and the number of images. These datasets encompass mainstream generative models, with data scales ranging from tens of thousands to millions of samples, providing comprehensive benchmark resources for algorithm development. Notably, significant variations exist across datasets in terms of image resolution, content diversity, and distribution of generation methods, which impose more stringent requirements on the generalization capability of detection models. Among them, ForenSynths, AIGCDetect-Benchmark, and GenImage, as frequently utilized datasets, can effectively evaluate method generalizability.

2.2 Deepfake generation and detection

Since the taking off of deep learning before 2012, deep learning architectures have rapidly evolved, driving significant advancements in deepfake research (Edwards et al. 2024). With the support of this technology, the types of synthetic fake faces have become diverse, and their quality continues to improve, making it impossible for the human eye to make judgments. Driven by this demand, detection technologies have progressed rapidly, and researchers have begun to leverage various feature information to enrich the body of research in this field. Therefore, in the following sections of this paper, a more in-depth discussion will be provided. Before that, this section will introduce some fundamental knowledge of this field, including deepfake generation technologies, and existing deepfake detection datasets.

2.2.1 Deepfake generation

From conventional techniques, autoencoders (AE), and GANs to DMs, image generation architectures have continuously improved. In existing datasets, these techniques can

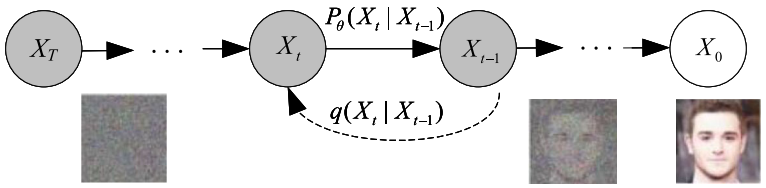

Fig. 3 Basic framework of diffusion models

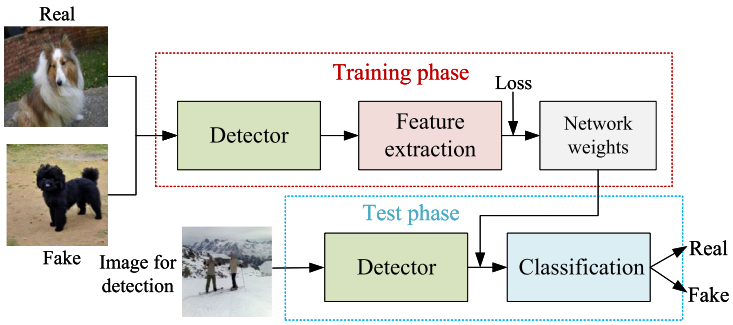

Fig. 4 Basic framework of AI-generated image detection on deep learning

generally be categorized into four types: identity swapping, face reenactment, attribute editing, and entire-face synthesis. Identity swapping involves directly replacing one person's facial features with another's, preserving all facial attributes. Face reenactment synchronizes one person's facial expressions or actions onto another person to create a forgery. Attribute editing allows modifications to certain facial features, such as age, gender, hairstyle, or skin color, without altering the identity of the person. Entire-face synthesis generates a completely virtual facial image using GANs or other generative models. Finally, we provide the generation process or examples of forged faces for these techniques, as shown in Fig. 5.

The four principal deepfake generation techniques represent a spectrum of facial manipulation capabilities as depicted in Fig. 5. Identity swapping (Fig. 5(a)) employs sophisticated encoder-decoder architectures where shared encoder weights learn universal facial structures while separate decoders maintain identity-specific features. During generation, the source face's latent features are redirected through the target's decoder, creating convincing identity transfers while preserving original expressions, lighting, and pose. Advanced implementations like DeepFaceLab enhance results through spatial attention mechanisms and refined blending algorithms that address boundary inconsistencies.

Face reenactment (Fig. 5(b)) operates by decomposing facial attributes into manipulable components—pose, illumination, expression, and identity—enabling precise expression transfer between individuals. Contemporary approaches have evolved from explicit 3D modeling to deep learning methods that implicitly model these transformations, with recent innovations achieving one-shot learning capabilities and improved temporal consistency across video frames. The technical pipeline typically involves expression extraction, mouth movement synchronization with original audio, and seamless compositing.

Attribute editing (Fig. 5(c)) enables more selective manipulation by targeting specific facial characteristics like age, gender, hairstyle, or skin tone while preserving core identity features. These methods leverage latent space manipulation techniques where facial attributes are disentangled into independently controllable dimensions, with StyleGAN-based approaches proving particularly effective through the identification of specific latent directions corresponding to desired attributes. More sophisticated systems now employ contrastive learning and semantic segmentation to achieve localized modifications with photorealistic transitions.

Entire-face synthesis (Fig. 5(d)) represents perhaps the most technically ambitious approach, creating completely artificial faces through generator-discriminator architectures

Table 1 Datasets for AI-generated image detection

Datasets	Year	Generator	Number of images (k)	
			Fake	True
ForenSynths (Wang et al. 2020)	2020	GANs	362	262
Diffusiondb (Wang et al. 2022b)	2022	DMs	14000	0
DE-Fake (Sha et al. 2024a)	2023	DMs	20	60
Artifact (Rahman et al. 2023)	2023	GANs, DMs	1522	962
AIGCDetectBenchmark (Zhong et al. 2023)	2023	GANs, DMs	433	433
Cifake (Bird and Lotfi 2024)	2024	DMs	60	60
Genimage (Zhu et al. 2023)	2024	GANs, DMs, Others	133	1350
Fake2M (Lu et al. 2023)	2024	GANs, DMs, Others	2000	0
Wildfake (Hong and Zhang 2024)	2024	GANs, DMs, Others	2577	1313

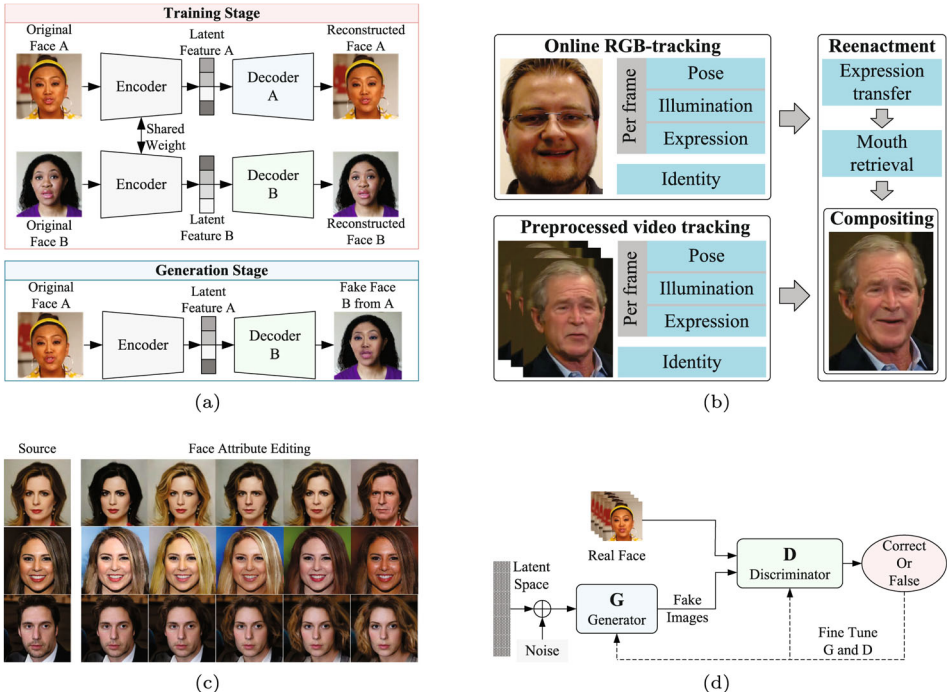

Fig. 5 Examples of the generation process or forged faces for the four forgery types: (**a**) Deepfakes (Deepfakes github 2018) forgery technique based on encoder-decoder architecture, belongs to faceswap; (**b**) Synthesis framework for Face2Face (Thies et al. 2016) in face reenactment, where the expression of the source face is modified; (**c**) Some examples of fake faces with attribute editing, where the hairstyle, gender or age has been modified; (**d**) General architecture for virtual face generation

that transform random noise vectors into real images. The latest diffusion models have further elevated quality through iterative denoising processes, producing results with remarkable detail including pore-level textures and consistent lighting interactions. These synthetic faces challenge even expert human inspection, incorporating subtle imperfections and asymmetries that characterize real human faces. Each technique presents unique forensic challenges—from inconsistent reflections in identity swapping to unnatural blinking patterns in reenactment, artificial boundaries in attribute editing, and geometric anomalies in synthetic faces—driving continuous advancement in both generation sophistication and corresponding detection countermeasures.

Although existing deepfake generation methods still face several technical challenges, such as persistent issues with temporal consistency in videos, where artifacts often appear during rapid movements or extreme expressions, and environmental interactions like shadows, reflections, and occlusions that frequently reveal the forged nature, the continuous improvement in computational power is leading to increasingly high-quality generated content. This may result in synthetic media that is nearly indistinguishable from reality, necessitating more robust detection methods to identify such forged content.

2.2.2 Existing deepfake datasets

The dataset is primarily used to train, validate, and evaluate the quality and performance of the model. The earliest deepfake detection datasets, such as UADFV (Korshunov and Marcel 2018) and Deepfake TIMIT (Li et al. 2018), not only had a small quantity but also poor quality, to the extent that the authenticity could be easily discerned by the human eye. However, with the rise of deepfake detection and the increasing involvement of researchers, various high-quality and challenging datasets have been proposed. Datasets such as FaceForensics++ (FF++) (Rossler et al. 2018), WildDeepfake (WDF) (Zi et al. 2020), and deepfake detection challenge (DFDC) (Dolhansky et al. 2020) have become widely used in scientific research, and the forgery techniques used are diverse, significantly increasing the difficulty of detection. Among them, the FF++ dataset uses five deepfake techniques: DeepFakes, Face2Face, FaceSwap, NeuralTextures, and FaceShifter, with three different compression levels—c0, c23, and c40. It has been used for model training in nearly all related studies. Therefore, to assist future researchers in understanding this field, we have organized the existing datasets based on four aspects: modality, real/fake quantity, source, and generation technique, as shown in Table 2.

2.3 Evaluation metrics

AI-generated image detection and deepfake detection are binary classification tasks, where the input data are classified as either real (positive) or fake (negative). Therefore, in actual classification, four possible outcomes can occur, including true positive, false negative, false positive and true negative, which can be represented by a confusion matrix, as shown in Table 3.

Table 2 A summary of existing deepfake datasets, including their modality, real/fake numbers, and generation techniques

Dataset	Year	Modality	Real/Fake	Source	Generation Technique
UADFV	2018	Video	49 / 49	YouTube	FakeAPP
Deepfake TIMIT	2018	Video	320 / 640	VidTIMIT	FaceSwap
FF++	2019	Video	1000 / 5000	YouTube	Deepfakes, Face2Face, NeuralTextures, FaceSwap, FaceShifter
DFD (Contributing data to deepfake detection research 2019)	2019	Video	363 / 3068	Live Action	Deepfakes
DFDC	2020	Video	23,654 / 104,500	Live Action	FaceSwap, NTH, FSGAN, StyleGAN
DFo (Jiang et al. 2020)	2020	Video	11,000 / 48,475	YouTube	FaceSwap
CDF-(v1, v2) (Li et al. 2020c)	2020	Video	590 / 5639	YouTube	DeepFake
WDF	2020	Video	3805 / 3509	Internet	Internet
KoDF (Kwon et al. 2021)	2021	Image	175,776 / 62,166	Live Action	FaceSwap, DeepFaceLab, FSGAN, FOMM, ATFHP, Wav2Lip
OpenForensics (Trung-Nghia et al. 2021)	2021	Image	45473 / 70325	Google Open Images	GAN
$FFIW_{10k}$ (Zhou et al. 2021)	2021	Video	10,000 / 10,000	Live Action	FaceSwap, FSGAN, DeepFaceLab
DFDM (Jia et al. 2023)	2022	Video	590 / 6450	YouTube	Facewap
DF-Platter (Narayan et al. 2023)	2023	Video	764 / 132,496	YouTube	FSGAN, FaceSwap, FaceShifter
Diffusion Deepfake (Bhattacharyya et al. 2024)	2024	Image	94120 / 112,627	DiffusionDB	Diffusion Model

In AI-created visual content detection field, commonly used evaluation metrics include accuracy, receiver operating characteristic curve, area under curve, equal error rate, and precision, which will be introduced below.

Accuracy (a) is the proportion of correctly predicted images out of all input images, a higher value indicates better model performance, which is expressed as (2):

$$a = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}} \tag{2}$$

Receiver operating characteristic (ROC) curve is commonly used to evaluate the performance of binary classifiers. First, we define the following three concepts: True positive rate (TPR, R_{TP}), true negative rate (TNR, R_{TN}), false positive rate (FPR, R_{FP}), and false negative rate (FNR, R_{FN}), which are expressed as (3)–(6):

$$R_{TP} = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{3}$$

$$R_{TN} = \frac{N_{TN}}{N_{TN} + N_{FP}} \tag{4}$$

$$R_{FP} = \frac{N_{FP}}{N_{FP} + N_{TN}} \tag{5}$$

$$R_{FN} = \frac{N_{FN}}{N_{FN} + N_{TP}} \tag{6}$$

The ROC curve plots the R_{TP} on the vertical axis and the R_{FP} on the horizontal axis. By setting different thresholds, diverse values of R_{TP} and R_{FP} are obtained. As the threshold decreases, more instances are classified as positive, but these positive instances also include some negatives, causing both

Table 3 Confusion matrix

True Label \ Prediction Label	Positive	Negative
Positive	True Positive (N_{TP})	False Negative (N_{FN})
Negative	False Positive (N_{FP})	True Negative (N_{TN})

R_{TP} and R_{FP} to increase simultaneously. When the threshold is at its maximum, the corresponding point on the ROC curve is (0, 0), and when the threshold is at its minimum, the corresponding point is (1, 1).

Area under curve (AUC) refers to the area under the ROC curve. Since the ROC curve typically lies above the diagonal line $y = x$, the AUC value ranges from 0.5 to 1. A larger AUC indicates better model performance. Specifically, an AUC value of 0.5 suggests a model with no discriminative ability (equivalent to random guessing), while an AUC value of 1 indicates perfect classification performance.

Equal error rate (EER) is the point on the ROC curve where the R_{FP} and R_{FN} are equal. Since $R_{FP} = 1 - R_{TP}$, EER occurs when $R_{FP} = R_{FN}$ on the ROC curve. This can be determined by drawing a line from (0, 1) to (1, 0) on the ROC curve. The lower the EER, the higher the classification accuracy of the model.

Precision (p), based on the prediction results, is the proportion of correctly predicted positive samples among those predicted as positive. The results predicted as positive can fall into two categories: either they are N_{TP} or N_{FP}. It can be expressed as (7):

$$p = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{7}$$

Recall (r), based on the actual samples, is the proportion of correctly predicted positive samples among all actual positive samples. The actual positive samples can be either correctly predicted as N_{TP} or incorrectly predicted as N_{FN}. be expressed as (8):

$$r = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{8}$$

3 AI-generated image detection methods based on deep learning

The AI-generated image detection methods can be categorized based on modality into unimodal and multimodal approaches. When classified according to the required features, unimodal methods can be further divided into spatial domain-based, frequency domain-based, and cross-domain features fusion-based. Multimodal methods mainly focus on image-text combinations. This section introduces various AI-generated image detection methods. At the end of this section, a table is provided that compares the characteristics, advantages, and limitations of these methods. An overview of deep learning-based AI-generated image detection methods is shown in Fig. 6.

