中图法分类号:TP18;TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2026)01-0013-32

论文引用格式:Li M L, Qian Z X and Zhang X P. 2026. Survey on artificial intelligence-generated image detection. Journal of Image and Graphics, 31(1):0013-0044(李美玲, 钱振兴, 张新鹏. 2026. 人工智能生成图像检测技术综述. 中国图象图形学报, 31(1):0013-0044)[DOI:10.11834/jig.250053]

人工智能生成图像检测技术综述

李美玲,钱振兴*,张新鹏

复旦大学计算与智能创新学院,上海　200438

摘　要:人工智能生成图像检测旨在判断图像是否由生成模型生成,是人工智能安全与治理领域一个重要研究方向。然而当前生成图像检测方法因生成模型结构的多样性、生成图像的复杂性以及对生成图像后处理操作的不确定性难以实现高效和鲁棒检测。早期的检测方法重点关注对生成对抗网络生成内容的检测。近年来,扩散模型生成图像的检测受到广泛关注,与之相关的检测方法涌现,并表现出优越性能。因此,首先对近年来的主流图像生成模型进行梳理,然后从监督范式、学习方式、检测依据、骨干网络、技术手段和可解释性多种维度对现有生成图像检测方法进行分类,并以检测依据(像素域特征、频域特征、预训练模型特征、融合特征和特定规则)作为主要划分标准,对各类研究工作的基本思想与特点进行详细阐述与综合分析。此外,列举了当前用于通用生成图像检测的基准数据集,从数据集结构和规模等方面进行综合比较,并对面向检测方法的综合评测基准进行汇总。关于评估维度,从域内准确性、域外泛化性和鲁棒性3个层面进行介绍。之后,对代表性检测方法进行横向比较,就检测结果进行分析。最后,对当前生成图像检测领域待解决的问题进行总结,并对未来的研究方向进行展望。

关键词:人工智能安全;人工智能生成图像(AIGI)检测;数字取证;图像取证;深度伪造检测;图像生成模型;深度学习

Survey on artificial intelligence-generated image detection

Li Meiling, Qian Zhenxing*, Zhang Xinpeng

College of Computer Science and Artificial Intelligence, Fudan University, Shanghai　200438, China

Abstract: In the era of rapid advancements in generative artificial intelligence, the explosive growth of artificial intelligence-generated content (AIGC), propelled by the widespread use of models such as ChatGPT and diffusion models, has given rise to significant challenges. The proliferation of false information, fraudulent content, and inappropriate remarks has severely undermined the authenticity and credibility of information dissemination, highlighting the urgent need for effective artificial intelligence-generated images (AIGI) detection. The detection of AIGI, which aims to determine whether an image is generated by a generative model, is a crucial research direction in the field of AI security and governance. Early detection methods focused on detecting the content generated by generative adversarial networks. In recent years, images generated by diffusion models have received extensive attention, and related detection methods have emerged, demonstrating superior performance. However, current detection methods still encounter difficulties in achieving high-generalization and robust detection performance due to the diversity of generative model structures, the complexity of generated images, and the uncertainty of post-processing operations on generated images. This comprehensive review aims to provide a thorough understanding of the research landscape in AIGI detection. The paper first commences with an in-depth exploration of

收稿日期:2025-02-21;**修回日期**:2025-06-03;**预印本日期**:2025-06-10

*通信作者:钱振兴　zxqian@fudan.edu.cn

基金项目:国家自然科学基金项目(62572125);上海市自然科学基金项目(25ZR1401019)

Supported by:National Natural Science Foundation of China(62572125);Shanghai Municipal Natural Science Foundation(25ZR1401019)

mainstream image-generation models. Generative adversarial networks, with their adversarial training mechanism between generators and discriminators, have evolved through various architectures such as conditional GAN (CGAN), CycleGAN, and StyleGAN, each enhancing the controllability and quality of generated images. Autoregressive models, inspired by the success of GPT-series in natural language processing, such as Image GPT and DALL-E, leverage powerful attention mechanisms for image generation. Based on the principles of diffusion and reverse diffusion processes, diffusion models have demonstrated remarkable capabilities in generating high-quality images, leading to the emergence of numerous models such as Stable Diffusion and its variants in recent years. Understanding these models is fundamental because their unique characteristics leave identifiable patterns in the generated images, which detection methods aim to identify. Subsequently, existing AIGI detection methods are systematically categorized from multiple perspectives, including supervision paradigm, learning methods, detection basis, backbone network, technical means, and interpretability. In terms of the supervision paradigm, most methods are based on supervised detection, with some incorporating additional contrastive learning losses or regularization terms to improve performance. Though less common, semi-supervised and unsupervised detection methods also contribute to the field. Regarding learning methods, eager learning-based approaches, especially those that using binary classifiers, are dominant, while lazy learning-based methods, such as those relying on the K-nearest neighbors algorithm, offer an alternative without requiring extensive training. When categorized by detection basis, methods based on pixel-domain features use characteristics such as texture, color, and reconstruction error. Frequency domain-based methods detect AIGI through the analysis of unique frequency patterns of generated images, such as the spectrum replication caused by upsampling in GAN models. Methods based on pre-trained model features leverage the powerful feature extraction capabilities of large pre-trained models such as contrastive language-image pre-training (CLIP) and Vision Transformer. Fusion feature-based methods combine different types of features, either from single-modality (e. g. , fusing pixel and frequency domain features) or multi-modality (e. g. , integrating image and text features), to improve detection accuracy. In contrast, rule-based methods rely on differences in the sensitivity of real and generated images to certain operations for detection. According to backbone network, early studies rely on traditional machine learning models such as random forest, support vector machine, and logistic regression. Convolutional neural networks such as ResNet have later become increasingly popular due to their powerful capabilities on local features. With the advent of pre-trained large models, visual language models such as CLIP, masked autoencoder (MAE), and large language and vision assistant (LLaVa) are increasingly being used as backbone networks, often fine-tuned with techniques such as LoRA to adapt to the detection task. Aiming to enhance the efficiency and effectiveness of detection algorithms, various technical means are employed, including incremental learning, knowledge distillation, transfer learning, ensemble learning, and few-shot learning. Additionally, data augmentation methods such as JPEG compression, Gaussian blurring, and random flipping are used to increase the diversity of training samples and improve the generalization capability of the detection model. In terms of interpretability of detection methods, several techniques have been developed. Layer-wise relevance propagation analyzes the contribution of each layer in the neural network to the final decision, helping to understand model prediction. Genetic programming optimizes the model structure to improve interpretability. Local interpretable model-agnostic explanations clarifies model's predictions at the local level without relying on the specific model structure. Grad-CAM generates visual explanations by highlighting the important regions in the image for the model's decision-making. Using the detection basis as the main classification criterion, a detailed introduction and analysis of the research status are presented. A detailed summary of benchmark datasets for general AIGI detection is then provided. These datasets substantially vary in structure, scale, image content, and the types of generators they cover. Some datasets are designed for specific types of images or generators. For example, several datasets (such as Forensic Synthetics) are dedicated to GAN-generated image detection, while others (such as Synthbuster) focus on detecting images generated by diffusion models. Comprehensive datasets can also be used for detecting images generated by multiple types of generators. The diversity in datasets reflects the complexity of the AIGI detection task and the need for a standardized evaluation environment. Moreover, the evaluation dimensions of detection methods, namely in-domain accuracy, out-of-domain generalization, and robustness, are introduced in detail. In-domain accuracy is measured using metrics such as accuracy, precision, recall, and average precision, which measure the performance of detection methods on images generated by generative models included in the training set. In contrast, out-

of-domain generalization evaluates the performance of detection methods on unknown generative models and categories outside the training set, which is crucial with the constant emergence of new generative models. This dimension is evaluated through cross-generator, cross-category, and other types of generalization tests. Robustness assesses the capability of detection methods to resist various image post-processing operations, including JPEG compression, Gaussian blurring, and adversarial sample attacks. Furthermore, a horizontal comparison of representative detection methods is conducted, and the results show that different methods exhibit diverse generalization capabilities. The choice of backbone network and training set substantially impacts detection results. For example, some methods perform better on specific generative models, while others demonstrate higher accuracy across multiple generative models. Detection methods based on fusion features generally exhibit superior performance in terms of detection accuracy and generalization. Finally, the paper highlights the challenges and open problems in the current AIGI detection field. These challenges include the construction of large-scale unbiased datasets, because the bias in existing training data can reduce the robustness and generalization of detection models. Designing robust detection methods against physical-level attacks, such as printing and social network-based image processing in real-world scenarios, is also necessary. Moreover, creating comprehensive detection methods that can effectively handle a variety of generative models is crucial to keep up with their rapid advancements. Additionally, research on anti-forensics attacks against detection methods is essential because it can help improve the resilience of detection techniques. Future research directions are outlined to address these challenges and drive the continuous development of AIGI detection technologies.

Key words: artificial intelligence security; artificial intelligence-generated image (AIGI) detection; digital forensics; image forensics; deepfake detection; image generative model; deep learning

0 引 言

当前，ChatGPT(chat generative pre-trained transformer)与扩散模型等生成式大模型快速发展和普及，直接导致人工智能生成内容(artificial intelligence-generated content, AIGC)规模呈爆炸式增长，虚假信息、欺诈内容和不当言论的传播与日俱增，严重影响了信息传播的真实性和可信度，阻碍了社会的良性发展。

图像作为信息的主要传播载体之一，对于引导舆论、维护社会秩序具有重要作用。然而，现如今以生成对抗网络和扩散模型为主的各种生成模型被广泛使用，使得社交网络中的图像真实性有待考量。此外，由于生成模型不断涌现，生成内容纷繁多样、效果逼真，人类对于图像真实性的区分变得十分困难(Mink 等，2022)。因此，检测并鉴别人工智能生成图像(artificial intelligence-generated image, AIGI)十分必要且紧迫。此外，如何达到较好的泛化性和鲁棒性一直是生成图像检测(synthetic image detection, SID)领域的痛点。

随着近年人工智能生成技术的快速发展，有综述文献(Wang 等，2024a)对生成式内容的安全与隐私问题进行了归纳，还有综述(曹申豪 等，2022；Zhang，2022；Wang 等，2024b)专注于 deepfake 检测的研究总结。何沛松等人(2022)对生成对抗网络生成图像的被动取证与反取证工作进行总结，但不包括对扩散模型生成图像的检测。此外，一些综述文献(Lin 等，2024；Tariang 等，2024；谢天圻 等，2024；Deng 等，2025)回顾了多种生成模型生成内容的检测与溯源现状。其中，Lin 等人(2024)对大模型生成内容(包括文本、图像、视频和音频等)的检测工作进行了详细分析；Deng 等人(2025)对 deepfake 和通用 AIGI 的研究工作进行了总结；Tariang 等人(2024)对生成图像的检测与溯源方法进行了汇总。然而，上述综述都没有涵盖 2024 年以来最新的检测方法，且缺乏对 AIGI 检测相关工作与数据集的全面介绍与细粒度分析。本文聚焦通用人工智能生成图像检测，首先介绍主流图像生成模型，然后对研究现状进行介绍分类，并对代表性方法的检测性能进行横向比较，最后总结并给出未来研究方向。

图 1 展示了生成图像检测领域的发展脉络及代表性方法。图 1 中，早期研究工作主要关注生成对抗网络(generative adversarial network, GAN)生成图像的检测，旨在提升检测方法在不同 GAN 模型生成图像上的精度和鲁棒性。2023 年以来，随着扩散模

型的广泛使用，研究人员开始研究面向不同模型架构的通用生成图像检测算法，以提升检测方法对不同模型架构的泛化性和对于多种后处理操作的鲁棒性，从而更好地服务于实际场景应用。

1 图像生成模型

在人工智能技术迅猛发展的进程中，生成式模型凭借其卓越的创造潜能，成为推动该领域前沿拓展的关键驱动力。近10年来，涌现出多种强大的生成式方法，包括变分自编码器（Kingma 和 Welling，2022）、流模型（Dinh 等，2015；Kingma 和 Dhariwal，2018）、生成对抗网络（Goodfellow 等，2014）、自回归模型（Vaswani 等，2017）以及扩散模型（Ho 等，2020）。在这些方法中，生成对抗网络、自回归模型和扩散模型在图像生成任务中展现出卓越的生成能力，并且支持文本到图像的生成任务。基于自然语言的图像编辑技术为图像生成领域赋予了前所未有的灵活性与可控性，为更新颖、更先进的相关应用的发展奠定了基础。

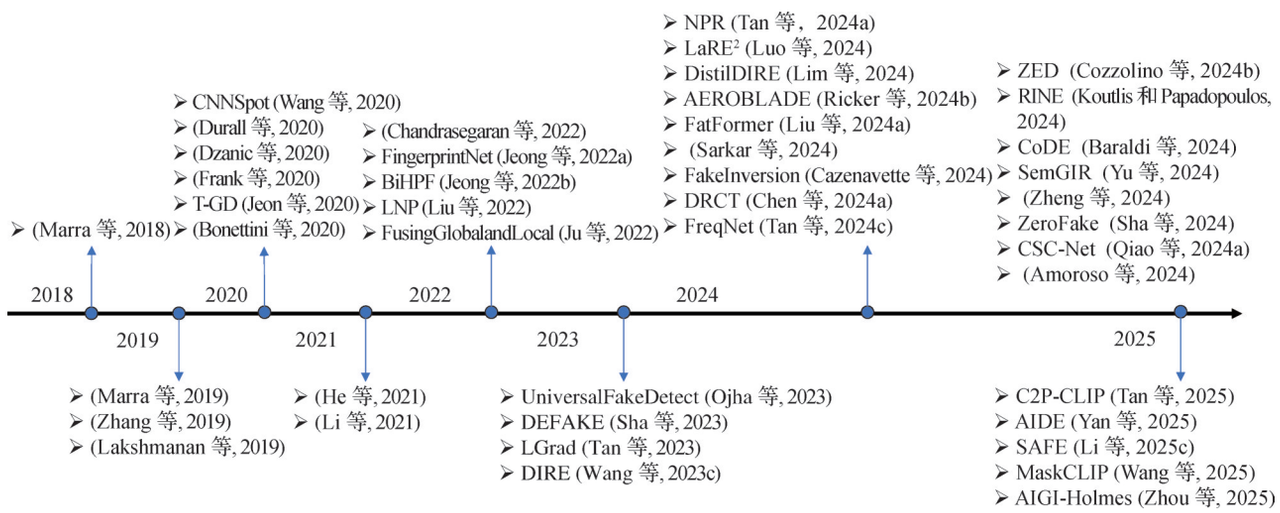

图1 代表性人工智能生成图像检测方法按时间顺序的概述

Fig. 1 Timeline of representative AI-generated image detection methods

1.1 生成对抗网络

生成对抗网络一般由生成器和判别器两部分组成，其核心思想是利用两个网络之间的对抗博弈，其中生成器学习生成逼真的数据，判别器鉴别数据的真实性。在训练过程中，生成器与判别器相互对抗，通过这种相互博弈的训练方式，在判别器更加有效的同时，生成器的生成能力也会提升，最终使得生成器生成的数据趋于逼真，难以被判别器所鉴别。

考虑到传统的GAN模型在生成过程中缺乏对生成内容特定属性的直接控制，Mirza 和 Osindero（2014）提出条件生成对抗网络（conditional GAN，CGAN），将额外的条件信息引入原始GAN架构中，使得生成器和判别器在训练过程中同时考虑条件变量，提高生成器生成结果的可控性。Zhu 等人（2017）率先将GAN应用到图像风格迁移领域，并提出CycleGAN以实现两个领域间的图像转换，使用两个生成器和两个判别器，交替训练生成器和判别器，并使用循环一致性损失确保转换的可逆性。进一步地，StarGAN（Choi 等，2018）为实现多领域图像转换，仅使用一个生成器和一个判别器，生成器通过一个共享的编码器将输入图像编码为一个低维向量，然后将特定标签作为额外信息与该向量共同输入到解码器中进行图像生成。为提高图像的生成质量，Karras 等人（2018）提出ProGAN，借助一种渐进增加图像分辨率的训练策略，使模型能够更加稳定地学习和生成高质量的图像。然而，ProGAN控制生成图像特定特征的能力有限。为此，StyleGAN系列模型（Karras 等，2019，2020）引入“Style”概念，能够实现对生成图像外观特征的独立控制。该方法不仅能够在无监督的情况下分离出高级属性（如表情、姿势和身份），还可以通过添加噪声学习图像的随机细节变化（如雀斑和头发）。这使得人们可以按特定属性的尺度控制图像内容，同时显著提升了图像质量。为生成具有更高分辨率、更加多样性的图像，Brock 等

人(2019)提出 BigGAN,通过增加批量数据与通道数,提高模型的建模能力,从而提升生成图像的质量。Park 等人(2019)基于 CGAN 的核心思想提出 GauGAN,并设计一种新的空间自适应归一化层,能够从草图生成逼真的生成图像。

近来,围绕 GAN 的研究重点已转向基于文本的图像生成,如 StyleGAN-T(Sauer 等,2023)和 GALIP(generative adversarial clips)(Tao 等,2023),二者均基于图像和文本进行联合训练,并利用 CLIP(contrastive language-image pre-training)(Radford 等,2021)作为基础语言模型,生成内容可控的高质量图像。GigaGAN(Kang 等,2023)是首个使用数十亿真实世界图像进行训练的 GAN 方法,不仅能够快速生成高分辨率图像,还支持潜在插值和风格化。

1.2 自回归模型

自回归模型(autoregressive model)凭借其强大的注意力机制,已成为处理序列相关任务的典范。其中,代表性工作是 Transformer(Vaswani 等,2017),采用多头自注意力机制,分别使用编码器和解码器进行编码与解码。

受到 GPT(generative pre-trained Transformer)系列模型(Radford 等,2018,2019;Brown 等,2020)在自然语言建模任务中成功的启发,OpenAI 推出 Image GPT(Chen 等,2020),将展平后的图像序列视为离散标记,采用 Transformer 进行自回归图像生成。生成图像表明,该方法能够模拟图像像素及高级属性(如纹理、语义和比例)之间的空间关系。为生成更高分辨率的图像,Esser 等人(2021)将卷积神经网路(convolutional neural network, CNN)的局部感知能力和 Transformer 的全局建模能力相结合,提出 Taming Transformer,使用 VQGAN(vector quantized generative adversarial network)将图像压缩并存储至离散密码本,基于此训练 Transformer,有效降低了随图像分辨率增长导致的资源消耗。2021 年,OpenAI 推出含有 120 亿参数的图像版 GPT-3——DALL-E(Ramesh 等,2021),使用 VQ-VAE 作为图像编码器,将图像分解为固定大小的块,并将每个块编码为一个离散向量,从而简化图像生成过程。随后,清华大学提出 CogView(Ding 等,2021),模型框架相较于 DALL-E 更为简洁,使用 VQ-VAE 与 GPT 联合训练。考虑到大规模文生图预训练的不稳定问题,针对 Loss NaN 和梯度弥散分别提出 Precision Bottleneck Relaxation 和 Sandwich Layernorm 应对方式。为进一步提升高分辨率图像的生成速度,CogView2(Ding 等,2022)采用分层 Transformer 和局部并行自回归生成,首先使用自监督任务预训练一个跨模态通用语言模型 CogLM,然后对其进行微调以实现快速超分辨率。