3.1 Spatial domain-based

Most methods are based on the spatial domain, so this section will categorize spatial domain-based methods into spatial feature-based, color feature-based, texture feature-based, and other methods.

3.1.1 Spatial features

Spatial feature-based detection methods directly analyze the geometric and structural characteristics of images in the

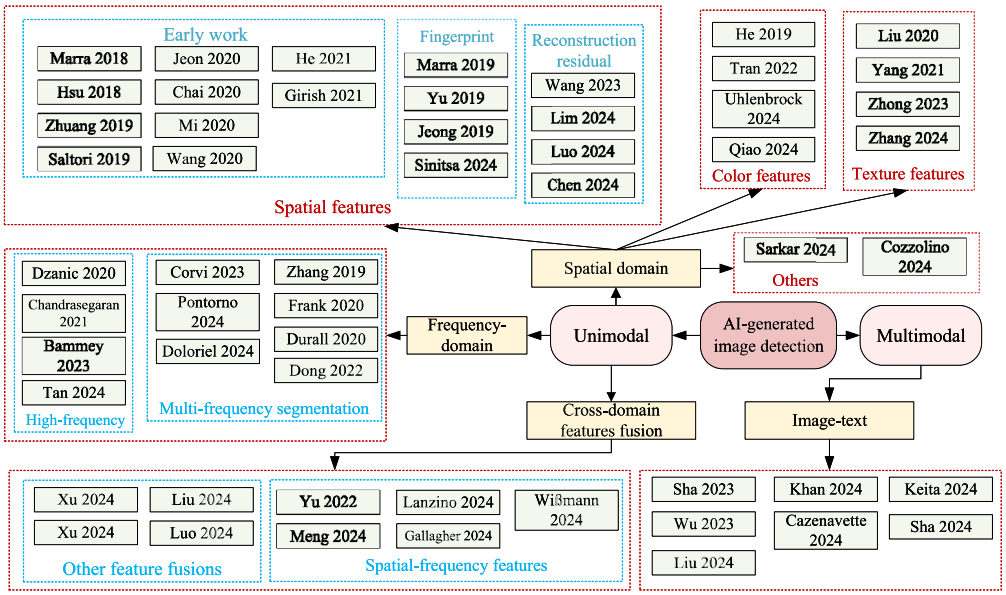

Fig. 6 Taxonomy of AI-generated image detection methods based on deep learning

spatial domain. These methods primarily focus on edge information, contour features, shape descriptors, and spatial distribution patterns within the image. Among these, Marra et al. (2018) conducted a comparison of several detectors for generated image detection, highlighting performance differences and finding that only deep learning-based detectors maintained high accuracy even on compressed data. Hsu et al. (2018) employed contrastive loss to identify typical features of synthetic images generated by different GANs, followed by the connection of a classifier to detect these computer-generated images. Yu et al. (2019) proposed the first study on learning GAN fingerprints for image attribution. They first reconstructed images through an autoencoder to obtain image fingerprints, then calculated the similarity between image fingerprints and GAN fingerprints to trace back to the image generation model. Zhuang and Hsu (2019) focused on detecting generated images based on GAN noise residual similarity, proving the existence of GAN fingerprints and discovering that training the same architecture on different datasets produced distinct fingerprints. Marra et al. (2019a) utilized triplet loss to learn discriminative features between real and generated images, and then introduced a novel dual network that accurately captured both local and global features of forged or authentic images. Marra et al. (2019b) employed an incremental learning approach for generated image detection, which also showed good performance when detecting previously unseen GAN models.

In 2020, Wang et al. (2020) incorporated data augmentation techniques into the training phase, which improved the model's performance and generalization ability. The GAN dataset they proposed has also been widely used. Jeon et al. (2020) introduced the transferable GAN-images detection framework (T-GD), which consists of a teacher model and a student model. These two models iteratively teach and evaluate each other, thereby enhancing detection performance. Chai et al. (2020) divided images into multiple patches and employed a patch-based classifier for generated image detection, finding visible artifacts in some image patches. Mi et al. (2020) applied a self-attention mechanism to extract global information for generated image detection. In 2021, He et al. (2021) employed super-resolution, denoising, and colorization as reconstruction methods to obtain residuals by regenerating denoised images and used a binary classifier for detection. Li et al. (2020b) constructed reference features for both real and generated images from datasets, and performed generated image detection by calculating the similarity between the input image and the reference features. However, this method depends on the quality and category of the datasets. Girish et al. (2021) employed an iterative algorithm to detect previously unseen GAN-generated images by leveraging the unique fingerprints left by GANs on the generated images.

In 2022, Liu et al. (2022) identified distinct characteristics in the noise patterns of real and generated images and designed an LNP extraction block to extract noise features from the images. By exploiting the noise patterns unique to real images, they designed a simple classifier for detection. Jeong et al. (2022) synthesized images from real images and used them to train a classifier. Zhang et al. (2022) proposed an unsupervised domain adaptation strategy to improve the generalization capability of the model. Ju et al. (2022) extracted global spatial information and local informative features from multiple image patches and utilized a multi-head attention mechanism to fuse the global and local features for image detection. Mandelli et al. (2022) divided images into patches and input them into multiple CNNs to obtain various features, then used ensemble learning to detect generated images.

In 2023, previous methods mainly targeted GANs, and due to the superior image generation quality of diffusion models, these methods performed poorly on images generated by diffusion models. To address this issue and improve generalization across unknown generative models, many new approaches have emerged in recent years. Tan et al. (2023) argued that the trend of pixel-level changes differs between real and generated images, and employed CNNs as a transformation model to convert images into gradients, using these gradients as input for image detection. Ojha et al. (2023) used CLIP (Radford et al. 2023) for generative image detection and found that CLIP exhibited good generalization capability for detecting generated images. To address the challenges faced by existing detectors in detecting images generated by diffusion models, Wang et al. (2023c) utilized the diffusion reconstruction error (DIRE) between real and synthesized images before and after reconstruction to detect diffusion model-generated images. However, this approach showed a decline in performance when detecting GAN-generated images due to its focus on diffusion models. Xu et al. (2023) developed a hybrid neural network based on attention-guided feature extraction (AGFE) and vision transformers-based feature extraction (ViTFE) modules, designed to capture both long-range and global features.

In 2024, Ju et al. (2024), building on the work in Ju et al. (2022), divided ResNet-50 into multiple layers to separately extract low-level and high-level features. They then constructed a patch selection module to extract high-energy patches, followed by the fusion of global and local features for image detection. Zhang et al. (2024c) developed a three-branch network, which alternately trains between the backbone network and auxiliary networks. Tan et al. (2024c) proposed neighboring pixel relationships (NPR) for image detection, based on the observation that upsampling causes correlations between neighboring pixels. Lim et al. (2024), building on the work in Wang et al. (2023c), introduced a lightweight diffusion-synthesized AI-generated

image detector with faster computation. Yan et al. (2024b) expanded the forgery space by constructing and simulating internal and inter-feature variations in the latent space, improving the model's generalization ability. Wang et al. (2024c) proposed a method to check if the examined images could be well-reconstructed using an inverted latent input to detect generated images. Chen et al. (2024a) reconstructed both real and generated images, and based on four categories of images-real, real-reconstructed, fake, and fake-reconstructed, they used contrastive learning loss to train a classifier, achieving a more accurate decision boundary. Sinitsa and Fried (2024) proposed the use of the deep image fingerprint method for generative image detection. He et al. (2024b) introduced RIGID, a model-agnostic, training-free method for detecting AI-generated images, which works by comparing the similarity between the original image and a slightly perturbed version in the visual model representation space. Liang et al. (2024) repeatedly employed contrastive learning to extract common features from real images, concatenating them with generated images features, and feeding them into a detector. Chen et al. (2024b) proposed randomly cropping the image, selecting the simplest patch, resizing it, and feeding it into the detector. Zhang and Xu (2023) utilized a pre-trained diffusion model to extract universal image representations for generative image detection. Finally, Yang et al. (2024) introduced the discrepancy AI-generated image detector framework, incorporating a parallel network branch that uses distorted images as additional discrepancy signals to supplement the original images, thereby learning general forgery features from multiple generators. However, spatial feature-based detection methods face significant limitations, such as limited effectiveness in detecting high-quality AI-generated images. As the quality of generative models improves, the discriminative power of traditional spatial features gradually diminishes. The comparison of spatial feature-based methods is shown in Table 4.

3.1.2 Color features

Some studies have found that generative models fail to capture the color characteristics of real images and have proposed detection methods based on color features.

In 2019, He et al. (2019a) utilized residual signals from chroma components in multiple color spaces and employed CNNs to learn robust deep feature representations. These features were then input into a random forest classifier to obtain the final detection results. In 2022, Chandrasegaran et al. (2022) analyzed transferable forensic features (T-FF) in universal detectors and proposed a novel forensic feature relevance statistic (FFRS) to quantify and discover T-FF in these detectors. Additionally, it was found that color is a key T-FF in the detectors. Uhlenbrock et al. (2024) found that natural and synthetic images differ in their color statistics, and thus, they used simple handcrafted color functions combined with random forests for generative image detection. Qiao et al. (2024) selected several color components that exhibited significant differences between real and synthetic images and employed the cross-color spatial co-occurrence matrix (CSCM) to extract color features for generative image detection. The comparison of color feature-based methods is shown in Table 5.

3.1.3 Texture features

AI-generated images (such as those generated by GANs) often exhibit differences from real images in aspects like texture consistency. Therefore, texture features can be helpful in distinguishing these images from authentic ones.

In these texture feature-based works, Liu et al. (2020) identified significant differences in texture features between generative and real images, and proposed the use of the gram matrix to construct a gram-net for generative image detection. Yang et al. (2021) introduced a novel multi-scale deep texture learning method, which captures multi-scale and deep texture features, incorporating an attention mechanism for feature fusion. Zhong et al. (2023) observed that for existing generative models, synthesizing realistic and rich textured regions is more challenging. To address this, they leveraged the pixel-wise correlation between rich and sparse texture regions in images to distinguish generative images. Zhang et al. (2024f) proposed a deep local binary pattern network (DLBPNet), where each branch contains filters and LBP feature extraction, followed by a central difference convolution module to learn more advanced features. The comparison of texture feature-based methods is shown in Table 6.

3.1.4 Others

Lorenz et al. (2024) proposed using the lightweight multi local intrinsic dimensionality (MultiLID) method for detecting generated images, demonstrating strong performance in detecting images generated by diffusion models. Lin et al. (2023) employed genetic programming for generative image detection, emphasizing the need for interpretability in the detection process. Sarkar et al. (2024) discovered that generative models cannot perfectly replicate projective geometric shapes, and developed a detection method using newer cues, such as object-shadow cues, perspective field cues, and line segment cues, to detect generated images. In order to address the need for detection without relying on generated images training and to improve the ability to detect unknown generative models, Cozzolino et al. (2024) found that generated images exhibit higher encoding costs and proposed a zero-shot detection method based on information entropy that utilizes encoding costs. To address the issue of detector bias towards training and testing sources, Tan et al. (2024a)

Table 4 A comparison of spatial feature-based methods

Ref.	Year	Method	Advantage	Deficiency
Yu et al. (2019)	2019	Fingerprint attribution	Trace the image back to the specific generative model	More complex computation when there are many models
Wang et al. (2020)	2020	Data augmentation	The generalization ability of GAN-generated images is good	Poor generalization ability on diffusion models
Jeon et al. (2020)	2020	Teacher-student model	Transferable model and good detection accuracy	Poor generalization ability on diffusion models
Chai et al. (2020)	2020	Image patch	Extracting local features of the image	Ignoring global information
Mi et al. (2020)	2020	Self-attention mechanism	Focus on the artifact regions of the features	Some generative models do not use upsampling operations
Girish et al. (2021)	2021	Fingerprint attribution	Good generalization to unseen GANs	As the number of generative models increases, the computational cost grows
Liu et al. (2022)	2022	Noise pattern	Good generalization	The noise information in the compressed image affects detection performance
Jeong et al. (2022)	2022	Fingerprint recognition	Only real images are needed for training, avoiding data dependency	As the number of generative models increases, the computational load grows
Tan et al. (2023)	2023	Gradient feature	Excellent detection performance on GAN-generated images	Poor detection performance on non-GAN generated images
Ojha et al. (2023)	2023	CLIP model	CLIP demonstrates good generalization capability in detecting generated images	The method is simple, and the accuracy is not high
Wang et al. (2023c)	2023	Reconstruction error	Performs well in detecting on diffusion models	Performs poorly on non-diffusion models
Tan et al. (2024c)	2024	Pixel correlation	Simple to compute, with good generalization	Relies on upsampling operations, with limitations
Lim et al. (2024)	2024	Reconstruction error	Lightweight network, faster computation	Has limitations for diffusion models
Yan et al. (2024b)	2024	Data augmentation	The method can be combined with other networks to improve generalization	It causes the computation time of other networks to increase
Chen et al. (2024a)	2024	Image reconstruction	The method can be combined with other detectors	Additional reconstruction dataset is required

employed invariant operators (such as the laplacian operator) alongside backbone networks for generative image detection, demonstrating a method that performs well without the need for training. The comparison of other methods is shown in Table 7.

3.2 Frequency domain-based

In the frequency domain, AI-generated images often exhibit unique artifacts and structural features, which provide strong clues for detection. Frequency-domain-based detection methods

Table 5 A comparison of color feature-based methods

Ref.	Year	Method	Advantage	Deficiency
He et al. (2019a)	2019	Chrominance components	Strong robustness	Limited generalization
Chandrasegaran et al. (2022)	2022	Relevance statistic	Discover that color is a critical feature in universal detectors	Images generated by diffusion models are similar to real images in terms of color
Uhlenbrock et al. (2024)	2024	Color statistics	High accuracy	Not tested on GAN datasets
Qiao et al. (2024)	2024	Co-occurrence matrix	Exhibits strong robustness	The experiment is simple

Table 6 A comparison of texture feature-based methods

Ref.	Year	Method	Advantage	Deficiency
Liu et al. (2020)	2019	Texture differences	Using texture differences for generated image detection	Limited generalization
Yang et al. (2021)	2021	Multi-scale texture	Extract multi-scale and deep texture information from the image	The network is complex and computationally intensive
Zhong et al. (2023)	2023	Texture contrast	Good generalization ability	Dependent on the high-frequency components of the image
Zhang et al. (2024f)	2024	Deep LBP network	Extract depth texture information	The experiment is simple and unable to validate the performance of the method

have gradually become an effective technical approach, and many researchers have proposed various frequency-domain analysis techniques to identify these traces of image generation. Zhang et al. (2019) utilized unique artifacts in the frequency domain generated by GAN upsampling operations for generative image detection. Frank et al. (2020) applied 2D-discrete cosine transform (DCT) to transform images into the frequency domain and performed generative image detection using ridge regression, optimizing parameters through grid search. Durall et al. (2020) designed a spectral regularization term and incorporated it into the training optimization objective. This approach enables the training of spectral-consistent GANs that avoid high-frequency errors. Dzanic et al. (2020) focused on high-frequency features for image detection and proposed a method to modify the high-frequency features of deep network-generated images to better mimic real images, also finding that spectral features are more easily distinguishable at high resolution and low compression rates. Bonettini et al. (2021a) used Benford's Law to extract a compact feature vector from images, which could be input into a very simple classifier for generative image detection.