2022 年,Google 提出 Parti(Yu 等,2022),与 DALL-E 和 CogView 类似,Parti 同样是一个两阶段模型,首先使用 ViT-VQGAN 训练一个图像分词器,用于在推理时重建图像,然后基于 Transformer 训练一个从文本生成图像的序列到序列模型。该模型最高可扩展至 200 亿参数,并且随着参数量的增长,输出的图像也更加逼真。2023 年,Google 推出的 Muse(Chang 等,2023)使用掩码图像建模进行文生图生成,与 Parti 相比,Muse 因使用并行解码策略生成效率更高。

1.3 扩散模型

扩散模型(diffusion model, DM)的理论基础主要源自统计物理中的扩散过程(Sohl-Dickstein 等,2015),其核心思想是通过模拟数据生成中的“扩散”正向过程和“反扩散”反向过程,学习从随机噪声逐步生成高质量的数据样本。具体而言,正向过程通过逐步添加高斯噪声将自然图像转变为随机噪声,然后通过一系列反向操作的神经网络逐步去除噪声,从而生成与真实数据分布相似的清晰数据样本。这一过程通常需要大量的计算资源进行多次迭代处理,以确保生成的数据质量。

2021 年以来,文生图大模型广泛涌现,比较有代表性的包括 OpenAI 的 GLIDE(guided language to image diffusion for generation and editing)(Nichol 等,2022)、DALL-E2(Ramesh 等,2022)和 DALL-E3(Betker 等,2023)、Google DeepMind 的 Imagen 系列模型(Saharia 等,2022)和 VQDM(vector quantized diffusion model)(Gu 等,2022)、Kandinsky 系列模型(Razzhigaev 等,2023;Vladimir 等,2024)、PixArt 系列模型(Chen 等,2024c,d,e)以及基于 LAION(large-scale artificial intelligence open network)数据集(Schuhmann 等,2022)中数十亿幅图像训练的 Stable Diffusion(Rombach 等,2022)。2024 年,Stability AI 推出 SDXL(stable diffusion xl)(Podell 等,2024),专用于高分辨率图像生成。相较于之前的模型,SDXL 在模型规模、结构和细节上都进行了较大程度的改进。随

着模型的参数量大规模增长，为更好地适配下游任务，DreamBooth（Ruiz等，2023）等针对文生图模型的高效微调方式相继被提出。

与此同时，许多在线文生图平台和工具走进大众视野，如NightCafe（https://creator.nightcafe.studio/）、Midjourney（https://www.midjourney.com/）、Stability AI旗下DeepFloyd Lab的DeepFloyd IF（https://github.com/deep-floyd/IF）、Adobe的Firefly（https://www.adobe.com/products/firefly/）和基于Gradio框架构建的开源AI绘画工具Fooocus（https://github.com/lllyasviel/Fooocus）。

就生成质量而言，扩散模型能够媲美GAN，且优于其他方法（Dhariwal和Nichol，2021）。相较于GAN，扩散模型训练时间更长，但其过程更加稳定，不会出现模式崩溃问题。扩散模型具有更高的灵活性，能够根据多样的文本描述生成复杂的图像，适用于文本到图像的生成（Tariang等，2024）。

2 研究方法

2.1 检测方法分类

人工智能生成图像检测任务可看做一个二分类问题，旨在判断待测图像是由模型生成的图像还是真实图像，后者一般指经相机采集的自然图像。根据不同的划分维度，现有检测方法可以划分为不同的类别，如图2所示。

1）按监督范式划分。当前检测工作可分为基于有监督的检测、基于半监督的检测（Jeon等，2020；Zhu等，2023a）和基于无监督的检测（Durall等，2020；Qiao等，2024b；Lyu等，2024）。大部分工作为基于有监督范式的检测，训练过程中使用的损失函数多为二元交叉熵（binary cross-entropy，BCE）损失，也有工作额外使用如Circle损失（Wu等，2025a）、对比损失（Liu等，2024a；Koutlis和Papadopoulos，2024；Amoroso等，2024；Liang等，2025；Tan等，2025）和InfoNCE（information noise contrastive estimation loss）损失（Baraldi等，2024）等对比学习损失提高检测性能，或引入正则项损失（Durall等，2020）提高训练效率。此外，生成图像的获取方式有以下几种：1）直接由生成模型生成；2）训练一个模拟生成器，如AutoGAN（Zhang等，2019）和Fingerprint-Net（Jeong等，2022b），模拟生成模型的生成特点；3）由真实图像修改获取（Li和Wang，2024；Coccomini等，2024）。

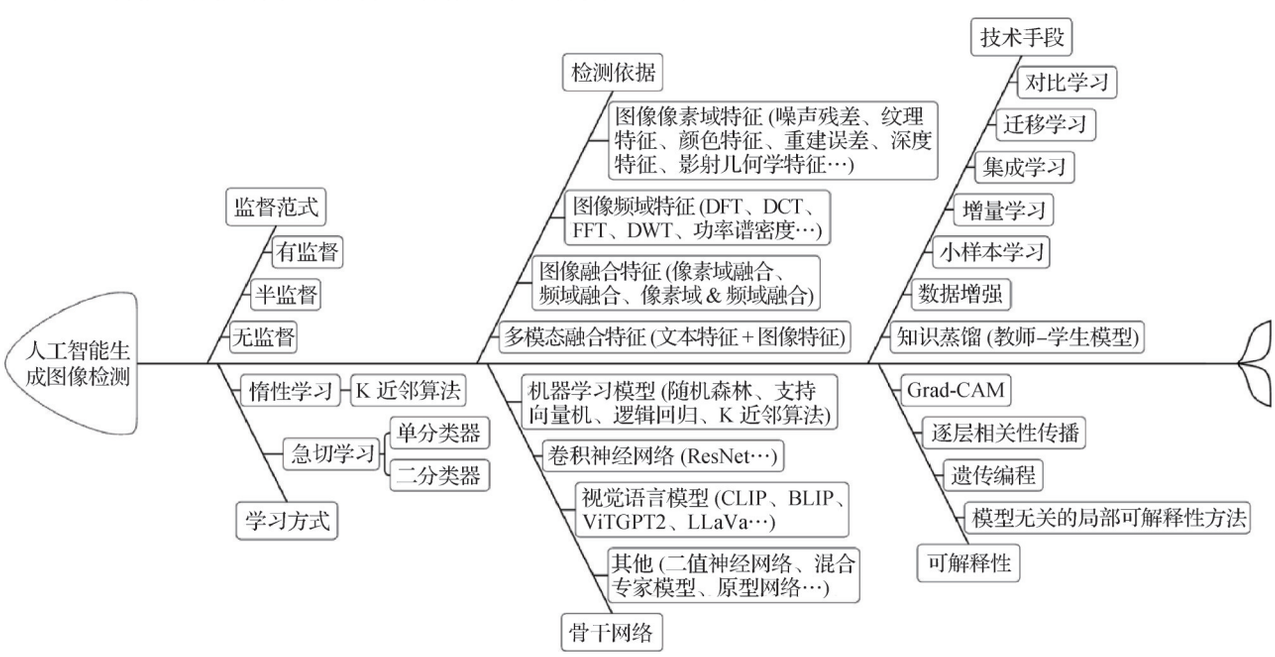

图2 人工智能生成图像检测方法的分类

Fig. 2 Classification of AI-generated image detection methods

2）按学习方式划分。当前检测工作可分为基于急切学习的检测和基于惰性学习（Aha，1997）的检测。基于急切学习的检测可进一步划分为基于单分类器的检测（Bi等，2023；Zhong等，2025）和基于二分类器的检测，现有工作主要为基于二分类器的检测。基于惰性学习的检测通常不需要显式训练过

程，代表性方法为基于 K 近邻（K-nearest neighbors，KNN）算法的检测，根据余弦相似度等相似度衡量指标判断图像的真实性。

3）按检测依据划分。大部分工作使用图像本身的像素域特征（纹理特征、颜色特征、局部特征、深度图特征、像素间相关性和统计特征等）、频域特征（功率谱密度、幅度谱和相位谱等）或融合特征作为检测依据；也有工作基于文本与图像的多模态融合特征完成检测。

4）按骨干网络划分。早期研究工作主要使用随机森林、支持向量机（support vector machine，SVM）、梯度提升树和逻辑回归等传统的机器学习模型作为检测分类器，之后，以 ResNet（residual neural network）（He 等，2016）为主的卷积神经网络逐渐被用做检测分类器，预训练大模型出现后，逐渐有工作使用 CLIP（Radford 等，2021）、MAE（masked autoencoder）（He 等，2022）、BLIP2（bootstrapping language-image pre-training 2）（Li 等，2023）和 LLaVa（large language and vision assistant）（Liu 等，2023）等视觉语言大模型作为骨干网络，采用 LoRA（low-rank adaptation）技术（Hu 等，2022）微调大模型完成检测。

5）按技术手段划分。为提高检测算法的效率和有效性，一些研究工作采用增量学习（Marra 等，2019）、知识蒸馏（Zhu 等，2023a；Lim 等，2024；Cavia 等，2024）、迁移学习（Zhang 等，2024）、集成学习（Raza 等，2024；Mehta 等，2025）和小样本学习（Yu 等，2024；Xu 等，2025a；Wu 等，2025b）等技术手段。一些工作（Mi 等，2020；Xu 等，2023，2024a，b）将注意力机制引入 CNN，有效提升检测器的全局信息提取能力。为增强检测方法的泛化性和鲁棒性，一些工作在将训练图像送入检测器前的预处理阶段会进行归一化、裁剪、随机翻转和尺寸变换等预处理操作（Marra 等，2019；Wang 等，2020；Mi 等，2020）；还有一些工作在检测器训练阶段加入数据增强模块，常用的数据增强方式包括 JPEG（joint photographic experts group）压缩（Wang 等，2020；Mandelli 等，2020；Jeon 等，2020；Chen 等，2024a）、高斯模糊（Wang 等，2020；Jeon 等，2020；Chen 等，2024a）、高斯噪声扰动（Chen 等，2024a）、随机翻转（Jeon 等，2020；Chen 等，2024a）和随机旋转（Chen 等，2024a；Li 等，2025c），也有部分研究工作使用类内 CutMix（Jeon 等，2020）、颜色抖动（Lanzino 等，2024；Li 等，2025c）、随机灰度（Chandrasegaran 等，2022）、随机掩码（Doloriel 和 Cheung，2024；Chen 等，2024a；Li 等，2025c）以及亮度和对比度调整（Chen 等，2024a）等方式提高训练样本的多样性。

6）按可解释性划分。为提高检测方法的可解释性，Chandrasegaran 等人（2022）使用逐层相关性传播（layer-wise relevance propagation，LRP）方法探究检测器的可解释性；Lin 等人（2023）采用遗传编程算法提高检测方法的可解释性；Pontorno 等人（2024）使用模型无关的局部可解释性方法对检测器的预测结果进行解释；还有研究工作（Bird 和 Lotfi，2024；Sarkar 等，2024；Zheng 等，2024；Mehta 等，2025）使用 Grad-CAM（Selvaraju 等，2017）对检测结果进行解释。

2.2　检测方法介绍

图 3 直观展示了当前通用人工智能生成图像检测流程，并给出每个阶段的关键要素。下面以检测依据作为划分标准对研究现状进行详细分类与介绍。

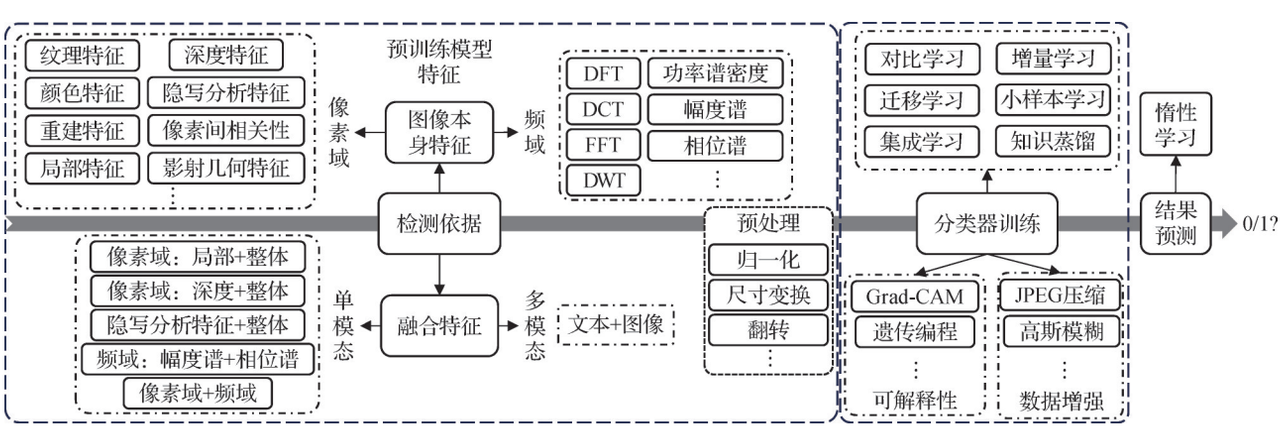

图 3　通用人工智能生成图像检测流程

Fig. 3　Illustration of general AIGI detection pipeline

2.2.1　基于图像像素域特征的检测方法

此类检测方法以真实图像或生成图像在像素域方面的共性为线索完成检测。

1)基于隐写分析特征的检测。在研究的早期阶段,研究者使用隐写分析领域广泛使用的统计特征进行检测。Marra等人(2018)率先提出社交网络中GAN生成图像的检测任务,分别对基于富模型的隐写分析特征、基于GAN判别器以及基于CNN(DenseNet、InceptionNet和XceptionNet)的检测方法准确性进行了评估。实验结果表明,在无损情况下,上述检测方法都能够取得较好的检测性能,然而在JPEG压缩场景下,上述检测方法均出现不同程度的性能下降。

2)基于图像纹理特征的检测。不同于上述直接将图像送入CNN的检测手段,一些工作考虑引入传统的图像取证特征,结合CNN实现检测。共生矩阵反映像素间的相关性,常用于捕捉图像的纹理信息。Nataraj等人(2019)将共生矩阵与卷积神经网络相结合,首先提取输入图像RGB 3个颜色通道的共生矩阵,然后将其堆叠送入一个浅层卷积神经网络进行检测。类似地,Goebel等人(2021)将图像三通道共生矩阵输入XceptionNet变体中,实现对5种GAN生成图像的检测。Qiao等人(2024a)提出CSC-Net(cross-color spatial co-occurrence matrix network),将跨颜色空间共生矩阵特征作为检测依据,并设计一个轻量级深度神经网络降低训练成本。Xu等人(2024a)将图像的全局纹理特征与局部纹理特征通过低秩张量表征融合进行检测。

由于网络结构或计算资源的限制,现有检测方法通常会调整输入图像的大小或居中裁剪输入图像,然而,这一操作会对检测高分辨率图像中出现的伪影造成影响。为解决这一限制,Konstantinidou等人(2025)提出图像预处理组件TextureCrop,聚焦生成伪影普遍存在的图像高频区域,挑选出原始图像中的纹理丰富部分送入分类器检测,将所有部分检测分数的均值作为最终检测结果。该组件可以插入任何预先训练的检测模型以提高其性能。

3)基于图像颜色特征的检测。一些工作利用图像的颜色信息完成检测。Chandrasegaran等人(2022)在使用LRP分析神经网络的决策过程时发现,对检测器决策结果影响最大的特征层大多具有相似的颜色表示,由此将颜色信息作为用于检测的可迁移特征,并有针对性地提出基于随机灰度化的数据增强方式。随后,Uhlenbrock等人(2024)发现,生成模型在训练过程中,感知损失的使用会导致生成图像与真实图像在亮度上产生比色度更大的差异,于是提出将颜色通道间的关系特征送入随机森林进行分类。Jia等人(2025)发现,真实图像与生成图像在颜色分布上有所差异,提出CoD(color-based detecting model),通过颜色量化与重建差异提取有效特征,构建轻量级检测模型。

4)基于图像重建的检测。此类工作围绕原始图像的重建图像设计检测方法,可进一步分为基于重建误差的检测和基于重建图像特征的检测。

关于基于重建误差的检测,He等人(2021)提出使用图像再生成的方式检测生成图像,将原始图像与使用图像上色、超分辨率和去噪等方式得到的再生成图像相减得到噪声残差,将其输入CNN训练一个二分类器。Zhang和Xu(2025)发现,扩散模型的反扩散过程会使真实图像和生成图像展现出不同的潜空间高斯表征,于是使用预训练扩散模型的"反扩散"过程提取的扩散噪声特征(diffusion noise feature, DNF)作为通用图像表征,将其送入CNN进行检测。Wang等人(2023c)关注到扩散模型的可逆特性,发现与真实图像相比,扩散模型可以更准确地重建扩散过程产生的图像,并在此基础上提出扩散重构误差(diffusion reconstruction error, DIRE),用于检测基于扩散模型的生成图像。Javaheri等人(2024)将DIRE作为输入训练检测分类器。为提高DIRE的检测效率,DistilDIRE(Lim等,2024)使用知识蒸馏策略,有效提升了DIRE框架的检测效率;LaRE^2(latent reconstruction error guided feature refinement)(Luo等,2024)使用潜空间重建损失提升检测速度。SeDID(stepwise error for diffusion-generated image detection)(Ma等,2023)利用扩散模型的可逆特性,将生成过程中特定时间步长的加噪样本和去噪样本之间的误差作为检测依据。Chu等人(2025)发现生成模型难以重建真实图像的中频信息,提出基于频域引导的重建误差(frequency-guided reconstruction error, FIRE)检测生成图像,具体而言,设计频域掩码优化模块分离原始图像中频部分,对其重建得到伪重建图像,然后将伪重建图像与原始图像的重建图像拼接送入CNN分类,并设计损失使得分离出的中频信息更加准确。AEROBLADE(Ricker等,

2024b)，针对潜空间扩散模型(latent diffusion model，LDM)，使用自编码器的重建误差检测图像是否为生成图像。然而，该方法仅限于检测基于自编码器模型生成的图像，因此存在局限。为更加充分地利用重建过程，Shen等人(2025)将单时间步重建误差扩展为多时间步重建误差，以提高检测精度。