Chandrasegaran et al. (2021) discovered that the high-frequency Fourier spectral decay differences are not inherent features of existing CNN-based generative models and, therefore, cannot be used for robust generative image detection. Dong et al. (2022) proposed a method to mitigate spectral artifacts, effectively reducing artifacts in the spectra of generated images, thus significantly improving the performance of frequency-based detectors. Corvi et al. (2023b) identified that diffusion models, like GANs, exhibit distinct fingerprints in the frequency domain, and incorporating a diffusion model during training helps detect images generated by diffusion-based models. Corvi et al. (2023a) further found significant differences between real and generated images in spatial autocorrelation functions, angular spectra, and radial spectra. Based on these differences, they designed high-performance detectors for generative image detection.

In 2024, Bammey (2024) used high-pass filtering to obtain image residuals, then applied Fourier transforms to generate spectral maps, which were subsequently fed into a classifier for detection. Pontorno et al. (2024) conducted a detailed examination of the statistical distribution of discrete cosine transform coefficients, training machine learning classifiers on different combinations of coefficients. Their experiments revealed that traces left by generative models are more distinguishable and persistent under JPEG attacks. Tan et al. (2024b) proposed a novel frequency-aware method called FreqNet, which centers on frequency learning and focuses on the high-frequency components of images. The method employs both phase and magnitude spectra for classification. Their designed frequency-domain learning module is capable of learning features that are independent of generative models, thus improving the model's generalization ability. To address the issue of model overfitting to training data, Doloriel and Cheung (2024) introduced a frequency-domain masking approach. By applying random masks to the image's frequency domain, their method prevents the detector from

Table 7 A comparison of other methods

Ref.	Year	Method	Advantage	Deficiency
Lorenz et al. (2024)	2023	Local intrinsic dimensionality	Good performance in diffusion models	Dependent on data augmentation
Lin et al. (2023)	2023	Genetic programming	Can improve accuracy to some extent	Limited generalization ability
Sarkar et al. (2024)	2024	Projective geometry	Having some level of generalization	Lacks effective defense against some attacks
Cozzolino et al. (2024)	2024	Coding cost	Good generalization ability	Weak robustness

overfocusing on all the image information. Weng (2024) proposed an innovative local frequency analysis (LFA) method, which combines medium-scale frequency domain analysis using Krawchouk moments and fine-scale frequency domain analysis via discrete cosine transform. This method enables multi-scale frequency analysis of images, allowing for the extraction of more comprehensive features. The comparison of frequency domain-based methods is shown in Table 8.

3.3 Cross-domain features fusion-based

As generative models continue to evolve, the quality of generated images has significantly improved, making them increasingly difficult for the human eye to distinguish. As a result, conventional detection methods have become less effective in identifying outputs from newer generative models. The performance of generators designed with only single-domain features is limited, and as a result, an increasing number of detectors are utilizing cross-domain feature fusion techniques for detecting generated images. Yu et al. (2022) proposed a method based on channel difference image (CDI) and spectrum image (SI), employing octave convolution and an attention-based fusion module. This approach effectively extracts intrinsic features from these two domains to detect AI-generated images. To capture GAN fingerprints at different levels, Liu et al. (2024b) introduced a decoupling representation framework, which is designed to separate and extract two types of GAN fingerprints from different domains. This framework also incorporates an adversarial data augmentation strategy and a transformation-invariant loss to enhance the robustness of the fingerprints against image perturbations.

In 2024, The cross-domain features fusion method has been rapidly developing, with an increasing number of approaches leaning towards identifying distinguishing features between real and generated images from multiple domains. Luo et al. (2024) proposed the latent reconstruction error (LaRE), the first feature based on reconstruction error in latent space. They also introduced an error-guided feature refinement module (EGRE) that refines image features guided by LaRE to improve their discriminability. The EGRE utilizes an alignment and refinement mechanism to effectively refine image features from both spatial and channel perspectives for generative image detection. Lanzino et al. (2024) constructed a classifier using a combination of fast Fourier transform (FFT) frequency-domain features, local binary pattern (LBP) texture features, and pixel-domain features. They also employed convolution to reduce the number of channels for detecting generated images. Wißmann et al. (2024) trained a classifier using DCT, power spectral density (PSD), and autocorrelation coefficients, finding that DCT and PSD demonstrated excellent performance in robust detection and high-precision attribution. Gallagher and Pugsley (2024) proposed a dual-branch neural network architecture that takes both images and their Fourier frequency decompositions as inputs. The network uses CNNs for feature fusion to detect generated images. Meng et al. (2024) introduced the artifact purification network (APN), separated and extracted artifact features from both the spectrum and spatial domain, and further aggregated the diluted artifact information in the features. Xu et al. (2024a) designed a global feature extraction module based on attention using MobileViT to learn deep representations of global tracking information. Additionally, multiple enhanced residual blocks are employed to extract

Table 8 A comparison of frequency domain-based methods

Ref.	Year	Method	Advantage	Deficiency
Zhang et al. (2019)	2019	Frequency artifacts	Detect frequency differences between real and generated images	Limited generalization
Frank et al. (2020)	2020	2D-DCT	Discover that color is a critical feature in universal detectors	Images generated by diffusion models are similar to real images in terms of color
Durall et al. (2020)	2020	High-frequency Fourier modes	Transferable model and good detection accuracy	Poor generalization ability on diffusion models
Corvi et al. (2023b)	2023	Training DM's images	Enhancing the performance of detecting diffusion model images	With the emergence of new generative models, updates are continuous
Tan et al. (2024b)	2024	Frequency learning	It can learn features unrelated to the generative model, enhancing the model's generalization ability	High computational cost and time-consuming
Doloriel and Cheung (2024)	2024	Frequency mask	Less dependence on detector data	The mask size affects the performance of the detector

distinctive multi-scale features. Leporoni et al. (2024) argued that the generation of fake content introduces potential inconsistencies in the depth of generated images. They proposed a method that inputs both RGB and depth images into the backbone network, which utilizes an RGB attention mechanism to perform final feature fusion. Yan et al. (2024a) concatenated DCT high and low-frequency local image block features with CLIP global semantic features for classification. However, such methods introduce significantly increased complexity and require substantial computational resources. The design of feature fusion strategies demands domain expertise and may suffer from issues such as feature redundancy and the curse of dimensionality. The comparison of cross-domain feature fusion methods is shown in Table 9.

3.4 Image-text

With the development of multimodal language models like CLIP and Flamingo, these models have demonstrated good generalization capabilities in AI-generated image detection. As text-to-image generation technology becomes more widespread, many studies have focused on utilizing multimodal models such as CLIP as the backbone network for detection. By combining the visual features of image content with the semantic features of text descriptions, detection is performed through the consistency, contrast, and alignment between the image and text, thus enabling the identification of generated images.

Based on the aforementioned multimodal language models, Sha et al. (2024a) investigated how the prompts used to generate generated images affect detection and attribution, and verified that generated images exhibit a higher correlation with text descriptions. To address the generalization issue of detectors, Wu et al. (2023a) designed textual labels to improve performance. Moreover, they transformed the synthetic image detection problem into a recognition task, fine-tuning CLIP's image encoder and text encoder based on contrastive learning loss, bringing the features of generated images closer to the "fake photo" prompt text features. Liu et al. (2024c) proposed the forgery-aware adaptive transformer approach (FatFormer). First, a forgery-aware adapter detects local forgery traces in both image and frequency domains. Then, FatFormer incorporates language-guided alignment, supervising forgery adaptation using both image and text prompts. Cazenavette et al. (2024) mapped the original image to latent space using the VAE encoder from stable diffusion. They then used DDIM for inversion and reconstruction, relying on the CLIP embeddings of BLIP-generated image captions. Finally, they concatenated the original image, decoded the noise image, and decoded the reconstruction image for generative image detection. Khan and Dang-Nguyen (2024) built four detectors based on different strategies using CLIP, including prompt tuning, adapters, fine-tuning, and linear probing, for generative image detection. Keita et al. (2024b) proposed an innovative method called Bi-LORA, which combines visual-language models (VLMs) and low-rank adaptation (LoRA) tuning techniques to improve the detection accuracy of synthetic images generated by unseen models. Building on this approach, Keita et al. (2024a) further used Bi-LORA to learn text features that align with image features and then employed a large language model to output real or fake text. Sha et al. (2024b) first generated DDIM inversion noise through adversarial prompts. They then reconstructed the image from the generated noise and compared the reconstructed image with the original image to determine whether the image was real or fake. Although the above methods achieve strong performance, they involve processing both visual and linguistic modalities, making model training more complex and increasing the demand for annotated data. Moreover, they are sensitive to the quality of text-image pairing and rely heavily on high-quality multimodal datasets, which may introduce bias issues in cross-lingual applications. The comparison of image-text-based methods is shown in Table 10.

Table 9 A comparison of cross-domain features fusion methods

Ref.	Year	Method	Advantage	Deficiency
Yu et al. (2022)	2022	Channel and spectrum difference	Effectively mine intrinsic features	Limited generalization ability
Luo et al. (2024)	2024	Reconstruction error	Able to extract refined features from images	Poor performance on non-diffusion models
Lanzino et al. (2024)	2024	Three types of feature fusion	Capture multiple features of the image with a simple network	Weak resistance to adversarial attacks
Xu et al. (2024a)	2024	Deep trace feature fusion	Good generalization performance	Complex network with long computation time
Leporoni et al. (2024)	2024	RGB-depth integration	RGB features capable of extracting depth	Weak resistance to adversarial attacks

Table 10 A comparison of image-text-based methods

Ref.	Year	Method	Advantage	Deficiency
Wu et al. (2023a)	2024	Contrastive learning	Transform the synthetic image detection problem into a recognition problem	Text description affects the performance of the detector
Liu et al. (2024c)	2024	Forgery-aware adaptive	Strong generalization ability	Computationally intensive and time-consuming
Cazenavette et al. (2024)	2024	Inverting stable diffusion	Good detection accuracy	Limited in the context of stable diffusion
Keita et al. (2024b)	2024	Technical optimizations	Combining BLIP and LoRA to enhance accuracy	Computationally intensive and time-consuming
Sha et al. (2024b)	2024	Image reconstruction	Requires no large training data and has good robustness	Mainly focused on diffusion models, with certain limitations

4 Deepfake detection based on feature selection

Deepfake detection methods can be broadly classified into four categories based on the feature information utilized, including methods based on spatial features, spatiotemporal features, biological features, and identity features. Among these, spatial features and spatio-temporal features can be considered general features, while biological features and identity features are considered special features, typically requiring specialized network architectures for extraction. Therefore, this section will present an article review in these four directions, and we have also summarized and organized the selected articles, as shown in Fig. 7.

4.1 Spatial feature-based

The initial deepfake detection methods primarily relied on spatial features, including texture features, tampering artifacts, etc. These methods focused on single-frame analysis, achieving simple and effective detection by extracting forgery clues from different domains. Based on the source of information, these methods can be categorized into space domain-based, frequency domain-based, and multi-domain fusion approaches.

4.1.1 Space domain information-based

In the field of deepfake detection, the space domain is a commonly used source of information, mainly containing visual cues such as edge blurring, illumination inconsistency, and abnormal shadows, using computer vision techniques, discriminative features such as texture, gradient, and statistical patterns can be extracted. Many researchers have improved model performance by designing network architectures, applying image preprocessing techniques, or leveraging specific spatial inconsistencies.

As an earlier method, Afchar et al. (2018) used a shallow convolutional network to extract mid-level features from images for forgery detection and achieved video-level detection through image aggregation. Li et al. (2020a) proposed Face X-ray, which detects the blending boundary in forged faces through synthetic data training. It performs classification while also locating the blending areas, but this approach does not apply to completely synthesized fake images. Bonettini et al. (2021b) used EfficientNetB4 as the backbone

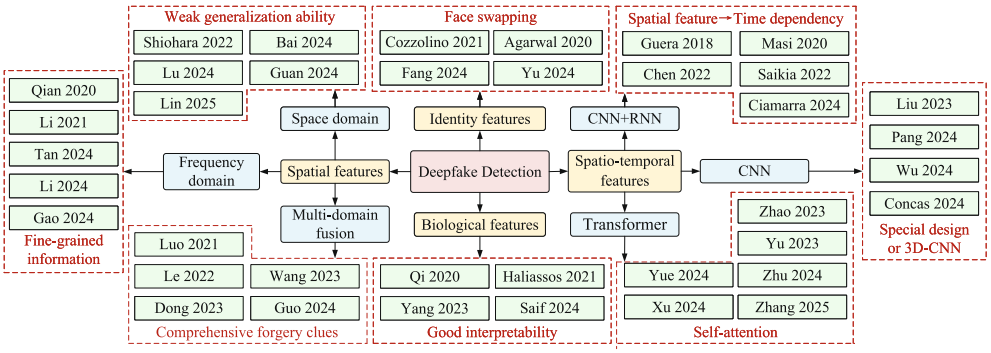

Fig. 7 Taxonomy of deepfake detection methods based on feature selection

network and incorporated attention layers and siamese training mechanisms, highlighting the role of these mechanisms through ablation studies.

To address the issue of limited generalization due to the commonly used fake data, researchers have proposed the idea of synthetic data, which helps make the trained models more generalizable. Shiohara and Yamasaki (2022) proposed self-blended images (SBI) to prevent the model from overfitting to specific forgery methods. By using a data synthesis strategy to reproduce general forgery artifacts, they improved the model's generalization ability. The general synthesis process of SBI is shown in Fig. 8. The basic idea is to generate the target image and the pseudo source image from the base image, and then blend the two images using a face mask, thereby creating general visual artifacts. Chen et al. (2022b) also employed data synthesis methods, but they enriched the diversity of synthetic data by introducing multi-configuration strategies and used adversarial training to enable the model to learn more robust feature representations. Lin et al. (2024b) designed self-shifted blending images to simply fuse temporal artifacts, searching for a suitable augmentation scheme during training. Their curriculum learning-based training strategy further enhanced model performance. Guan et al. (2021) introduced a gradient regularization term into the original loss function to reduce the model's sensitivity to texture features. The new loss function improved the model's robustness to shallow feature statistical perturbations and could be combined with existing backbone networks or methods to further enhance detection performance.

Additionally, Gao et al. (2024a) proposed separating texture and artifact information in the features and performing face and background separation using estimated masks obtained through self-supervised learning strategies. This allowed for the extraction of more detailed texture information, which was then combined with artifacts for detection. Lu et al. (2024b) proposed a long-distance attention mechanism based on fine-grained classification and designed spatial and temporal attention modules to obtain local region attention maps from single and consecutive frames. To address the generalization issue, Zheng et al. (2024) combined unsupervised-supervised contrastive learning for deepfake detection. They mined features from both original and data-augmented images, performed multi-scale fusion, and applied contrastive loss constraints between individual samples and diverse class features, achieving effective and stable detection. Since forgeries disrupt the consistency of regional noise, Bai et al. (2024) proposed a method that leverages the noise pattern differences between the face and background regions. They performed noise enhancement and multi-scale integration to effectively detect forged images. Ma et al. (2024) utilized incremental learning strategies to improve the model's generalization performance with limited samples and combined human perception saliency with self-attention to highlight important regions. Lu et al. (2024a) designed a multi-scale texture feature extraction module using central difference convolution, effectively enhancing the quality of texture features. They also introduced region-specific separable self-consistency loss to constrain the representation learning of different regions and emphasize important areas. These methods have the advantage of high computational efficiency and relatively simple implementation. However, they exhibit limited capability in detecting high-quality forgeries, struggle to adapt to the rapidly evolving forgery techniques, and suffer from poor robustness, making them vulnerable to post-processing attacks. Table 11 summarizes these methods from three aspects: key idea, backbone, and dataset.

4.1.2 Frequency domain information-based

Space domain-based detection methods can achieve good detection performance; however, when subjected to common attacks such as noise or compression, the forgery clues become harder to detect. Additionally, the traces left by different forgery methods vary, which limits further improvement in generalization performance. On the other hand, the frequency domain information of an image, especially high-frequency components, contains edges and other fine details that are more resilient to attacks. Furthermore, different forgery methods tend to generate unnatural artifacts in the frequency domain, which makes it possible to detect

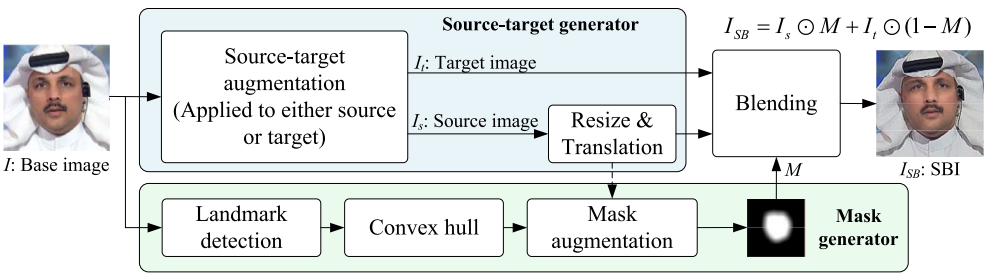

Fig. 8 The general synthesis process of SBI. $\odot$ denote element-wise multiplication operation

Table 11 A comparison of space domain information-based deepfake detection methods

Ref.	Year	Key Idea	Backbone	Dataset
Shiohara and Yamasaki (2022)	2022	Synthetic data	EfficientNetB4	FF++, CDF, DFD, DFDCp (Dolhansky et al. 2019), DFDC, $FFIW_{10k}$
Bai et al. (2024)	2024	Regional noise inconsistency	Xception	FF++, CDF, DFDC
Gao et al. (2024a)	2024	Feature decomposition	Convolution layer	FF++, WDF, CDF, DFDC
Lu et al. (2024b)	2024	Long-distance attention	Xception	FF++, CDF
Lin et al. (2024b)	2024	Synthetic data, curriculum learning	Transformer	FF++, CDF, DFDCp, DFDC, WDF

forgery traces that are difficult to identify in the space domain. Therefore, incorporating frequency domain information can enhance the robustness and generalization of the detection model.

In these methods, common frequency domain information extraction techniques include FFT, DCT, and discrete wavelet transform (DWT). Peng et al. (2023) designed a high-frequency residual extraction module based on the Laplacian pyramid, utilizing the high-frequency components of shallow features to extract visual artifacts. Qian et al. (2020) used DCT for domain transformation and integrated frequency-aware decomposition images and local frequency statistics through a dual-stream collaborative learning framework to mine forgery clues. Li et al. (2021) restructured DCT coefficients across different frequency bands while preserving the original spatial relationships, allowing the use of convolutional networks for frequency feature extraction. They also employed a single-center loss to compress intra-class variations and expand inter-class differences. Gao et al. (2024b) addressed the difficulty of detecting compressed data by proposing a high-frequency enhancement framework that integrates comprehensive frequency-domain information from block-wise DCT and DWT. Using a two-stage cross-fusion strategy, they effectively merged information and achieved high accuracy on highly compressed data. To address the limitation of self-attention in capturing subtle clues, Miao et al. (2023) introduced the central difference operators to extract fine-grained feature details and used DWT to supplement local high-frequency information, achieving strong accuracy and robustness. To supplement fine-grained information in transformer, Li et al. (2024) embedded wavelet transforms into self-attention and designed down-sampling strategies for information enhancement across stages. Through optimal data augmentation, they effectively improved generalization performance. Hasanaath et al. (2024) extracted discriminative generic features from self-blended images using DWT and fed them into a CNN for deepfake classification. Zhao et al. (2023b) introduced an adaptive Fourier neural operator to learn frequency-domain forgery clues and applied an efficient attention mechanism to enhance detailed information while reducing the computation.

A comparison of these methods is shown in Table 12. Table 12 describes the methods from three aspects: frequency domain information extraction methods, backbone and datasets.

4.1.3 Multi-domain information fusion-based

To leverage complementary information from different domains, many researchers have fused multi-domain features to obtain more comprehensive feature representations. Table 13 compares the multi-domain information fusion-based methods in terms of information sources, fusion methods, backbone, and datasets.

Wang et al. (2023a) calculated the residual between the original grayscale image and the low-frequency components of the DWT to obtain the mid-high frequency image,

Table 12 A comparison of frequency domain information-based deepfake detection methods

Ref.	Year	Transform Type	Backbone	Dataset
Qian et al. (2020)	2020	DCT	Xception	FF++
Li et al. (2021)	2021	DCT	Xception	FF++
Miao et al. (2023)	2023	DWT	Transformer	FF++, CDF, DFDC, Deepfake TIMIT
Zhao et al. (2023b)	2023	FFT	Convolution and attention layer	FF++-Deepfakes, FFHQ, CelebA
Gao et al. (2024b)	2024	DCT, DWT	Convolution and fusion layer	FF++, CDF, OpenForensics
Hasanaath et al. (2024)	2024	DWT	EfficientNetB5	FF++, CDF
Li et al. (2024)	2024	DWT	Transformer	FF++, CDF, DFDC, Deepfake TIMIT, DFo

Table 13 A comparison of multi-domain information fusion-based deepfake detection methods

Ref.	Year	Information Source	Fusion Method	Backbone	Dataset
Luo et al. (2021)	2021	RGB, SRM noise	Concatenation	Xception	FF++, DFD, DFDC, CDF, DFo
Fei et al. (2022)	2022	RGB, SRM noise	Attention-guided	ResNet-18	FF++, CDF, DFD
Wang et al. (2023a)	2023	RGB, DWT	Concatenation	Xception	FF++, CDF, UADFV
Guo et al. (2024)	2024	RGB, High-frequency	Interaction, concatenation	ResNet-26	HFF, FF++, DFDC, CDF
Wang et al. (2024a)	2024	RGB, DCT	Attention-guided	Xception	FF++, CDF
Wang et al. (2024b)	2024	RGB, DWT, Residual feature	Attention-guided	ResNet-34	FF++, CDF, UADFV, DFD
Zhang et al. (2024a)	2024	RGB, SRM noise	Attention-guided	EfficientNet	FF++, DFDC, CDF, WDF
Zhou et al. (2024b)	2024	RGB, FFT	Multihead-attention	EfficientNetB4	FF++, CDF, WDF

which was then concatenated with the RGB image and fed into a convolution network for classification. Wang et al. (2024b) integrated deep-frequency domain information extracted from residual maps reflecting facial edge information with wavelet frequency domain and RGB domain information. Zhou et al. (2024b) fused multi-scale RGB features with frequency-domain-aware features based on FFT. Le and Woo (2022) used attention distillation to transfer high-frequency components learned by a teacher model trained on high-quality data to a student model, enhancing feature discrimination under low-quality data conditions. To restore the model's attention to compressed artifacts, Wang et al. (2024a) designed a spatial-frequency feature fusion architecture and also employed knowledge distillation to transfer feature representations from a teacher model to a student model. Most existing methods focus on improving conventional convolutional backbones, but Guo et al. (2024) designed a new space-frequency interactive convolution module that integrates space domain information and high-frequency information through interaction, resulting in more refined feature representations.

High-pass filters in spatial rich model (SRM) can extract high-frequency noise from images, removing color textures and revealing the differences between the real and forgery regions. The three commonly used filtering kernels and their resulting noise residual images are shown in Fig. 9. Based on this observation, Luo et al. (2021) used SRM to extract multi-scale high-frequency residuals as an information branch and generated residual attention maps to highlight forgery clues in the RGB branch features. Their cross-modal fusion achieved efficient utilization of the dual-branch information. Zhang et al. (2024a) also fused high-frequency noise with spatial texture features and used local attention to enhance forgery traces. Fei et al. (2022) supplemented noise features while calculating first and second order local anomaly maps in the RGB branch, magnifying and learning the local anomaly information of forged images for more generalizable detection. Dong et al. (2023) treated SRM high-frequency noise as data augmentation and employed supervised contrastive learning to minimize the positive pair distance, improving the generalization performance of the model. Although these methods significantly improve detection performance, they also incur substantial computational overhead, increase the complexity of feature engineering, and may introduce redundancy among different feature representations.

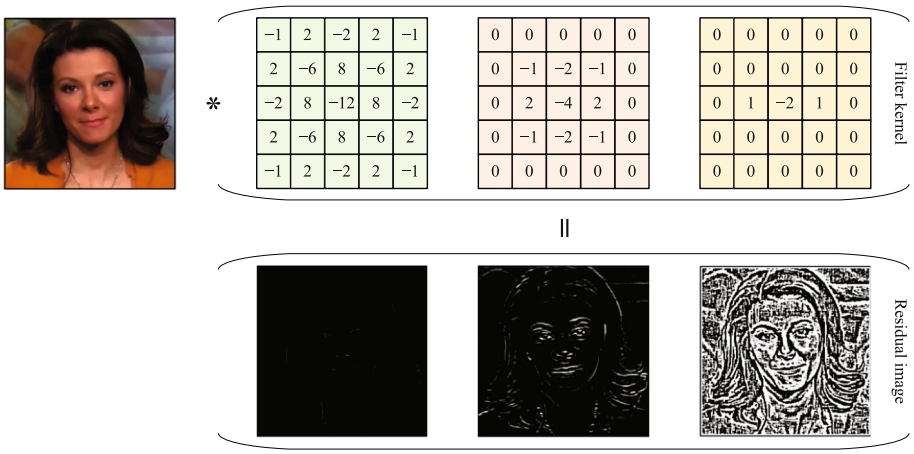

Fig. 9 Three commonly used high-pass filtering kernels in SRM and their resulting noise residual images

4.2 Spatio-temporal feature-based

The consecutive frames of an original video have natural consistency, but deepfake videos are composed of individual forged images linked together, which disrupts the original spatio-temporal consistency and introduces forgery traces in the temporal domain. Spatial feature-based detection methods fail to account for this disruption, making them unsuitable for video-level detection. As a result, some researchers have started designing frameworks for extracting spatio-temporal features. The backbone networks used in these methods mainly include CNN, recurrent neural networks (RNN), and transformer, so these methods can be divided into three categories: CNN-based, CNN+RNN-based, and transformer-based.

4.2.1 CNN-based

CNN-based methods typically involve special designs for feature extraction or the use of 3D CNNs. Liu et al. (2023) integrated RGB domain and frequency domain information, utilizing locally sensitive regions to enhance forgery features, and employed a 3D CNN to supplement temporal domain information. Concas et al. (2024) proposed an innovative method for extracting forgery artifacts by performing facial quality estimation on the face region of single-frame or consecutive-frame images and generating a quality feature matrix that is input into a CNN for forgery detection. Existing methods tend to capture spatio-temporal information with fixed time steps, rarely focusing on the extraction of dynamic spatio-temporal inconsistencies. Pang et al. (2023) designed a video sampling strategy, bipartite group samplin (BGS), which used different sampling rates to obtain multiple video frame sets and extracted short-term and long-term spatio-temporal information in subsequent networks, enabling full utilization of forgery clues. Zhang et al. (2024d) proposed a frame sampling strategy with temporal diversification and used self-contrastive learning to extract short-term and long-term temporal artifacts, reducing the model's sensitivity to binary labels. Yu et al. (2024b) applied multi-path dynamic inconsistency magnification to multiple groups of sampled frames to extract local-consecutive fine-grained features, used graph convolution network (GCN) to obtain global temporal views across multiple groups and designed a domain alignment module to improve generalization performance.

To fill the gap in frequency domain spatio-temporal information for deepfake detection, Wang et al. (2023b) proposed a frequency domain forgery clue augmentation strategy based on DCT. They first enhanced the high-frequency components of the DCT spectrum, then divided it into multiple blocks along the spatial dimension, replacing the original spectrum with the maximum response to reduce computational complexity. The attention map obtained from the frequency temporal attention module enhanced temporal clues. Wu et al. (2023b) designed patch-wise decomposable DCT to extract finer-grained high-frequency clues and extracted comprehensive spatio-temporal representations of both RGB and frequency branches in stages. An interaction module was used to eliminate cross-modality feature inconsistencies, achieving effective feature fusion.

The state-of-the-art CNN-based deepfake detection methods are described in Table 14. Table 14 describes the methods from three aspects: improved methods, backbone, and dataset.

4.2.2 CNN+RNN-based

Since CNNs are primarily used for extracting spatial information from images and are not effective at capturing temporal dependencies, applying RNNs for temporal modeling of the features extracted by CNNs has become an effective solution. RNNs, especially long short-term memory (LSTM) networks, are well-suited for modeling sequential data and capturing the temporal relationships between frames in video, which helps improve the detection of temporal inconsistencies in deepfake videos. By combining CNNs for spatial feature extraction with RNNs for temporal sequence modeling, such hybrid approaches can better leverage both spatial and temporal information for more accurate and robust deepfake detection.

The general process of using a CNN and RNN hybrid model for detection is shown in Fig. 10. In this process,

Table 14 Summary of methods for extracting spatio-temporal features using CNNs

Ref.	Year	Improved method	Backbone	Dataset
Liu et al. (2023)	2023	Local attention augmentation	3D ResNet-50	FF++, CDF, DFDC
Pang et al. (2023)	2023	Sampling strategy	ResNet-34	FF++, CDF, DFDC, WDF
Wang et al. (2023b)	2023	Attention augmentation	ResNet-50	FF++, CDF, WDF, DeepfakeNIR
Concas et al. (2024)	2024	Quality feature	Convolution layer	FF++
Yu et al. (2024b)	2024	Multilevel spatio-temporal features	ResNet-50, GCN	FF++, DFD, DFDC, CDF, DFo
Zhang et al. (2024d)	2024	Sampling strategy	EfficientNetB3	FF++, CDF, DFDC, WDF

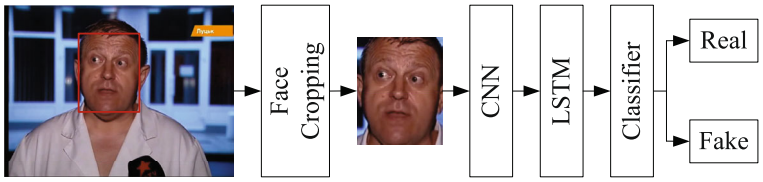

Fig. 10 The general detection process based on the CNN and RNN hybrid model

the CNN extracts spatial features from the facial image, the RNN performs temporal modeling on the spatial features, and finally, classification is performed. Based on this process, Guera and Delp (2018) used InceptionV3 to extract frame features and input them into an LSTM to learn inter-frame inconsistencies, achieving video-level classification. Saikia et al. (2022) utilized optical flow from consecutive face frames and fed it into a hybrid model of CNN and LSTM to extract temporal information. Chen et al. (2022a) introduced a spatio-temporal attention mechanism to enhance the temporal correlation between frames, and the augmented frames were input into Xception and ConvLSTM to extract spatial and temporal inconsistencies. Jayashre and Amsaprabhaa (2024) optimized network weights using the spotted hyena optimizer during the hybrid model training. Amerini and Caldelli (2020) used image prediction errors as inputs to incorporate temporal information; however, the increase in complexity led to a decrease in generalization ability. Masi et al. (2020) proposed a dual-stream network, with one branch extracting RGB features and the other using the laplacian of gaussian (LoG) operator to suppress facial visual content and extract high-frequency edge information. They also designed a new loss function based on the concept of one-class classifiers, which pulls positive samples closer while pushing negative samples further apart. Since deepfake videos often fail to preserve the inherent features left during the camera capture process, Ciamarra et al. (2024) used UprightNet to estimate camera orientation and generate surface frames. They leveraged temporal anomalies in these frames to detect forgery.