关于基于重建图像特征的检测，FakeInversion(Cazenavette等，2024)将原始图像、解码潜空间噪声图和解码重建图像一同送入CNN进行分类。DRCT(Chen等，2024a)首先得到真实图像和生成图像各自的重建图像，然后基于真实图像、真实图像重建图像、生成图像和生成图像重建图像4类图像，使用对比损失训练分类器，得到更加准确的决策边界。Yu等人(2024)将重建图像特征与文本和原始图像特征融合，送入CNN完成检测。受Lempitsky等人(2018)使用深度图像先验进行图像重建的启发，Sinitsa和Fried(2024)提出DIF(deep image fingerprint)，首先训练一个模型提取真实图像和生成图像的指纹特征，期间使用皮尔森相关系数约束重建过程。测试时，通过比较待测图像分别与真实图像和生成图像的相似度，判定图像的真实性。为检测扩散模型生成图像，Yang等人(2024)设计了一种面向文本模态的特征提取方法TOFE(text modality-oriented feature extraction)，以文生图模型中的文本作为高层语义表示为灵感，利用去噪扩散隐式模型(denoising diffusion implicit models，DDIM)反演获得图像潜向量的轨迹，通过最小化生成潜向量与真实潜向量的均方误差迭代优化文本条件嵌入，使其包含细粒度细节等低级特征，最终得到能引导目标图像生成的文本表示，该表示融合图像高级特征和低级特征，可用于构建检测器。

5)基于图像局部特征的检测。不同于上述工作使用图像的全局特征作为检测依据，一些研究工作仅使用图像的局部信息完成检测。其中，SSP(single simple patch)(Chen等，2024f)挑选出纹理最为贫乏的单个图像块，基于局部感受野的块分类器完成检测。LaDeDa(Cavia等，2024)通过限制CNN的感受野，得到图像块级别特征，然后通过平均池化得到图像级别的特征，用于二分类。为实现不同分辨率图像的检测，SPAI(spectral AI-generated image detection)(Karageorgiou等，2025)分别提取不同图像块的频谱特征，使用注意力机制融合送入多层感知机(multi-layer perceptron，MLP)检测。Xiao等人(2025)将输入图像分割成多个块，通过预测每个块的可检测性分数，选择分数最高的块送入检测器。

6)基于图像内部相关性的检测。一些研究工作基于图像中像素的相关性进行检测，其中，PatchCraft(Zhong等，2024)利用图像内纹理丰富区域和纹理贫乏区域的像素相关性对比度检测生成图像。Zhong等人(2024)认为相较于生成图像，真实图像的两个区域的像素相关性更高。具体而言，首先将图像划分为纹理丰富区域和纹理贫乏区域，然后使用高通滤波器分别提取两个区域的噪声模式，最后将其送入CNN完成分类。与Zhang等人(2019)类似，Tan等人(2024a)探究上采样产生的伪影在像素域的影响，借此提出将邻近像素关系(neighboring pixel relationships，NPR)作为检测的通用表征。Chen等人(2024b)提出IPD-Net，借助图像块之间的依赖关系完成检测。

7)基于图像语义特征抑制的检测。为降低图像的语义内容对检测结果的干扰，Amoroso等人(2024)使用对比学习解耦图像语义特征和风格特征，Zheng等人(2024)在图像预处理时对图像块进行置乱操作，零填充至固定维度，并训练一个基于图像块的分类器提取各图像块特征，最后将所有图像块特征进行拼接送入线性分类器进行检测。类似地，Gye等人(2025)通过多级图像块置乱捕捉图像低级纹理特征，结合完整图像的高级语义特征进行检测。Tao等人(2025)提出SAGNet(semantic-agnostic artifact network)，采用对抗训练的思想，通过梯度反转技术，迫使共享特征提取器学习无法被类别判别器识别的特征，从而抹除特征中的语义信息，实现语义特征的解耦。为减轻预训练模型中与判别无关特征带来的偏差与泛化能力下降问题，Zhang等人(2025)基于变分信息瓶颈(variational information bottleneck，VIB)，提出VIB-Net，通过对特征进行过滤与压缩，有效保留关键的判别性信息。

8)其他基于图像像素域的检测方法。Tan等人(2023)首次提出基于梯度学习的检测方法LGrad，使用预训练CNN模型将图像转换成梯度图并归一化，然后在梯度图数据集上训练得到一个区分真实图像和生成图像的二分类器。预训练CNN模型的梯度，即损失函数相对于输入图像的导数，表示输入图像中每个像素对模型输出的影响。这种做法排除

了图像语义对于检测器的影响，巧妙解决了检测器对于训练数据的依赖问题，然而由于特定预训练模型的使用，LGrad并不能保证在非基于CNN架构的模型生成图像上表现良好。Lorenz等人（2023）将对抗样本检测领域常用的局部内在维度（local intrinsic dimensionality，LID）作为检测依据，首先使用CNN提取特征图计算LID，然后将其送入随机森林训练。Sarkar等人（2024）挖掘真实图像与生成图像间的几何差异，将影射几何学特征如阴影、透视场和线段作为线索检测生成图像。Guarnera等人（2024）通过微调CNN完成生成图像级联检测与溯源任务。Tan等人（2024b）考虑到不同来源的数据存在的分布差异，提出一种新型图像伪影表征提取方法，使用高通滤波器、低通滤波器、预训练卷积层和随机初始化卷积层等多种无需训练的数据无关算子（data-independent operator，DIO）将多个域的伪影进行对齐，减少特征提取器对训练数据的依赖。Xu等人（2025b）先后使用图像的RGB空间特征与描述文本特征、中高频特征、纹理特征以及图像布局作为检测依据，并对上述检测和溯源方法的性能进行了深度分析。Nguyen等人（2025）通过自监督学习从真实图像中获取一组预测滤波器，以提取包含来源信息的残差，再通过一个紧凑模型对这些残差进行多尺度建模，将其参数作为图像的“法证自描述”直接刻画图像来源，从而在无需见过任何合成图像的情况下，实现强大的零样本检测、开集溯源和无监督聚类。

2.2.2　基于图像频域特征的检测方法

此类检测方法以真实图像或生成图像在频域方面的共性为线索完成检测，主要通过将图像转化为频谱图送入检测网络中进行检测。

1）基于离散傅里叶变换的检测。Zhang等人（2019）指出主流GAN模型的上采样模块会产生独特伪影，并从理论上证明这些伪影在频域中表现为频谱的复制。基于此，他们使用二维离散傅里叶变换（discrete Fourier transform，DFT）提取出的频域伪影作为检测依据，训练基于频谱输入的分类器模型。此外，他们构建了一个AutoGAN，其上采样模块包括反卷积和最近邻插值，用于模拟多个GAN模型的生成过程生成图像样本。AutoGAN无需访问实际GAN模型便可生成用于训练的“假”图像，因此一定程度上缓解了训练过程中特定模型的生成样本缺失问题。在测试阶段，当待测样本由与AutoGAN网络结构相似的模型生成时，该方法能够取得较好的检测性能，因此具有一定的泛化能力。与Zhang等人（2019）类似，Durall等人（2020）将DFT得到图像频谱送入SVM分类器或使用K-Means算法完成检测。此外，他们提出一种正则化项损失，使生成图像与真实图像的频谱分布尽可能接近。Dzanic等人（2020）将高频频谱与KNN算法结合，发现仅用少量训练数据即可识别生成图像，且当图像的分辨率较高或压缩率较低时，产生的频谱特征更加容易区分。Jeong等人（2022a）使用对抗训练的方式得到图像的伪影压缩图，证实了生成图像中伪影的存在，基于此发现，提出一种基于双边高通滤波的检测方法BiHPF（bilateral high-pass filters），首先使用二维傅里叶变换将输入图像转换为频率图的幅度谱，然后对幅度谱分别进行像素级高通滤波和频率级高通滤波，重点关注背景区域的伪影和高频分量，最后将滤波后的幅度谱图作为输入送至CNN检测。受Corvi等人（2023a）启发，Li等人（2024）提出基于频谱相似性的检测方法MaskSim，首先使用去噪网络DnCNN（Zhang等，2017）对图像进行预处理，然后对降噪图像进行DFT得到频谱图，并设计可训练掩码，关注频谱图中对区分自然图像和生成图像贡献最大的频率，得到增强频谱。之后，使用可训练参考频谱使得提取出的增强频谱更加准确。最后，通过计算参考频谱与增强频谱的相似性训练逻辑回归分类器完成检测。

2）基于离散余弦变换的检测。Frank等人（2020）指出，GAN模型的上采样过程会使生成图像的相邻像素具备一定的相关性，直接导致图像在频域产生异常的中高频分量，因此首先对输入图像经离散余弦变换（discrete cosine transform，DCT）得到频谱图，然后对其进行对数缩放并归一化送入CNN，结果表明基于频谱的CNN分类器相较于基于像素域的CNN分类器能够取得更好的检测结果。Bonettini等人（2020）分析图像像素间的关系发现，自然图像对应的DCT系数的首位数字分布遵循本福特定律，而GAN模型生成图像的DCT系数不符合上述规律，借此提出将图像DCT系数的首位数字分布作为检测依据，结合随机森林算法进行检测。Pontorno等人（2024）发现不同的DCT系数对检测结果的贡献不同，并使用K近邻方法、梯度提升和随机

森林对DCT系数分类完成检测。为抵御黑盒对抗扰动对检测器性能的影响，Hooda等人(2024)提出不相交扩散深度伪造检测方法，利用频域特征的冗余特性，根据显著性分数将原始图像的DCT特征划分成多份，分别送入分类器检测。Karageorgiou等人(2025)认为真实图像的频谱分布具有更高的不变性，并提出基于谱学习的检测方法SPAI，首先使用自监督学习的方式训练一个频谱重建模型，然后使用频谱重建相似度捕捉生成图像。

3)基于快速傅里叶变换的检测。Jeong等人(2022b)认为生成图像的指纹具有相似性，因此设计FingerprintNet，首先基于真实图像集得到生成图像集，然后将二者共同作为训练集，并将图像经快速傅里叶变换(fast Fourier transform，FFT)作为检测分类器的输入。这种方法在训练阶段并不依赖于任何特定生成网络的图像，因此具有较高的泛化性。Bammey(2024)提出Synthbuster，首先将图像通过交叉差分进行高通滤波，然后将FFT得到的幅度谱峰值特征送入基于直方图的梯度提升树分类器检测DM生成的图像。