4.2.3 Transformer-based

Transformer was first applied to natural language processing (NLP) tasks and achieved significant improvements (Vaswani et al. 2017). To extend their application to vision tasks, researchers designed the vision transformer (ViT) (Dosovitskiy et al. 2021) and swin transformer (SwinT) (Liu et al. 2021), utilizing self-attention mechanisms to capture long-distance dependencies between different frames.

Given their powerful spatio-temporal modeling capabilities, many researchers have started applying these models to the field of deepfake detection. Yu et al. (2023) designed a multi-view modeling strategy based on transformer, where for multiple groups of consecutive frames, they first establish local spatio-temporal fusion features for each set and then connect them along the temporal channel to create global spatio-temporal fusion features. Huang and Zhang (2024) introduced an improved meta-learning approach to the spatio-temporal backbone, effectively enhancing generalization to unseen forgery methods. Zhao et al. (2023a) decomposed the computation of self-attention, using a self-subtract mechanism to make the model focus more on inter-frame distortions based on feature residuals, thus reducing redundant information. Inspired by correlation propagation algorithms, they also designed a visualization algorithm to improve the interpretability of the transformer. Liu et al. (2024a) used RGB images and motion flow to provide spatio-temporal information, modeling spatio-temporal feature connections with SwinT, and designed identity-decoupling attention to extract more general spatio-temporal feature representations that are independent of identity, thus effectively improving the model's generalization ability. Yue et al. (2024) used UniformerV2 as the backbone to extract global features and leveraged local frequency dynamic information, generated region-of-interest (ROI) attention maps by local region alignment, guided the global features to extract more refined forgery clues. Zhu et al. (2024) employed knowledge distillation to transfer fine-grained spatial-frequency knowledge and spatio-temporal structural knowledge to the student model, effectively improving the model's robustness to compression. Zhang et al. (2024b) proposed a self-supervised learning approach to learn the natural consistency representation of real face videos and used the fact that the consistency of deepfake videos is disrupted to distinguish authenticity, designing corresponding natural consistency enhancement strategies to improve detection accuracy.

In addition to the aforementioned methods, some studies detect deepfakes by preprocessing data or utilizing special information. Choi et al. (2024) found temporal variations in the style latent vectors of generated facial videos, so they used a StyleGRU module to capture the style latent vector and established a style flow based on the differences for subsequent input. Tian et al. (2024) extracted rich and robust forgery information by leveraging the temporal variation of local and global lighting information and the dynamic spatio-temporal inconsistencies of intra-frame/inter-frame forgery cues. Tu et al. (2024) employed optical flow difference algorithms to locate key facial expression frames as input, which, compared to using the entire video sequence, improved accuracy while reducing training time by nearly 75%. To reduce computational requirements, Xu et al. (2024b) designed a

thumbnail layout method that transforms consecutive frames into a predefined layout while preserving the original spatio-temporal relationships. After embedding modified positions, it effectively utilizes the transformer to learn spatio-temporal information.

These methods are particularly well-suited for detecting high-quality deepfake content and show clear advantages in complex scenarios requiring global contextual information. However, Transformer-based approaches also face challenges such as high computational resource requirements and increased overall model complexity, which limit their applicability in resource-constrained environments. Finally, we summarize the transformer-based methods, including three aspects: whether transformer network architecture design, whether input design, and datasets were performed, as shown in Table 15.

4.3 Biological feature-based

The detection methods based on general features have achieved good performance, but they suffer from a significant lack of interpretability. In contrast, biological features, due to their inherent regularity, are much easier to interpret when subjected to forgery distortion. They provide a more intuitive understanding and align better with human cognition. Therefore, detection methods based on biological features represent a promising research direction.

Next, some methods based on biological features are introduced. Yang et al. (2019) trained a classifier using the head pose differences between real and fake facial images for deepfake detection. Haliassos et al. (2021) detected fakes by exploiting the high-level semantic irregularities of mouth movements in forged videos, and their method is robust to most data corruptions. Demir and Ciftci (2021) extensively analyzed features related to eyes and gaze, integrating visual, geometric, and temporal information, achieving superior detection results compared to methods based on a single biological feature. Peng et al. (2024) aggregated gaze direction, facial attributes, and texture information as spatio-temporal features to enhance the model's ability to mine discriminative information. He et al. (2024a) proposed GazeForensics, which incorporates MSE and applies constraints on general spatial features using 3D gaze features, achieving an accuracy of 0.9942 on the CDF dataset. Qi et al. (2020) introduced remote photoplethysmography (rPPG) into deepfake detection by observing the heart rate differences between real and fake facial videos, enhancing facial detail information with eulerian video magnification. Wu et al. (2024) extracted multi-region rPPG maps and highlighted significant information using local attention, with adjacent features input into a transformer to extract temporal knowledge. Yang et al. (2023) used CPPG signals to provide temporal information, supplementing spatial information through correlations between image pixels reflected by AR coefficients, extracting pixel-level discriminative features for forgery detection. Motion in deepfake videos often contains apparent errors. To capture this anomaly for forgery detection, Saif et al. (2024) constructed spatio-temporal graphs from facial landmarks in both single frames and across frames, using GCNs for detection, which is parameter-efficient and computationally effective. Zhang et al. (2024e) used facial landmarks and face region information as nodes for GCNs, effectively identifying anomalous regions by analyzing both explicit and latent geometric relationships. However, such methods typically require high-quality input videos and are susceptible to environmental factors such as lighting and viewpoint variations. Moreover, as forgery techniques continue to advance, these biometric anomaly cues are highly likely to be eliminated or obfuscated. Table 16 describes these methods from three aspects: biosignal type, backbone, and datasets.

4.4 Identity feature-based

Certain forgery methods can cause identity discrepancies in video subjects. Therefore, using identity-aware frameworks to extract such identity features can help leverage this anomaly information for deepfake detection. Agarwal et al. (2018) enabled the model to learn spatio-temporal

Table 15 A summary of transformer-based methods

Ref.	Year	Network Architecture Design	Input Design	Dataset
Yu et al. (2023)	2023	✖	✔	FF++, DFD, DFDC, DFo, CDF, WDF
Zhao et al. (2023a)	2023	✔	✖	FF++, CDF, DFDC
Choi et al. (2024)	2024	✖	✔	FF++, DFo, CDF, DFD
Liu et al. (2024a)	2024	✖	✔	DFGC, FF++, DFo, CDF, DFD, UADFV
Tian et al. (2024)	2024	✖	✔	FF++, CDF, DFDC
Tu et al. (2024)	2024	✖	✔	FF++
Xu et al. (2024b)	2024	✔	✔	FF++, CDF, DFDC, DFo, WDF, KoDF, DLB
Yue et al. (2024)	2024	✖	✔	FF++, CDF, DFDC, DiffFace, DiffSwap

Table 16 A comparison of biological feature-based deepfake detection methods

Ref.	Year	Biosignal Type	Backbone	Dataset
Yang et al. (2019)	2019	Head poses	SVM	UADFV, DARPA GAN (Guan et al. 2019)
Qi et al. (2020)	2020	rPPG	DNN, GRU	FF++, DFDCp
Demir and Ciftci (2021)	2021	Eye, gaze	DNN	FF++, DF Datasets (Ciftci et al. 2020), CDF, DFo
Haliassos et al. (2021)	2021	Mouth movements	ResNet-18, MS-TCN	FF++, CDF, DFDC
Yang et al. (2023)	2023	CPPG	ACBlock-based DenseNet	FF++, FF, DFDC, CDF, FakeAVCeleb (Kong et al. 2022)
Peng et al. (2024)	2024	Gaze	ResNet-34,ResNet-50, Res2Net-101	FF++, WDF, CDF, DFDCp
He et al. (2024a)	2024	Gaze	ResNet-18	FF++, WDF, CDF
Saif et al. (2024)	2024	Facial landmarks	GCN	FF++, CDF, DFDC
Wu et al. (2024)	2024	rPPG	MLA, Transformer	FF++, CDF
Zhang et al. (2024e)	2024	Facial landmarks, informative regions	GCN	FF++, CDF, WDF, DFDCp, DFD, DFo, ForgeryNIR (Wang et al. 2022a)

biological behavioral features related to identity, distinguishing between different individuals and achieving face-swapping detection. Cozzolino et al. (2021) introduced adversarial training to generate feature vectors consistent with the input individual's identity information, using a temporal ID network as a discriminator for identity recognition. Ramachandran et al. (2021) trained face recognition models with multiple loss functions to extract identity features. To address the identity representation bias in the extracted features, Fang et al. (2024a) designed a bias rectification module and implemented attention-based feature fusion, also utilizing the inconsistency between reference-query images. Additionally, Fang et al. (2024b) proposed a knowledge distillation framework, supervising the identity extractor with region-sensitive spatial features and cross-modality audio's temporal representations to obtain rich spatio-temporal information. Attribute bias can cause errors in the extracted identity features, and Yu et al. (2024a) aligned reference and test images to the same attribute space to extract identity differences, quantifying pixel differences to discern authenticity. However, such methods typically rely heavily on comprehensive identity database support and are mainly tailored for identity-swapping forgery types, exhibiting limited applicability to other forms of forgery. The key ideas of several methods and their AUC scores comparison on the FF++, CDF and DFDCp datasets are shown in Table 17.

5 Challenges, future research directions and conclusions

5.1 Challenges

The rapid advancement of generative AI has led to unprecedented realism in synthetic media, posing significant challenges for detection systems. Below, we outline key challenges that researchers must address in the coming years, expanding beyond the initial scope to include emerging threats and technological gaps.

Detection of large-scale generative models Between 2024 and 2025, powerful generative models such as Videocrafter2, Sora, and Face2Diffusion have achieved near-photorealistic quality in videos. This evolution renders traditional image detection methods increasingly ineffective. 1) Diminishing forensic traces: Newer models produce fewer detectable anomalies (e.g., unnatural textures, inconsistent lighting); 2) Model generalization: Detection systems trained on older generative architectures struggle with newer, more sophisticated models; 3) Real-time processing demands: High-resolution, dynamic content requires faster and more scalable detection frameworks.

Anti-forensics and evasion techniques Adversaries are developing sophisticated methods to bypass detection. 1)

Table 17 The AUC score comparison of identity-based methods on the FF++, CDF and DFDCp datasets

Ref.	Year	Key Idea	FF++	CDF	**DFDCp**
Cozzolino et al. (2021)	2021	Adversarial training	—	0.840	0.910
Fang et al. (2024a)	2024	Identity bias rectification	0.996	0.945	0.983
Fang et al. (2024b)	2024	Multi-modal knowledge distillation	0.958	0.921	0.994
Yu et al. (2024a)	2024	Attribute alignment	0.991	0.911	—

Ensemble Forgery: Combining multiple generation techniques to obscure model-specific fingerprints; 2) Physically-Based Adversarial Patches: Manipulating real-world objects to deceive spatial, frequency, or boundary-based detectors; 3) Prompt Injection Attacks: Exploiting multimodal models to generate inconsistencies that evade cross-modal verification; 4) Post-Processing Obfuscation: Applying noise, compression, or stylization to erase detectable traces while preserving visual plausibility. These evasion strategies severely undermine the robustness and reliability of existing detectors, necessitating more adaptive and resilient approaches.

Generalization across domains and modalities Current detection models often fail when applied to unseen generative techniques or cross-domain data. 1) Overfitting to training datasets: Many detectors perform well only on the specific forgeries they were trained on; 2) Multimodal forgery detection: Deepfakes increasingly incorporate audio, text, and behavioral cues, requiring unified multimodal detection frameworks; 3) Domain shift in real-world deployment: Social media platforms apply varying compression, cropping, and filters, altering forensic signatures.

Real-time and scalable detection As synthetic media spreads rapidly across social networks, news platforms, and live streams, detection systems must evolve to handle. 1) Low-latency processing: Many current methods are too computationally intensive for real-time applications; 2) Lightweight deployment: Mobile and edge-device compatibility remains an open challenge; 3) Adaptation to streaming content: Detecting manipulated live videos (e.g., deepfake video calls) requires frame-level analysis without prior context.

Ethical and adversarial use cases Beyond technical hurdles, emerging ethical and adversarial concerns complicate detection efforts. 1) "White-hat" vs. "black-hat" generative AI: Defensive tools may inadvertently aid malicious actors in refining forgeries; 2) Legal and privacy implications: Detection systems that analyze biometric data raise concerns about surveillance and misuse; 3) Disinformation at scale: AI-generated propaganda and fake news could outpace detection capabilities, requiring automated fact-checking integration.

Proactive defense mechanisms Rather than purely reactive detection, future research should explore preemptive mitigation strategies. 1) Digital watermarking & blockchain-based provenance: Embedding tamper-proof signatures in authentic media; 2) Adversarial training: Training detectors against known evasion techniques to improve robustness.

5.2 Future research directions

The previous sections of this survey provide a comprehensive overview of AI-created visual content detection, including generation technologies, datasets, and related detection methods, along with a detailed classification of detection techniques, offering essential guidance for future researchers. Based on the study of existing challenges, this section will discuss the future directions for AI-created visual content detection.

Regarding AI-generated image detection algorithms, although there has been some development in this area, challenges such as low generalization ability and poor robustness still persist. The fundamental issue in AI-generated image detection is the design of a detector that can effectively identify the differences between real and fake images, while also maintaining strong generalization on unknown generative models. Since the performance of AI-generated image detection algorithms based on deep learning largely depends on the specific generative models in the training datasets, their performance typically drops significantly when tested on samples from different generative models. This necessitates a deeper analysis of the intrinsic relationships between different generated images and improvements in network architectures to learn more effective, generalized features. When images are subjected to certain post-processing attacks (such as scaling, rotation, or JPEG compression), the model's ability to detect general features of generated images also diminishes. In the future, many methods are likely to focus on extracting universal features of generated images from multiple domain-specific features of images. Some methods aim to design detectors that do not require training on fake images, thus avoiding the reliance on specific generative data and offering higher generalization. With the development of image-text models such as CLIP, some research is inclined towards using both images and text for detecting generated images.

For deepfake detection, firstly, existing detection methods have achieved good performance within datasets, but due to the differences in datasets and various forgery techniques, the generalization performance remains insufficient. Therefore, it is necessary to explore methods to improve generalization, such as applying learning strategies like meta-learning and incremental learning, combining data augmentation techniques, or utilizing self-supervised learning to enhance generalization. These are promising directions for future research. Secondly, most current detection technologies are based on a single modality, limited to video and image data. However, many forgery techniques also involve multimodal data such as audio and text. Relying solely on single-modality information may limit detection performance. It is crucial to effectively integrate multimodal knowledge and perform multimodal collaborative learning to fully leverage forgery cues, thereby improving detection performance. Thirdly, with the widespread dissemination of deepfake content across social media, news reports, and live streaming, deepfake detection will move towards real-time

online detection to meet practical demands. However, most current methods focus primarily on improving detection accuracy, with little attention given to model efficiency and lightweight design. This gap remains to be addressed in future research. Finally, in addition to passive detection of forged content, researchers are beginning to focus on active defense methods, such as adding watermarks or noise through preprocessing techniques to make images and videos resistant to forgeries or easily detectable when forged, fundamentally preventing the generation and spread of deepfake content.

5.3 Conclusions

In this review, we provide an overview of existing research on AI-created visual content detection, including AI-generated image detection techniques and deepfake forensics based on deep learning.

For AI-generated image detection, the key approach is to train detectors to explore distinctive feature patterns between real and fake images. In this investigation, we analyze and review the latest techniques for AI-generated image detection based on deep learning. First, we introduce a deep learning-based framework for AI-generated image detection and commonly used datasets. Based on the type of detection features, AI-generated image detection methods can be classified into three categories: spatial-domain-based detection methods, frequency-domain-based detection methods, cross-domain feature fusion detection methods, and image-text-based detection methods. Next, we compare and analyze the state-of-the-art algorithms from three aspects: detection methods, advantages, and limitations. Finally, we address the challenges in current AI-generated image detection algorithms and explore future research directions.

In deepfake detection, we conducted a comprehensive study of existing detection technologies and summarized these methods into four categories based on feature selection: spatial feature-based, spatio-temporal feature-based, biological feature-based, and identity feature-based. Methods based on spatial features focus on mining forgery cues within single-frame images, and they can be further subdivided into space domain-based, frequency domain-based, and multi-domain fusion methods. These approaches generally lack generalization ability and overlook temporal information between video frames, making them unsuitable for detecting dynamic content. Methods based on spatio-temporal features integrate temporal information to enable video-level detection. Depending on the backbone network used, these can be classified into CNN-based, CNN+RNN-based, and transformer-based methods. Among these, transformer-based methods have stronger spatio-temporal modeling capabilities, and most current spatio-temporal methods use transformer networks as the backbone. Both the aforementioned categories are based on general features. While they achieve good detection performance, they lack interpretability. To address this, biological feature-based methods have been introduced. These methods leverage the biological regularities inherent in faces, such as mouth movement, gaze direction, and heart rate, to detect deepfakes. They are more interpretable and easier to understand. Additionally, some forgery techniques alter the identity features of the video subject, prompting researchers to use facial recognition models for identity consistency verification to detect deepfakes. Several improvements have been made in this area, but these methods rely on identity differences, so they are not suitable for detecting forgery techniques that do not change the identity. Finally, based on the existing challenges, a brief analysis of future directions for deepfake detection is provided.

Author Contributions Y.Z., Z.P. and C.W. chose the topic and designed the structure of the paper. Z.P. and C.W. sorted out and analyzed AI-generated images detection techniques. Y.Z. and C.W. reviewed the deepfake detection methods. Y.Z., Z.P. and X.Z. classified the involved algorithms and analyzed the data. C.W., S.H. and X.Z. revised the manuscript. C.W. performed the project administration and supervision. C.W. and X.Z. is funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding This work was supported in part by Natural Science Foundation under Grant ZR2021MF060, in part by the National Key Research and Development Program of China under Grant 2023YFC3321601, in part by the Joint Fund of Shandong Provincial Natural Science Foundation under Grant ZR2021LZH003, in part by the National Natural Science Foundation of China under Grant 61702303, and in part by the 20th Student Research Training Program (SRTP) at Shandong University, Weihai, under Grant A25318, A25315.

Data Availability Data availability is not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

References

Afchar D, Nozick V, Yamagishi J et al (2018) MesoNet: A compact facial video forgery detection network. In: IEEE International workshop on information forensics and security, 10–13 December 2018

Agarwal S, Farid H, El-Gaaly T et al (2018) Detecting deep-fake videos from appearance and behavior. In: IEEE International workshop on information forensics and security, 6–11 December 2020

Amerini I, Caldelli R (2020) Exploiting prediction error inconsistencies through LSTM-based classifiers to detect deepfake videos. In: ACM Workshop on information hiding and multimedia security, 22–24 June 2020

Bai W, Liu Y, Zhang Z et al (2024) Learn from noise: Detecting deepfakes via regional noise consistency. In: International joint conference on neural network, 30 June–5 July 2024

Bammey Q (2024) Synthbuster: towards detection of diffusion model generated images. IEEE Open J Signal Process 5:1–9. https://doi.org/10.1109/OJSP.2023.3337714

Bellemare MG, Danihelka I, Dabney W et al (2018) The cramer distance as a solution to biased wasserstein gradients. In: International conference on learning representations, 30 April–3 May 2018

Bhattacharyya C, Wang H, Zhang F et al (2024) Diffusion deepfake. arXiv:2404.01579. https://doi.org/10.48550/arXiv.2404.01579

Bird JJ, Lotfi A (2024) Cifake: Image classification and explainable identification of AI-generated synthetic images. IEEE Access 12:15642–1565. https://doi.org/10.1109/ACCESS.2024.3356122

Bonettini N, Bestagini P, Milani S et al (2021a) On the use of benford's law to detect GAN-generated images. In: International conference on pattern recognition, 18–21 July 2021

Bonettini N, Cannas E, Mandelli S et al (2021b) Video face manipulation detection through ensemble of CNNs. In: International conference on pattern recognition, 18–21 July 2021

Brock A, Donahue J, Simonyan K (2019) Large scale GAN training for high fidelity natural image synthesis. In: International conference on learning representations, 6–9 May 2019

Cao C, Cai Y, Dong Q et al (2024) LeftRefill: Filling right canvas based on left reference through generalized text-to-image diffusion model. In: IEEE/CVF Conference on computer vision and pattern recognition, 17–21 June 2024

Cazenavette G, Sud A, Leung T et al (2024) FakeInversion: Learning to detect images from unseen text-to-image models by inverting stable diffusion. In: IEEE/CVF Conference on computer vision and pattern recognition, 17–21 June 2024

Chai L, Bau D, Lim S et al (2020) What makes fake images detectable? Understanding properties that generalize. In: European conference on computer vision, 23–28 August 2020

Chandrasegaran K, Tran NT, Cheung NM (2021) A closer look at Fourier spectrum discrepancies for CNN-generated images detection. In: IEEE/CVF Conference on computer vision and pattern recognition, 19–25 June 2021

Chandrasegaran K, Tran NT, Binder A et al (2022) Discovering transferable forensic features for CNN-generated images detection. In: European conference on computer vision, 23–27 October 2022

Chen B, Li T, Ding W (2022a) Detecting deepfake videos based on spatiotemporal attention and convolutional LSTM. Inf Sci 601:58–70. https://doi.org/10.1016/j.ins.2022.04.014

Chen L, Zhang Y, Song Y et al (2022b) Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In: IEEE/CVF Conference on computer vision and pattern recognition, 18–24 June 2022

Chen B, Zeng J, Yang J et al (2024a) DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In: International conference on machine learning, 21–27 July 2024

Chen J, Yao J, Niu L (2024b) A single simple patch is all you need for AI-generated image detection. arXiv:2402.01123. https://doi.org/10.48550/arXiv.2402.01123

Choi Y, Choi M, Kim M et al (2018) StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: IEEE/CVF Conference on computer vision and pattern recognition, 18–22 June 2018

Choi J, Kim T, Jeong Y et al (2024) Exploiting style latent flows for generalizing deepfake video detection. In: IEEE/CVF Conference on computer vision and pattern recognition, 17–21 June 2024

Chollet F (2017) Xception: Deep learning with depthwise separable convolutions. In: IEEE/CVF Conference on computer vision and pattern recognition, 21–26 July 2017

Ciamarra A, Caldelli R, Bimbo AD (2024) Temporal surface frame anomalies for deepfake video detection. In: IEEE/CVF Conference on computer vision and pattern recognition workshops, 17–21 June 2024

Ciftci U, Demir I, Yin L (2020) FakeCatcher: Detection of synthetic portrait videos using biological signals. IEEE Trans Pattern Anal Mach Intell 166:1–1. https://doi.org/10.1109/TPAMI.2020.3009287

Concas S, La Cava S, Casula R et al (2024) Quality-based artifact modeling for facial deepfake detection in videos. In: IEEE/CVF Conference on computer vision and pattern recognition workshops, 17–21 June 2024

Contributing data to deepfake detection research (2019) https://research.google/blog/contributing-data-to-deepfake-detection-research

Corvi R, Cozzolino D, Poggi G, et al. (2023a) Intriguing properties of synthetic images: From generative adversarial networks to diffusion models. In: IEEE/CVF Conference on computer vision and pattern recognition workshops, 18–22 June 2023

Corvi R, Cozzolino D, Zingarini G et al (2023b) On the detection of synthetic images generated by diffusion models. In: IEEE International conference on acoustics, speech and signal processing, 4–10 June 2023

Cozzolino D, Rössler A, Thies J et al (2021) ID-Reveal: Identity-aware deepfake video detection. In: IEEE/CVF International conference on computer vision, 10–17 October 2021

Cozzolino D, Poggi G, Nießner M et al (2024) Zero-shot detection of AI-generated images. In: European conference on computer vision, 29 September–4 October 2024

Deepfakes github (2018). https://github.com/deepfakes/faceswap

Demir I, Ciftci UA (2021) Where do deep fakes look? Synthetic face detection via gaze tracking. In: IEEE/CVF Conference on computer vision and pattern recognition, 25–27 May 2021

Deng J, Lin C, Zhao Z et al (2023) A survey of defenses against AI-generated visual media: Detection, disruption, and authentication. arXiv:2407.10575. https://doi.org/10.48550/arXiv.2407.10575

Dhariwal P, Nichol A (2021) Diffusion models beat GANs on image synthesis. In: Annual conference on neural information processing systems, 6–14 December 2021

Dolhansky B, Howes R, Pflaum B et al (2019) The deepfake detection challenge (DFDC) preview dataset. arXiv preprint arXiv:1910.08854. https://doi.org/10.48550/arXiv.1910.08854

Dolhansky B, Bitton J, Pflaum B et al (2020) The deepfake detection challenge (DFDC) dataset. arXiv:2006.07397. https://doi.org/10.48550/arXiv.2006.07397

Doloriel CT, Cheung NM (2024) Frequency masking for universal deepfake detection. In: IEEE International conference on acoustics, speech and signal processing, 14–19 April 2024

Dong C, Kumar A, Liu E (2022) Think twice before detecting GAN-generated fake images from their spectral domain imprints. In: IEEE/CVF Conference on computer vision and pattern recognition, 18–24 June 2022

Dong F, Zou X, Wang J et al (2023) Contrastive learning-based general deepfake detection with multi-scale RGB frequency clues. J King Saud Univ-Comput Inf Sci 35(4):90–9. https://doi.org/10.1016/j.jksuci.2023.03.005

Dosovitskiy A, Beyer L, Kolesnikov A, et al. (2021) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929

Durall R, Keuper M, Keuper J (2020) Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In: IEEE/CVF Conference on computer vision and pattern recognition, 13–19 June 2020

Dzanic T, Shah K, Witherden F (2020) Fourier spectrum discrepancies in deep network generated images. In: 34th International conference on neural information processing systems, 6–12 December 2020

Edwards P, Nebel JC, Greenhill D et al (2024) A review of deepfake techniques: Architecture, detection, and datasets. IEEE Access 12:154718–154742. https://doi.org/10.1109/ACCESS.2024.3477257

Fang M, Yu L, Song Y et al (2024a) IEIRNet: Inconsistency exploiting based identity rectification for face forgery detection. IEEE Trans Multimed 26:11232–11245. https://doi.org/10.1109/TMM.2024.3453066

Fang M, Yu L, Xie H et al (2024b) STIDNet: Identity-aware face forgery detection with spatiotemporal knowledge distillation. IEEE Trans Comput Soc Syst 11(4):5354–536. https://doi.org/10.1109/TCSS.2024.3356549

Fei J, Dai Y, Yu P et al (2022) Learning second order local anomaly for general face forgery detection. In: IEEE/CVF Conference on computer vision and pattern recognition, 18–24 June 2022

Frank J, Eisenhofer T, Schönherr L et al (2020) Leveraging frequency analysis for deep fake image recognition. In: International conference on machine learning, 12–18 July 2020

Gallagher J, Pugsley W (2024) Development of a dual-input neural model for detecting AI-generated imagery. arXiv:2406.13688. https://doi.org/10.48550/arXiv.2406.13688

Gao J, Micheletto M, Orru G et al (2024a) Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection. Eng Appl Artif Intell 133:108450. https://doi.org/10.1016/j.engappai.2024.108450

Gao J, Xia Z, Marcialis GL et al (2024b) Deepfake detection based on high-frequency enhancement network for highly compressed content. Expert Syst Appl 249(10):12373. https://doi.org/10.1016/j.eswa.2024.123732

Girish S, Suri S, Rambhatla SS et al (2021) Towards discovery and attribution of open-world GAN generated images. In: IEEE/CVF International conference on computer vision, 11–17 October 2021

Golda A, Mekonen K, Pandey A et al (2024) Privacy and security concerns in generative AI: A comprehensive survey. IEEE Access 12:48126–48144. https://doi.org/10.1109/ACCESS.2024.3381611

Gong LY, Li XJ (2024) A contemporary survey on deepfake detection: Datasets, algorithms, and challenges. Electronics 13(3):585. https://doi.org/10.3390/electronics13030585

Goodfellow I, Pouget AJ, Mirza M et al (2014) Generative adversarial nets. In: Annual conference on neural information processing systems, 8–11 December 2014

Gu S, Chen D, Bao J et al (2022) Vector quantized diffusion model for text-to-image synthesis. In: IEEE/CVF Conference on computer vision and pattern recognition, 18–24 June 2022

Guan H, Kozak M, Robertson E et al (2019) MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In: IEEE Winter applications of computer vision workshops, 7–11 January 2019

Guan W, Wang W, Dong J et al (2021) Improving generalization of deepfake detectors by imposing gradient regularization. IEEE Trans Inf Forensic Secur 19:5345–535. https://doi.org/10.1109/TIFS.2024.3396064

Guera D, Delp EJ (2018) Deepfake video detection using recurrent neural networks. In: 15th IEEE International conference on advanced video and signal surveillance, 27–30 November 2018

Guo M, Hu Y, Jiang Z et al (2023) AI-generated image detection: Passive or watermark? arXiv:2411.13553. https://doi.org/10.48550/arXiv.2411.13553

Guo Z, Jia Z, Wang L et al (2024) Constructing new backbone networks via space-frequency interactive convolution for deepfake detection. IEEE Trans Inf Forensic Secur 19:401–41. https://doi.org/10.1109/TIFS.2023.3324739

Haliassos A, Vougioukas K, Petridis S et al (2021) Lips don't lie: A generalisable and robust approach to face forgery detection. In: IEEE/CVF Conference on computer vision and pattern recognition, 20–25 June 2021