2.2.3　基于预训练模型的检测方法

此类检测方法利用预训练模型的特征提取能力完成检测，主要包括两类策略：直接使用预训练模型提取的特征进行检测，或在此基础上微调预训练模型以构建检测器。

1)基于预训练模型特征的检测。随着ViT(vision Transformer)(Dosovitskiy等，2021)、CLIP(Radford等，2021)和BLIP系列(Li等，2022a，2023；Xue等，2025)大模型技术的发展，预训练模型展现出强大的特征提取能力，有研究工作直接借助其表征能力完成检测。其中，Ojha等人(2023)使用预训练的CLIP模型提取图像特征，用于训练检测器，使其学习到更加平衡的决策边界，以此提升检测器对于GAN和扩散模型等不同生成模型的泛化性。类似地，Cozzolino等人(2024a)直接将CLIP提取出的图像特征送入SVM进行检测，发现使用少量样本就能够达到很好的检测效果。Gaintseva等人(2024)使用CLIP提取出图像特征的残差向量投影作为特征送入逻辑回归模型进行分类，并提出特征移除和注意力机制两种方法提高检测器的鲁棒性。Yang等人(2025)提出差异伪造检测器(discrepancy deepfake detector，D3)，设计并行网络分支，使用预训练CLIP分别提取原始图像特征和置乱图像特征，然后将两类特征拼接，训练自注意力层和全连接层进行分类。这种方法使得检测器能够学习多个生成器的通用特征，从而在检测各种生成模型时实现更好的泛化性和鲁棒性。Herur等人(2024)同样将CLIP提取出的特征用于检测，进一步判断图像是否为篡改图像。Koutlis和Papadopoulos(2024)提出RINE(representations from intermediate encoder-blocks)，提取CLIP图像编码器中间Transformer模块的图像表示，设计轻量级网络将其映射到一个可学习伪造感知向量空间，并使用一个可训练模块预测每个Transformer模块的重要性。Lamichhane(2025)微调ViT检测GAN生成的图像。Huang等人(2025)提出一个集成检测、定位与解释功能的大模型框架SIDA(social media image detection，localization，and explanation assistant)，将指令与待测图像送入视觉语言模型(vision-language model，VLM)，将特定token的特征作为检测依据。Wang等人(2025)提出Mask-CLIP，通过自注意力模块融合CLIP视觉编码器的多层级特征生成全局图像表征，并利用可学习的连续提示向量对CLIP进行提示调优。在测试阶段，通过图像—文本在CLIP共享语义空间内的相似度完成分类，展现出优异的跨生成模型泛化能力。

2)基于预训练模型微调的检测。一些研究工作通过微调视觉语言模型完成检测任务。Wu等人(2025a)提出文本引导的检测方法LASTED(language-guided synthesis detection)，首先设计图像文本对，基于对比学习损失微调预训练模型的图像编码器和文本编码器。在测试阶段，将测试图像与锚图像在图像编码器的特征层面进行相似度比对，通过阈值决定检测结果。Xu等人(2025a)提出FAMSeC(lora-based forgery awareness module and semantic feature-guided contrastive learning strategy)，首先基于对比学习和LoRA微调CLIP得到用于检测的特征提取器，在测试阶段，通过比较测试图像分别与自然图像集和生成图像集的距离判定图像的真实性。Chang等人(2024)使用提示调优策略微调InstructBLIP(Dai等，2023)模型得到检测分类器。Moskowitz等人(2024)通过微调CLIP证实了其强大的检测能力。Khan和Dang-Nguyen等人(2024)进一步围绕CLIP预训练模型，提出提示调优、适配器调优、微调和线性探测4种微调方式完成检测。Liu等

人(2024b)提出基于混合专家模型的检测方法,使用LoRA高效微调CLIP-ViT完成检测。Keita等人(2024b)提出FIDAVL(fake image detection and attribution using vision-language model),利用视觉语言模型的零样本学习能力,将Q-Former和Vicuna模型(https://github.com/lm-sys/FastChat)相结合作为骨干网络,借助软提示微调策略检测生成图像。Keita等人(2024a)将生成图像检测问题转化为一个图像描述任务,通过微调ViTGPT2(https://ankur3107.github.io/blogs/the-illustrated-image-captioning-using-transformers/)等视觉语言模型检测扩散模型生成图像。类似地,Bi-LORA(Keita等,2025)使用LoRA微调BLIP2模型(Li等,2023)完成扩散模型生成图像的检测。Tan等人(2025)发现,CLIP对生成图像强大的检测能力是源于其对近似概念的识别,因此提出C2P-CLIP以提高检测泛化性,通过引入类别通用提示,使用对比学习对真实图像和生成图像概念进行强化,借助LoRA微调CLIP,打造通用生成图像检测器。还有工作将生成图像检测任务适配大语言模型完成检测。Fan等人(2024)提出Fake-GPT,将生成图像检测任务转换为序列预测任务,将图像的RGB像素值表示成序列作为输入,固定提示文本微调大语言模型以预测图像真假,该方法无需位置编码或特征对齐,验证了大语言模型在非文本领域任务的出色性能。为提高检测器的可解释性,Zhou等人(2025)提出AIGI-Holmes三阶段训练框架,首先通过视觉专家预训练赋予检测器领域特定的图像感知能力;随后利用监督微调引导检测器学习生成结构化的检测报告;最后通过直接偏好优化使检测器的解释能力与人类判断标准对齐。经过这一系列阶段训练,模型在检测精度、跨域泛化性及可解释性方面均表现出显著提升。

2.2.4 基于融合特征的检测方法

按模态数量划分,基于融合特征的检测方法可进一步分为基于单模态融合特征的检测方法和基于多模态融合特征的检测方法,前者指仅对图像的不同特征进行融合,后者指对图像和文本等多种模态的特征进行融合。

1)基于单模态融合特征的检测方法。部分工作将像素域的特征融合作为检测依据。Xi等人(2023)使用交叉注意力机制,将图像的隐写分析特征与整体特征进行融合得到检测依据。Ju等人(2022)首先使用CNN提取图像的全局特征,然后基于全局特征图,设计一个图像块选择模块选取部分图像块提取细粒度局部特征,最后通过一个注意力机制模块融合全局特征与局部特征,将得到的特征向量送入CNN进行检测。之后,Ju等人(2024)对上述算法中的特征提取方法进行了改进,提出GLFF(global and local feature fusion),将图像整体的多尺度全局特征与信息丰富图像块的精细局部特征通过注意力机制进行结合。Leporoni等人(2024)将图像的深度特征与RGB空间特征通过注意力机制融合完成检测。MMGANGuard(Raza等,2024)将纹理特征和CNN提取的图像特征同时用于检测,类似地,HFMF(Mehta等,2025)将CNN和CLIP提取的多尺度图像特征与边缘等局部特征结合完成生成图像的检测。Baraldi等人(2024)指出,以往基于CLIP等预训练模型特征的检测方法只关注到了全局特征的提取,提出CoDE(contrastive deepfake embeddings),同时考虑图像的全局和局部特征,使用对比学习对齐图像全局和局部区域的相似性。

一些工作将频域的特征融合作为检测依据。Liu等人(2022)观察到真实图像的噪声模式在频域中表现出相似的特征,而不同模型生成的图像相较而言却大不相同。因此,他们通过检查图像是否遵循真实图像的模式来判定图像的真实性:首先使用降噪网络CycleISP(Zamir等,2020)提取图像噪声,与原图像相减得到学习噪声模式(learned noise patterns, LNP),然后利用LNP的幅度谱和相位谱的融合特征训练CNN检测器。Tan等人(2024c)提出频率感知的轻量级检测网络FreqNet,利用图像本身的高频表征以及图像的高频特征在空间和通道维度上的高频表征作为检测依据,具体而言,FreqNet包含一个频域学习模块,对FFT得到的相位谱和幅度谱进行卷积操作,使检测模型聚焦于图像的高频特征。

此外,有工作同时将图像的频域与像素域特征用于检测。Liu等人(2024a)提出FatFormer,分别使用轻量级卷积网络和离散小波变换(discrete wavelet transform, DWT)提取像素域和不同频段的频域伪造特征。此外,引入语义引导的对齐模块提升检测方法的泛化性。Lanzino等人(2024)将图像RGB空间整体特征、FFT提取的频域幅度谱特征图以及局部二值模式提取的纹理特征图的融合特征作为检测依据,并首次提出基于二值神经网络的检测方法,选

取 BNext（Guo 等，2022）作为骨干网络，大大提升了检测效率，以适应实际场景中的实时检测需求。Meng 等人（2024）分别提取图像的像素域和 DFT 频域特征，然后对像素域特征采取可学习正交分解，对频域特征提取可疑频带，将图像特征分解为伪影相关特征和伪影不相关特征，并使用互信息估计器增大两种特征之间的距离。最后将伪影相关特征送入分类器。Wißmann 等人（2024）提出 Whodunit，将图像的 DCT 频谱、功率谱密度以及像素域的自相关系数用于检测。Gallagher 和 Pugsley（2024）设计基于图像 DFT 频域特征和像素域特征的双路分类器。受 PatchCraft（Zhong 等，2024）启发，Chen 和 Yashtini（2024）对图像的纹理丰富区域和纹理贫乏区域，分别使用位置编码和 DFT 得到像素域特征和频域特征，然后对二类区域的特征进行拼接作为检测依据。Li 等人（2025b）提出 CSFD（collaborative spatial and frequency detector），首先将图像划分为纹理丰富区域和纹理贫乏区域，然后分别对这两个区域聚合频域中的不同分量，使用加权通道注意力提升空间推理能力，最后融合两区域特征送入分类器检测。Liu 等人（2025）将图像空间特征、梯度图特征和多尺度高频特征融合，送入 HRNet（high-resolution network）（Sun 等，2019）进行检测。