Hasanaath AA, Luqman H, Katib R et al (2024) FSBI: Deepfakes detection with frequency enhanced self-blended images. arXiv preprint arXiv:2406.08625. https://doi.org/10.48550/arXiv.2406.08625

He P, Li H, Wang H (2019a) Detection of fake images via the ensemble of deep representations from multi color spaces. In: IEEE International conference on image processing, 22–25 September 2019

He Z, Zuo W, Kan M et al (2019b) AttGAN: Facial attribute editing by only changing what you want. IEEE Trans Img Proc 28(11):5464–5478. https://doi.org/10.1109/TIP.2019.2916751

He Y, Yu N, Keuper M et al (2021) Beyond the spectrum: Detecting deepfakes via re-synthesis. In: International joint conference on artificial intelligence, 21–26 August 2021

He Q, Peng C, Liu D et al (2024a) GazeForensics: Deepfake detection via gaze-guided spatial inconsistency learning. Neural Netw 180:10663. https://doi.org/10.1016/j.neunet.2024.106636

He Z, Chen P, Ho TY (2024b) RIGID: A training-free and model-agnostic framework for robust AI-generated image detection. arXiv:2405.20112. https://doi.org/10.48550/arXiv.2405.20112

Heidari A, Navimipour NJ, Dag H et al (2024) Deepfake detection using deep learning methods: A systematic and comprehensive review. Wiley Interdiscip Rev-Data Mining Knowl Discov 14(2):e152. https://doi.org/10.3390/electronics13030585

Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. In: Annual conference on neural information processing systems, 6–12 December 2020

Ho S (2023) From development to dissemination: Social and ethical issues with text-to-image AI-generated art. In: Canadian conference on artificial intelligence, 5–9 June 2023

Hoe JT, Jiang X, Chan CS et al (2024) InteractDiffusion: Interaction control in text-to-image diffusion models. In: IEEE/CVF Conference on computer vision and pattern recognition, 17–21 June 2024

Hong Y, Zhang J (2024) Wildfake: A large-scale challenging dataset for AI-generated images detection. arXiv:2402.11843. https://doi.org/10.48550/arXiv.2402.11843

Hsu CC, Lee CY, Zhuang YX (2018) Learning to detect fake face images in the wild. In: International symposium on computer, consumer and control, 6–8 December 2018

Hu B, Wang J (2020) Deep learning for distinguishing computer generated images and natural images: A survey. J Inf Hiding Privacy Protection 2(2):37–47. https://doi.org/10.32604/jihpp.2020.010464

Huang D, Zhang Y (2024) Learning meta model for strong generalization deepfake detection. In: International joint conference on neural networks, 30 June–5 July 2024

Höllein L, Božič A, Müller N et al (2024) ViewDiff: 3D-consistent image generation with text-to-image models. In: IEEE/CVF Conference on computer vision and pattern recognition, 17–21 June 2024

Jayashre K, Amsaprabhaa M (2024) Safeguarding media integrity: A hybrid optimized deep feature fusion based deepfake detection in

videos. Comput Secur 142:103860. https://doi.org/10.1016/j.cose.2024.103860
Jeon H, Bang Y, Kim J et al (2020) T-GD: Transferable GAN-generated images detection framework. In: International conference on machine learning, 12–18 July 2020
Jeong Y, Kim D, Ro Y et al (2022) Fingerprintnet: Synthesized fingerprints for generated image detection. In: European conference on computer vision, 23–27 October 2022
Jia S, Li X, Lyu S (2023) Model attribution of face-swap deepfake videos. In: IEEE International conference on image processing, 16–19 October 2022
Jiang L, Li R, Wu W et al (2020) Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In: IEEE/CVF Conference on computer vision and pattern recognition, 13–19 June 2020
Ju Y, Jia S, Ke L et al (2022) Fusing global and local features for generalized AI-synthesized image detection. In: IEEE International conference on image processing, 16–19 October 2022
Ju Y, Jia S, Cai J et al (2024) GLFF: Global and local feature fusion for AI-synthesized image detection. IEEE Trans Multimed 26:4073–4085. https://doi.org/10.1109/TMM.2023.3313503
Karras T, Aila T, Laine S et al (2018) Progressive growing of GANs for improved quality, stability, and variation. In: International conference on learning representations, 30 April–3 May 2018
Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: IEEE/CVF Conference on computer vision and pattern recognition, 16–20 June 2019
Karras T, Laine S, Aittala M et al (2020) Analyzing and improving the image quality of styleGAN. In: IEEE/CVF Conference on computer vision and pattern recognition, 13–19 June 2020
Kaur A, Hoshyar A, Saikrishna V et al (2024) Deepfake video detection: Challenges and opportunities. Artif Intell Rev 57(6):159. https://doi.org/10.1007/s10462-024-10810-6
Keita M, Hamidouche W, Bougueffa H et al (2024a) Harnessing the power of large vision language models for synthetic image detection. arXiv:2404.02726. https://doi.org/10.48550/arXiv.2404.02726
Keita M, Hamidouche W, Eutamene HB et al (2024b) Bi-LORA: A vision-language approach for synthetic image detection. arXiv:2404.01959. https://doi.org/10.48550/arXiv.2404.01959
Kertysova K (2018) Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. Security and Human Rights 29:55–8. https://doi.org/10.1163/18750230-02901005
Khan SA, Dang-Nguyen DT (2024) CLIPping the deception: Adapting vision-language models for universal deepfake detection. In: International conference on multimedia retrieval, 10–14 June 2024
Kong C, Chen B, Yang W et al (2022) Appearance matters, so does audio: Revealing the hidden face via cross-modality transfer. IEEE Trans Circuits Syst Video Technol 32(1):423–43. https://doi.org/10.1109/TCSVT.2021.3057457
Korshunov P, Marcel S (2018) Deepfakes: A new threat to face recognition? Assessment and detection. arXiv:1812.08685. https://doi.org/10.48550/arXiv.1812.08685
Kwon P, You J, Nam G et al (2021) Kodf: A large-scale korean deepfake detection dataset. In: IEEE/CVF International conference on computer vision, 10–17 October 2021
Lanzino R, Fontana F, Diko A et al (2024) Faster than lies: Real-time deepfake detection using binary neural networks. In: IEEE/CVF Conference on computer vision and pattern recognition, 17–21 June 2024
Le M, Woo S (2022) ADD: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images. In: AAAI Cconference on artificial intelligence, 22 February–1 March, 2022
Lee KS, Tran NT, Cheung NM (2021) Infomax-GAN: Improved adversarial image generation via information maximization and contrastive learning. In: IEEE/CVF Winter conference on applications of computer vision, 5–9 January 2021
Leporoni G, Maiano L, Papa L et al (2024) A guided-based approach for deepfake detection: RGB-depth integration via features fusion. Pattern Recognit Lett 112:99–10. https://doi.org/10.1016/j.patrec.2024.03.025
Li Y, Chang MC, Lyu S (2018) In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking. In: IEEE International workshop on information forensics and security, 10–13 December 2018
Li L, Bao J, Zhang T et al (2020a) Face X-ray for more general face forgery detection. In: IEEE/CVF Conference on computer vision and pattern recognition, 13–19 June 2020
Li W, He P, Li H et al (2020b) Detection of GAN-generated images by estimating artifact similarity. IEEE Signal Process Lett 21:862–86. https://doi.org/10.1109/LSP.2021.3130525
Li Y, Yang X, Sun P et al (2020c) Celeb-DF: A large-scale challenging dataset for deepfake forensics. In: IEEE/CVF Conference on computer vision and pattern recognition, 13–19 June 2020
Li J, Xie H, Li J et al (2021) Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In: IEEE/CVF Conference on computer vision and pattern recognition, 20–25 June 2021
Li Y, Zhang Y, Yang H et al (2024) SA^{3}WT: Adaptive wavelet-based transformer with self-paced auto augmentation for face forgery detection. Int J Comput Vis 132(10):4417–443. https://doi.org/10.1007/s11263-024-02091-x
Liang Z, Wang R, Liu W et al(2024) Let real images be as a judger, spotting fake images synthesized with generative models. arXiv:2403.16513. https://doi.org/10.48550/arXiv.2403.16513
Lim Y, Lee C, Kim A et al (2024) DistilDIRE: A small, fast, cheap and lightweight diffusion synthesized deepfake detection. In: International conference on machine learning, 21–27 July 2024
Lin M, Shang L, Gao X (2023) Enhancing interpretability in AI-generated image detection with genetic programming. In: IEEE International conference on data mining workshops, 1–4 December 2023
Lin L, Gupta N, Zhang Y et al (2024a) Detecting multimedia generated by large AI models: A survey. arXiv:2402.00045. https://doi.org/10.48550/arXiv.2402.00045
Lin Y, Song W, Li B et al (2024b) Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection. In: European conference on computer vision, 29 September–4 October 2024
Liu Z, Qi X, Torr PHS (2020) Global texture enhancement for fake face detection in the wild. In: IEEE/CVF Conference on computer vision and pattern recognition, 13–19 June 2020
Liu Z, Lin Y, Cao Y et al (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE/CVF International conference on computer vision, 10–17 October 2021
Liu B, Yang F, Bi X et al (2022) Detecting generated images by real images. In: European conference on computer vision, 11–17 October 2021
Liu Z, Wang H, Wang S (2023) Cross-domain local characteristic enhanced deepfake video detection. In: 16th Asian conference on computer vision, 4–8 December 2023
Liu B, Liu B, Ding M et al (2024a) MeST-Former: Motion-enhanced spatiotemporal transformer for generalizable deepfake detection. Neurocomputing 610:12858. https://doi.org/10.1016/j.neucom.2024.128588
Liu C, Zhu T, Zhao Y et al (2024b) Disentangling different levels of GAN fingerprints for task-specific forensics. Comput Stand Interfaces 89:10382. https://doi.org/10.1016/j.csi.2023.103825
Liu H, Tan Z, Tan C et al (2024c) Forgery-aware adaptive transformer for generalizable synthetic image detection. In: IEEE/CVF Con-

ference on computer vision and pattern recognition, 17–21 June 2024
Lorenz P, Durall R, Keuper J (2024) Detecting images generated by deep diffusion models using their local intrinsic dimensionality. In: IEEE/CVF International conference on computer vision, 2–6 October 2024
Lu Z, Huang D, Bai L et al (2023) Seeing is not always believing: Benchmarking human and model perception of AI-generated images. In: Annual conference on neural information processing systems, 10–16 December 2023
Lu L, Wang Y, Zhuo W et al (2024a) Deepfake detection via separable self-consistency learning. In: IEEE International conference on image processing, 27–30 October 2024
Lu W, Liu L, Zhang B et al (2024b) Detection of deepfake videos using long-distance attention. IEEE Trans Neural Netw Learn Syst 35(7):9366–937. https://doi.org/10.1109/TNNLS.2022.3233063
Luo Y, Zhang Y, Yan J et al (2021) Generalizing face forgery detection with high-frequency features. In: IEEE/CVF Conference on computer vision and pattern recognition, 20–25 June 2021
Luo Y, Du J, Yan K et al (2024) LaRE^2: Latent reconstruction error based method for diffusion-generated image detection. In: IEEE/CVF Conference on computer vision and pattern recognition, 17–21 June 2024
Ma X, Tian J, Cai Y et al (2024) HIDD: Human-perception-centric incremental deepfake detection. In: International joint conference on neural networks, 30 June–5 July 2024
Mandelli S, Bonettini N, Bestagini P et al (2022) Detecting GAN-generated images by orthogonal training of multiple CNNs. In: IEEE International conference on image processing, 16–19 October 2022
Marra F, Gragnaniello D, Cozzolino D et al (2018) Detection of GAN-generated fake images over social networks. In: IEEE Conference on multimedia information processing and retrieval, 10–12 April 2018
Marra F, Gragnaniello D, Verdoliva L et al (2019a) Do GANs leave artificial fingerprints? In: IEEE Conference on multimedia information processing and retrieval, 28–30 March 2019
Marra F, Saltori C, Boato G et al (2019b) Incremental learning for the detection and classification of GAN-generated images. In: IEEE International workshop on information forensics and security, 9–12 December 2019
Masi I, Killekar A, Mascarenhas RM et al (2020) Two-branch recurrent network for isolating deepfakes in videos. In: European conference on computer vision, 23–28 August 2020
Meng Z, Peng B, Dong J et al (2024) Artifact feature purification for cross-domain detection of AI-generated images. Comput Vis Image Underst 247:10407. https://doi.org/10.1016/j.cviu.2024.104078
Mi Z, Jiang X, Sun T et al (2020) GAN-generated image detection with self-attention mechanism against GAN generator defect. IEEE J Sel Top Signal Process 14(5):969–98. https://doi.org/10.1109/JSTSP.2020.2994523
Miao C, Tan Z, Chu Q et al (2023) F2Trans: High-frequency fine-grained transformer for face forgery detection. IEEE Trans Inf Forensic Secur 18:1039–105. https://doi.org/10.1109/TIFS.2022.3233774
Midjourney (2022) https://www.midjourney.com/home
Narayan K, Agarwal H, Thakral K et al (2023) DF-Platter: Multi-face heterogeneous deepfake dataset. In: IEEE/CVF Conference on computer vision and pattern recognition, 18–22 June 2023
Nichol A, Dhariwal P, Ramesh A et al (2022) GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In: International conference on machine learning, 17–23 July 2022
Ojha U, Li Y, Lee YJ (2023) Towards universal fake image detectors that generalize across generative models. In: IEEE/CVF Conference on computer vision and pattern recognition, 17–24 June 2023
Pang G, Zhang B, Teng Z et al (2023) MRE-Net: Multi-rate excitation network for deepfake video detection. IEEE Trans Circuits Syst Video Technol 33(8):3663–367. https://doi.org/10.1109/TCSVT.2023.3239607
Park T, Liu MY, Wang TC et al (2019) Semantic image synthesis with spatially-adaptive normalization. In: IEEE/CVF Conference on computer vision and pattern recognition, 16–20 June 2019
Peng C, Miao Z, Liu D et al (2024) Where deepfakes gaze at? Spatial–temporal gaze inconsistency analysis for video face forgery detection. IEEE Trans Inf Forensic Secur 19:4507–4517. https://doi.org/10.1109/TIFS.2024.3381823
Peng S, Cai M, Ma R et al (2023) Deepfake detection algorithm for high-frequency components of shallow features. Laser Optoelectron Prog 60(10):101500. https://doi.org/10.3788/LOP213318
Pontorno O, Guarnera L, Battiato S (2024) On the exploitation of DCT-traces in the generative-AI domain. In: IEEE International conference on image processing, 27–30 October 2024
Qi H, Guo Q, Juefei-Xu F et al (2020) DeepRhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In: 18th ACM International conference on multimedia, 12–26 October 2020
Qian Y, Yin G, Sheng L et al (2020) Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: European conference on computer vision, 23–28 August 2020
Qiao T, Chen Y, Zhou X et al (2024) CSC-Net: Cross-color spatial co-occurrence matrix network for detecting synthesized fake images. IEEE Trans Cognit Dev Syst 16(1):369–37. https://doi.org/10.1109/TCDS.2023.3274450
Radford A, Kim J, Hallacy C et al (2023) Learning transferable visual models from natural language supervision. In: International conference on machine learning, 18–24 July 2021
Rahman MA, Paul B, Sarker NH et al (2023) Artifact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection. In: IEEE International conference on image processing, 8–11 October 2023
Ramachandran S, Nadimpalli A, Rattani A (2021) An experimental evaluation on deepfake detection using deep face recognition. In: International carnahan conference on security technology, 11–15 October 2021
Ramesh A, Pavlov M, Goh G et al (2021) Zero-shot text-to-image generation. In: International conference on machine learning, 18–24 July 2021
Rana MS, Nobi MN, Murali B et al (2022) Deepfake detection: A systematic literature review. IEEE Access 10:25494–25513. https://doi.org/10.1109/ACCESS.2022.3154404
Rombach R, Blattmann A, Lorenz D et al (2022) High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on computer vision and pattern recognition, 18–24 June 2022
Rossler A, Cozzolino D, Verdoliva L et al (2018) FaceForensics++: Learning to detect manipulated facial images. In: IEEE/CVF International conference on computer vision, 27 October–2 November 2018
Saharia C, Chan W, Saxena S et al (2022) Photorealistic text-to-image diffusion models with deep language understanding. In: Annual conference on neural information processing systems, 28 November–9 December 2022
Saif S, Tehseen S, Ali SS (2024) Fake news or real? Detecting deepfake videos using geometric facial structure and graph neural network. Technol Forecast Soc Chang 205:12347. https://doi.org/10.1016/j.techfore.2024.123471
Saikia P, Dholaria D, Yadav P et al (2022) A hybrid CNN-LSTM model for video deepfake detection by leveraging optical flow features. In: International joint conference on neural network, 27–30 November 2018

Sandotra N, Arora B (2024) A comprehensive evaluation of feature-based AI techniques for deepfake detection. Neural Comput Appl 36(8):3859–3887. https://doi.org/10.1007/s00521-023-09288-0

Sarkar A, Mai H, Mahapatra A et al (2024) Shadows don't lie and lines can't bend! Generative models don't know projective geometry . . . for now. In: IEEE/CVF Conference on computer vision and pattern recognition, 17–21 June 2024

Seow JW, Lim MK, Phan RCW et al (2022) A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities. Neurocomputing 513:351–371. https://doi.org/10.1016/j.neucom.2022.09.135

Sha Z, Li Z, Yu N et al (2024a) DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models. In: ACM SIGSAC Conference on computer and communications security, 14–18 October 2024

Sha Z, Tan Y, Li M et al (2024b) ZeroFake: Zero-shot detection of fake images generated and edited by text-to-image generation models. In: ACM SIGSAC conference on computer and communications security, 14–18 October 2024

Shiohara K, Yamasaki T (2022) Detecting deepfakes with self-blended images. In: IEEE/CVF Conference on computer vision and pattern recognition, 18–24 June 2022

Shiohara K, Yamasaki T (2024) Face2Diffusion for fast and editable face personalization. In: IEEE/CVF Conference on computer vision and pattern recognition, 17–21 June 2024

Shirakawa T, Uchida S (2024) NoiseCollage: A layout-aware text-to-image diffusion model based on noise cropping and merging. In: IEEE/CVF Conference on computer vision and pattern recognition, 17–21 June 2024

Sinitsa S, Fried O (2024) Deep image fingerprint: Towards low budget synthetic image detection and model lineage analysis. In: IEEE/CVF Winter conference on applications of computer vision, 3–8 January 2024

Tan C, Zhao Y, Wei S et al (2023) Learning on gradients: Generalized artifacts representation for GAN-generated images detection. In: IEEE/CVF Conference on computer vision and pattern recognition, 17–24 June 2023

Tan C, Liu P, Tao R et al (2024a) Data-independent operator: A training-free artifact representation extractor for generalizable deepfake detection. arXiv:2403.06803. https://doi.org/10.48550/arXiv.2403.06803

Tan C, Zhao Y, Wei S et al (2024b) Frequency-aware deepfake detection: Improving generalizability through frequency space learning. In: AAAI Conference on artifical intelligence, 20–27 February 2024

Tan C, Zhao Y, Wei S et al (2024c) Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. In: IEEE/CVF Conference on computer vision and pattern recognition, 17–21 June 2024

Thies J, Zollhöfer M, Stamminger M et al (2016) Face2Face: Real-time face capture and reenactment of RGB videos. In: IEEE/CVF Conference on computer vision and pattern recognition, 27–30 June 2016

Tian K, Chen C, Zhou Y et al (2024) Illumination enlightened spatial-temporal inconsistency for deepfake video detection. In: IEEE International conference on multimedia and expo, 15–19 July 2024

Trung-Nghia L, Nguyen HH, Yamagishi J et al (2021) Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In: IEEE/CVF International conference on computer vision, 10–17 October 2021

Tu Y, Wu J, Lu L et al (2024) Face forgery video detection based on expression key sequences. J King Saud Univ-Comput Inf Sci 36(7):102142. https://doi.org/10.1016/j.jksuci.2024.102142

Uhlenbrock L, Cozzolino D, Moussa D et al (2024) Did you note my palette? Unveiling synthetic images through color statistics. In: ACM Workshop on information hiding and multimedia security, 24–26 June 2024

Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. In: Annual conference on neural information processing systems, 4–9 December 2017

Wang SY, Wang O, Zhang R et al (2020) CNN-generated images are surprisingly easy to spot...for now. In: IEEE/CVF Conference on computer vision and pattern recognition, 13–19 June 2020

Wang Y, Peng C, Liu D et al (2022a) ForgeryNIR: Deep face forgery and detection in near-infrared scenario. IEEE Trans Inf Forensic Secur 17:500–51. https://doi.org/10.1109/TIFS.2022.3146766

Wang ZJ, Montoya E, Munechika D et al (2022b) DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. arXiv:2210.14896. https://doi.org/10.48550/arXiv.2210.14896

Wang B, Wu X, Tang Y et al (2023a) Frequency domain filtered residual network for deepfake detection. Mathematics 11(4):816. https://doi.org/10.3390/math11040816

Wang Y, Peng C, Liu D et al (2023b) Spatial-temporal frequency forgery clue for video forgery detection in VIS and NIR scenario. IEEE Trans Circuits Syst Video Technol 33(12):7943–7956. https://doi.org/10.1109/TCSVT.2023.3281475

Wang Z, Bao J, Zhou W et al (2023c) Dire for diffusion-generated image detection. In: IEEE/CVF International conference on computer vision, 1–6 October 2023

Wang B, Wu X, Wang F et al (2024a) Spatial-frequency feature fusion based deepfake detection through knowledge distillation. Eng Appl Artif Intell 133:10834. https://doi.org/10.1016/j.engappai.2024.108341

Wang F, Chen Q, Jing B et al (2024b) Deepfake detection based on the adaptive fusion of spatial-frequency features. Int J Intell Syst 2024:757803. https://doi.org/10.1155/2024/7578036

Wang Z, Sehwag V, Chen C et al (2024c) How to trace latent generative model generated images without artificial watermark? In: International conference on machine learning, 21–27 July 2024

Weng J (2024) Local frequency analysis for diffusion-generated image detection. In: International conference on image processing and artificial intelligence, 19–21 April 2024

Wißmann A, Zeiler S, Nickel R et al (2024) Whodunit: Detection and attribution of synthetic images by leveraging model-specific fingerprints. In: ACM International workshop on multimedia AI against disinformation, 10–13 June 2024

Wu PW, Lin YJ, Chang CH et al (2019) RelGAN: Multi-domain image-to-image translation via relative attributes. In: IEEE/CVF International conference on computer vision, 6–9 May 2019

Wu H, Zhou J, Zhang S (2023a) Generalizable synthetic image detection via language-guided contrastive learning. arXiv:2305.13800. https://doi.org/10.48550/arXiv.2305.13800

Wu J, Zhang B, Li Z et al (2023b) Interactive two-stream network across modalities for deepfake detection. IEEE Trans Circuits Syst Video Technol 33(11):6418–643. https://doi.org/10.1109/TCSVT.2023.3269841

Wu J, Zhu Y, Jiang X et al (2024) Local attention and long-distance interaction of rPPG for deepfake detection. Visual Comput 40(2):1083–1094. https://doi.org/10.1007/s00371-023-02833-x

Wukong (2023) https://xihe.mindspore.cn/modelzoo/wukong

Xu Q, Wang H, Meng L et al (2023) Exposing fake images generated by text-to-image diffusion models. Pattern Recognit Lett 176:76–82. https://doi.org/10.1016/j.patrec.2023.10.021

Xu Q, Jiang X, Sun T et al (2024a) Detecting artificial intelligence-generated images via deep trace representations and interactive feature fusion. Inf Fusion 112:10257. https://doi.org/10.1016/j.inffus.2024.10257

Xu Y, Liang J, Sheng L et al (2024b) Learning spatiotemporal inconsistency via thumbnail layout for face deepfake detection. Int J Comput Vis 132:5663–568. https://doi.org/10.1007/s11263-024-02054-2

Yadav A, Vishwakarma DK (2024) Datasets, clues and state-of-the-arts for multimedia forensics: An extensive review. Expert Syst Appl 249:123756. https://doi.org/10.1016/j.eswa.2024.123756

Yan S, Li O, Cai J, et al. (2024a) A sanity check for AI-generated image detection. arXiv:2406.19435. https://doi.org/10.48550/arXiv.2406.19435

Yan Z, Luo Y, Lyu S et al (2024b) Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In: IEEE/CVF Conference on computer vision and pattern recognition, 17–21 June 2024

Yang X, Li Y, Lyu S (2019) Exposing deep fakes using inconsistent head poses. In: IEEE International conference on acoustics, speech and signal processing, 12–17 May 201

Yang J, Li A, Xiao S et al (2021) MTD-Net: Learning to detect deepfakes images by multi-scale texture difference. IEEE Trans Inf Forensic Secur 16:4234–424. https://doi.org/10.1109/TIFS.2021.3102487

Yang J, Sun Y, Mao M et al (2023) Model-agnostic method: Exposing deepfake using pixel-wise spatial and temporal fingerprints. IEEE Trans Big Data 9(6):1496–150. https://doi.org/10.1109/TBDATA.2023.3284272

Yang Y, Qian Z, Zhu Y, et al. (2024) Scaling up deepfake detection by learning from discrepancy. arXiv:2404.04584. https://doi.org/10.48550/arXiv.2404.04584

Yu N, Davis LS, Fritz M (2019) Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In: IEEE/CVF International conference on computer vision, 27 October–2 November 2019

Yu Y, Ni R, Li W et al (2022) Detection of AI-manipulated fake faces via mining generalized features. ACM Trans Multimed Comput Commun Appl 18:94. https://doi.org/10.1145/3499026

Yu Y, Ni R, Zhao Y et al (2023) MSVT: Multiple spatiotemporal views transformer for deepfake video detection. IEEE Trans Circuits Syst Video Technol 33(9):4462–447. https://doi.org/10.1109/TCSVT.2023.3281448

Yu C, Zhang X, Duan Y et al (2024a) Diff-ID: An explainable identity difference quantification framework for deepfake detection. IEEE Trans Dependable Secur Comput 21(5):5029–5045. https://doi.org/10.1109/TDSC.2024.3364679

Yu Y, Ni R, Yang S et al (2024b) Mining generalized multi-timescale inconsistency for detecting deepfake videos. Int J Comput Vis 2024:1–1. https://doi.org/10.1007/s11263-024-02091-x

Yue P, Chen B, Fu Z (2024) Local region frequency guided dynamic inconsistency network for deepfake video detection. Big Data Min Anal 7(3):889–904. https://doi.org/10.26599/BDMA.2024.9020030

Zhang X, Karaman S, Chang SF (2019) Detecting and simulating artifacts in GAN fake images. In: IEEE International workshop on information forensics and security, 9–12 December 2019

Zhang M, Wang H, He P et al (2022) Improving GAN-generated image detection generalization using unsupervised domain adaptation. In: IEEE International conference on multimedia and expo, 18–22 July 2022

Zhang Y, Xu X (2023) Diffusion noise feature: Accurate and fast generated image detection. arXiv:2312.02625. https://doi.org/10.48550/arXiv.2312.02625

Zhang D, Chen J, Liao X et al (2024a) Face forgery detection via multi-feature fusion and local enhancement. IEEE Trans Circuits Syst Video Technol 34(9):8972–8977. https://doi.org/10.1109/TCSVT.2024.3390945

Zhang D, Xiao Z, Li S et al (2024b) Learning natural consistency representation for face forgery video detection. In: European conference on computer vision, 29 September–4 October 2024

Zhang L, Chen H, Hu S et al (2024c) X-Transfer: A transfer learning-based framework for GAN-generated fake image detection. In: International joint conference on nneural networks, 30 June–5 July 2024

Zhang R, He P, Li H et al (2024d) Temporal diversified self-contrastive learning for generalized face forgery detection. IEEE Trans Circuits Syst Video Technol 34(12):12782–12795. https://doi.org/10.1109/TCSVT.2024.3436554

Zhang R, Wang H, Liu H et al (2024e) Generalized face forgery detection with self-supervised face geometry information analysis network. Appl Soft Comput 166:11214. https://doi.org/10.1016/j.asoc.2024.112143

Zhang Y, Zhu N, Zhang X et al (2024f) Computer-generated image detection based on deep LBP network. In: International conference on computer application and information security, 20–22 December 2024

Zhao C, Wang C, Hu G et al (2023a) ISTVT: Interpretable spatial-temporal video transformer for deepfake detection. IEEE Trans Inf Forensic Secur 18:1335–134. https://doi.org/10.1109/TIFS.2023.3239223

Zhao Y, Li J, Wang L (2023b) Harmonizing dynamic frequency analysis with attention mechanisms for efficient facial image authenticity detection. In: International conference on computer science and automation technology, 6–8 October 2023

Zheng J, Zhou Y, Hu X et al (2024) Deepfake detection with combined unsupervised-supervised contrastive learning. In: IEEE International conference on image processing, 27–30 October 2024

Zhong N, Xu Y, Qian Z et al (2023) Rich and poor texture contrast: A simple yet effective approach for AI-generated image detection. arXiv:2311.12397. https://doi.org/10.48550/arXiv.2311.12397

Zhou T, Wang W, Liang Z et al (2021) Face forensics in the wild. In: IEEE/CVF Conference on computer vision and pattern recognition, 20–25 June 2021

Zhou D, Li Y, Ma F et al (2024a) MIGC: Multi-instance generation controller for text-to-image synthesis. In: IEEE/CVF Conference on computer vision and pattern recognition, 17–21 June 2024

Zhou J, Zhao X, Xu Q et al (2024b) MDCF-Net: Multi-scale dual-branch network for compressed face forgery detection. IEEE Access 12:58740–5874. https://doi.org/10.1109/ACCESS.2024.3390217

Zhu JY, Park T, Isola P et al (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE/CVF International conference on computer vision, 22–29 October 2017

Zhu M, Chen H, Yan Q et al (2023) Genimage: A million-scale benchmark for detecting AI-generated image. In: Annual conference on neural information processing systems, 10–16 December 2023

Zhu Y, Zhang C, Gao J et al (2024) High-compressed deepfake video detection with contrastive spatiotemporal distillation. Neurocomputing 565:126872. https://doi.org/10.1016/j.neucom.2023.126872

Zhuang YX, Hsu CC (2019) Detecting generated image based on a coupled network with two-step pairwise learning. In: IEEE International conference on image processing, 22–25 September 2019

Zi B, Chang M, Chen J et al (2020) WildDeepfake: A challenging real-world dataset for deepfake detection. In: 18th ACM International conference on multimedia, 12–26 October 2020