还有工作将图像的空间特征与语义特征融合。例如，AIDE（ai-generated image detector with hybrid features）（Yan 等，2025）将根据 DCT 系数划分的高低频图像块特征作为局部空间特征，与 CLIP 提取出的全局语义特征拼接后送入 MLP 分类。

2）基于多模态融合特征的检测方法。Sha 等人（2023）聚焦文生图模型的生成图像检测任务，指出相较于真实图像，生成图像与提示文本之间的距离更加接近，提出 DEFAKE，将 CLIP 提取出的文本与图像特征进行拼接，然后训练一个 MLP 进行分类。在 DEFAKE 的基础上，Santosh 等人（2024）提出一个双目标损失框架，将条件风险值损失和受试者工作特征曲线下面积损失协同作用，旨在解决训练集中困难样本和样本不均衡问题带来的挑战。此外，Santosh 等人（2024）引入锐度感知最小化的参数优化策略，以提高检测方法的泛化性。Yu 等人（2024）提出基于图像语义的图像重建方法 SemGIR（semantic-guided image regeneration），该方法面向小样本学习场景，首先通过反推原始图像的提示词生成参考图像，然后将文本特征、原始图像特征与重建图像特征进行融合，送入 CNN 完成生成图像的检测。

2.2.5 基于特定规则的检测方法

此类检测方法通常将两类图像对于相同操作的敏感程度差异作为检测依据，一般无需训练，且由于检测过程不依赖分类器，检测结果往往无偏。Ricker 等人（2024b）提出 AEROBLADE，使用自编码器的重建误差，根据 LPIPS（learned perceptual image patch similarity）指标检测图像是否为生成图像。Choi 等人（2024）发现 AEROBLADE 在检测具有简单背景的图像时表现不佳，指出 AEROBLADE 本质上是对背景图像的过拟合，为解决这一问题，他们提出 HFI（high-frequency influence），通过测量重建图像高频信息的失真检测生成图像，即分别对原始图像和低通滤波后的原始图像进行重建，根据重建距离的差异完成检测。He 等人（2024）指出，相较于生成图像，真实图像在特征层面对细微的扰动更加鲁棒，因此提出 RIGID（robust ai-generated image detection），使用自监督模型 DINOv2（Oquab 等，2024）提取高斯噪声扰动前和扰动后的特征，根据二者相似度判断图像是否由模型生成。Tsai 等人（2024）进一步对 RIGID 进行了改进，根据高斯模糊图像和去高斯模糊图像特征之间的距离检测生成人脸图像，并提出 MINDER，消除不同噪声导致的检测偏见，提高检测方法在通用图像上的泛化能力。Sha 等人（2024）针对文生图模型的生成图像检测提出 ZeroFake，通过对抗提示的方式得到参考图像，计算测试图像与参考图像的结构相似性（structure similarity index measure，SSIM）分数进行检测。

2.2.6 其他检测方法

Li 等人（2022b）提出基于 GAN 图像伪影相似度评估的检测方法 GASE-Net，基于锚图像分别与生成图像和真实图像的特征相似度构建打分网络，并设计一种类别和域感知损失函数对打分网络进行训练。该工作首次将测试图像与真实图像的距离作为评判的依据之一，根据距离的相对性得到检测结果。Dogoulis 等人（2023）首先基于真实图像训练一个图像质量评估模型，然后基于此模型选取部分高质量生成图像与真实图像一起送入 CNN 检测。与 Liu 等人（2022）类似，Liang 等人（2025）认为，真实图像相较于生成图像具备更加稳定的统计特征，并基于此

设计检测方法NTF(natural trace forensics),首先将自然图像特征解耦为同质特征和异质特征,然后通过对比学习,使得自然图像的同质特征与生成图像特征差异更为明显,便于检测器的检测。Cozzolino等人(2024b)借鉴文本生成任务中对下一个token的预测任务,提出图像"编码代价"概念并据此完成检测。

Karageorgiou等人(2024)发现,随着生成图像在社交媒体中的传播,现有方法的检测性能均出现不同程度的降低趋势,且随着时间的推移,检测性能持续下降。基于此,他们提出基于检索辅助的检测方法RASID(retrieval-assisted synthetic image detection),首先对初始时段内的图像检测并保存检测结果,在对后续图像进行检测时,查询待测图像与检测结果中图像的相似度。若相似度大于某一阈值则基于现有的检测结果进行检测;否则,将待测图像当做新样本送入检测器进行检测。

Wu等人(2025b)提出基于小样本学习的检测方法FSD(few-shot detector),首先训练一个原型网络生成每个类别的原型表示,在测试阶段,根据查询样本与原型表示的距离得到检测结果,该方法能够实现扩散模型生成图像的检测与溯源。Li等人(2025c)发现,当前训练范式中存在伪影特征弱化和伪影特征过拟合两种偏差,且生成图像的成像机制使得像素间的局部相关性增强。鉴于此,提出一种轻量有效的检测器SAFE(simple preserved and augmented features),采用3种简单的图像变换。首先,针对伪影特征弱化问题,在图像预处理中使用裁剪操作替代下采样操作,以避免伪影失真。其次,针对伪影特征过拟合问题,引入颜色抖动和随机旋转作为额外的数据增强手段,以减轻有限训练样本中颜色差异和语义差异带来的无关偏差。然后,为实现局部感知,设计一种基于图像块的随机掩蔽策略,在训练过程中使得检测器关注局部区域。

以上详细介绍了当前人工智能生成图像检测方法,表1对代表性研究工作进行了归纳总结。

2.3 反取证方法介绍

随着各种检测方法陆续提出,一些工作开始关注AIGI的反取证问题。Chandrasegaran等人(2021)发现,上采样操作导致的频谱差异并非CNN生成图像所固有,因此并不能作为检测依据。类似地,Dong等人(2022)提出针对基于频域特征检测的对抗方法。此外,还有工作使用基于对抗样本的反取证方法逃避检测,如Stealthdiffusion(Zhou等,2024)和R^2BA(realistic-like robust black-box adversarial attack)(Xie等,2024),其中,Xie等人(2024)使用高斯模糊、JPEG压缩、高斯噪声和光照等真实场景中常见的图像后处理操作构建对抗样本。

3 数据集

3.1 基准数据集介绍及分类

为实现多种检测方法的公平比较,一些工作致力于构建完备可靠的评价基准数据集。表2按时间顺序汇总了当前用于通用AIGI检测的基准数据集及相关信息,生成模型的不同变体可看做不同生成模型。可以看出,不同数据集在结构、规模、图像内容和格式以及生成模型种类和数量等方面存在较大差异,且侧重点有所不同。

关于数据集结构和规模,大部分数据集包含数十万幅图像,也有少量数据集包含千万幅图像。图像的生成来源涵盖从数个到数千个。Hong等人(2025)首次提出级联结构的数据集WildFake,从种类、提出时间、架构、权重和版本5个维度对生成模型进行了细致划分。Park和Owens(2025)指出,训练数据多样性不足是限制检测器泛化能力的主要因素。为此,他们构建了Community Forensics数据集,包含270万幅图像,覆盖4 803个生成模型,在场景内容、生成架构及图像处理设置上均具有高度多样性。

关于图像内容和格式,一些数据集如DeepArt(Wang等,2023a)和DDDB(deepart detection database)(Wang等,2023b)专用于生成式艺术图像检测;部分数据集如Artifact(Rahman等,2023)和ImagiNet(Boychev和Cholakov,2025)包含艺术图像、场景图像和物体图像等多种图像。还有数据集关注到图像分辨率对检测结果的影响,如DMimageDetection(Corvi等,2023b)对检测器在256×256像素、512×512像素以及1 024×1 024像素这3种图像尺寸上的性能进行了评估。为增加数据集的多样性,D^3(diffusion-generated deepfake detection dataset)(Baraldi等,2024)包含256×256像素,512×512像素,640×480像素和640×360像素等多种尺寸的生成图像,且使用BMP、GIF、JPEG、TIFF和PNG多种编码和压缩方式模拟真实图像的分布。

表 1 人工智能生成内容检测代表性研究归纳

Table 1 A summary of representative studies on AI-generated image detection

方法	检测模型	检测依据	骨干网络	监督范式	技术手段	访问链接
Marra等人(2018)	GAN	图像隐写分析特征	CNN	有监督	–	*
Zhang等人(2019)	GAN	图像DFT频域伪影	CNN	有监督	–	https://github.com/ColumbiaDVMM/AutoGAN
Marra等人(2019)	GAN	–	CNN	有监督	增量学习	*
CNNSpot(Wang等,2020)	GAN	–	CNN	有监督	数据增强	https://github.com/peterwang512/CNNDetection
Durall等人(2020)	GAN	图像DFT频域特征	SVM/K-Means	有监督/无监督	正则项损失	https://github.com/cc-hpc-itwm/UpConv
T-GD(Jeon等,2020)	GAN	–	CNN	半监督	教师–学生模型	https://github.com/cutz-j/T-GD
Bonettini等人(2020)	GAN	图像DCT系数分布	随机森林	有监督	–	https://github.com/polimi-ispl/icpr-benford-gan
He等人(2021)	GAN	图像再生成残差	CNN	有监督	–	https://github.com/SSAW14/BeyondtheSpectrum
LNP(Liu等,2022)	GAN	图像频域融合特征	CNN	有监督	–	https://github. com/Tangsenghenshou/Detecting-Generated-Images-by-Real-Images
Chandrasegaran等,2022)	GAN	图像颜色信息	CNN	有监督	数据增强	https://github. com/sutd-visual-computing-group/transferable-forensic-features
FingerprintNet (Jeong等,2022b)	GAN	图像FFT频域特征	CNN	有监督	–	*
FusingGlobalandLocal (Ju等,2022)	GAN	图像全局/局部融合特征	CNN	有监督	注意力机制	https://github.com/littlejuyan/FusingGlobalandLocal
LGrad(Tan等,2023)	GAN	图像梯度图特征	CNN	有监督	–	https://github.com/chuangchuangtan/lgrad
UniversalFakeDetect (Ojha等,2023)	GAN, DM	图像预训练模型特征	CLIP+KNN	有监督	余弦相似度	https://github. com/WisconsinAIVision/Universal-FakeDetect
DEFAKE(Sha等,2023)	DM	多模态融合特征	CLIP+MLP	有监督	–	https://github.com/zeyangsha/De-Fake
DIRE(Wang等,2023c)	DM	图像重建误差	CNN	有监督	–	https://github.com/ZhendongWang6/DIRE
NPR(Tan等,2024a)	GAN, DM	图像邻近像素关系	CNN	有监督	–	https://github. com/chuangchuangtan/NPR-DeepfakeDetection
FatFormer(Liu等,2024a)	GAN, DM	图像像素域/频域融合特征	CLIP	有监督	对比学习	https://github.com/Michel-liu/FatFormer
Sarkar等人(2024)	DM	图像影射几何特征	CNN	有监督	–	*
AEROBLADE (Ricker等,2024b)	DM	图像重建距离	–	–	SSIM	https://github.com/jonasricker/aeroblade
FakeInversion (Cazenavette等,2024)	DM	原始图+噪声图+重建图特征	CNN	有监督	–	https://fake-inversion.github.io/
DRCT(Chen等,2024a)	DM	真实/生成原始图+重建图	CNN/CLIP	有监督	对比学习	https://github.com/beibuwandeluori/DRCT
FreqNet(Tan等,2024c)	GAN	图像FFT频域融合特征	CNN	有监督	频域学习模块	https://github. com/chuangchuangtan/FreqNet-DeepfakeDetection
ZeroFake(Sha等,2024)	DM	–	–	–	–	https://github.com/TrustAIRLab/ZeroFake
ZED(Cozzolino等,2024b)	GAN, DM	–	–	–	–	https://github.com/grip-unina/ZED
RINE(Koutlis和Papadopoulos,2024)	GAN, DM	图像预训练模型特征	CLIP	有监督	对比学习	https://github.com/mever-team/rine
CoDE(Baraldi等,2024)	DM	图像全局/局部特征相似性	ViT	有监督	对比学习	https://github.com/aimagelab/CoDE
SemGIR(Yu等,2024)	GAN, DM	图像重建+多模态融合特征	CNN	有监督	–	*
Zheng等人(2024)	GAN, DM	图像语义特征抑制	CNN	有监督	–	https://github.com/Zig-HS/FakeImageDetection
C2P-CLIP(Tan等,2025)	GAN, DM	预训练模型特征	CLIP	有监督	对比学习	https://github. com/chuangchuangtan/C2P-CLIP-DeepfakeDetection
AIDE(Yan等,2025)	GAN, DM	图像空间与语义融合特征	CNN	有监督	–	https://github.com/shilinyan99/AIDE
SAFE(Li等,2025c)	GAN, DM	图像局部相关性特征	CNN	有监督	数据变换	https://github.com/Ouxiang-Li/SAFE
MaskCLIP(Wang等,2025)	GAN, DM	图像预训练模型特征	CLIP	有监督	提示调优	https://github.com/iamwangyabin/OpenSDI
AIGI-Holmes (Zhou等,2025)	GAN, DM	CLIP特征/NPR特征	LLaVA	有监督	直接偏好优化	https://github.com/wyczzy/AIGI-Holmes

注:“–”表示不涉及该项内容。“*”表示未公开源码。

表2　通用人工智能生成图像检测数据集

Table 2　The description of datasets for general AI-generated image detection

名称	图像内容	规模	生成模型			真实图像	多模态	多任务	访问链接
			GAN	DM	其他				
Forensic Synthetics（Wang等，2020）	物体	790 k	11	–	–	√	×	×	https://github.com/peterwang512/CNNDetection
IEEE VIP Cup*（Verdoliva等，2022）	物体	14 k	5	6	3	√	×	×	*
Genimage（Zhu等，2023b）	通用	2.7 M	1	7	–	√	×	×	https://github.com/GenImage-Dataset/GenImage
Fake2M（Lu等，2023）	通用	2.3 M	1	9	1	×	×	×	https://github.com/Inf-imagine/Sentry
TWIGMA（Chen和Zou，2023）	通用	800 k	–	–	–	√	×	×	https://yiqunchen.github.io/TWIGMA/
DEFAKE*（Sha等，2023）	通用	222 k	–	4	–	√	√	√	*
UniversalFakeDetect（Ojha等，2023）	通用	870 k	11	7	1	√	×	×	https://github. com/WisconsinAIVision/UniversalFakeDetect
DiffusionForensics（Wang等，2023c）	通用	615 k	–	11	–	√	×	×	https://github.com/ZhendongWang6/DIRE
DiffusionDB（Wang等，2023d）	通用	14 M	–	1	–	×	√	×	https://github.com/poloclub/diffusiondb
DMimageDetection（Corvi等，2023b）	通用	421 k	4	8	–	√	×	×	https://github.com/grip-unina/DMimageDetection
Artifact（Rahman等，2023）	混合	2.5 M	13	7	5	√	×	×	https://github.com/awsaf49/artifact/
Synthbuster（Bammey，2024）	通用	10 k	–	9	–	×	√	×	https://github.com/qbammey/synthbuster
DeepArt*（Wang等，2023a）	艺术	–	–	5	–	√	×	×	*
DDDB*（Wang等，2023b）	艺术	138 k	–	4	1	√	√	×	*
COCOFake（Amoroso等，2024）	物体	1.2 M	–	2	–	√	√	×	https://github.com/aimagelab/COCOFake
CiFAKE（Bird和Lotfi，2024）	物体	120 k	–	1	–	√	×	×	https://github.com/andrew-lin-chang/cifake
D^3（Baraldi等，2024）	通用	12 M	–	12	–	√	√	×	https://github.com/aimagelab/CoDE
FOSID（Karageorgiou等，2024）	通用	1.7 k	–	–	–	×	×	√	https://github.com/mever-team/fosid
LSUNDB（Ricker等，2024a）	通用	500 k	5	5	–	√	×	×	https://github. com/jonasricker/diffusion-model-deepfake-detection
DIF（Sinitsa和Fried，2024）	通用	169 k	56	6	–	√	×	×	https://github.com/Sergo2020/DIF_pytorch_official
AIGCBenchmark（Zhong等，2024）	通用	870 k	7	9	–	√	×	×	https://github.com/Ekko-zn/AIGCDetectBenchmark
WildRF（Cavia等，2024）	通用	5.3 k	–	–	–	√	×	×	https://github.com/barcavia/RealTime-Deepfake Detection-in-the-RealWorld
ImagiNet（Boychev和Cholakov，2025）	混合	200 k	3	5	–	√	×	√	https://github.com/delyan-boychev/imaginet
AntifakePrompt（Chang等，2024）	通用	306 k	–	9	5	√	√	×	https://github.com/nctu-eva-lab/AntifakePrompt
COCOXGEN（Moeßner和Adel，2024）	物体	5.3 k	–	2	–	√	√	×	https://github.com/heikeadel/cocoxgen
FakeBench*（Li等，2025d）	通用	6 k	2	7	1	√	√	√	测试集 https://github.com/Yixuanli423/FakeBench
WildFake（Hong等，2025）	通用	3.6 M	6	13	4	√	×	×	https://github. com/hy-zpg/AIGC-Image-Detection-Dataset
SID-Set（Huang等，2025）	混合	300 k	–	1	–	√	×	√	https://github.com/hzlsaber/SIDA
OpenSDID（Wang等，2025）	通用	300 k	–	5	–	√	×	√	https://github.com/iamwangyabin/OpenSDI
Community Forensics（Park和Owens，2025）	通用	2.7 M	–	–	–	×	×	×	https://github.com/JeongsooP/Community-Forensics
Holmes-Set（Zhou等，2025）	通用	69 k	–	–	–	×	√	√	https://github.com/wyczzy/AIGI-Holmes

注：“–”表示不涉及该项内容；“*”表示非公开数据集；“√”和“×”分别表示具备和不具备该特性。

关于待检测图像源模型的类型，一些数据集（如Forensic Synthetics（Wang 等，2020））专用于 GAN 生成图像的检测；一些数据集（如 Synthbuster（Bammey，2024）、D^3（Baraldi 等，2024）、SID-Set（Huang 等，2025）和 OpenSDID（Wang 等，2025））专用于扩散模型生成图像检测。此外，也有一些数据集（Verdoliva 等，2022；Lu 等，2023；Ojha 等，2023；Rahman 等，2023；Li 等，2025d；Hong 等，2025）可同时用于检测 GAN、扩散模型和自回归模型的生成图像。

为排除数据集自身偏差对检测结果的影响，一些工作致力于构建无偏数据集，以提高检测方法的性能。Grommelt 等人（2024）提出，数据集本身具有的偏差信息可能会导致在该数据集上训练的检测器在预测图像来源时具备一定的倾向性。据此，他们构造了无偏 GenImage 数据集，保证自然图像和生成图像的压缩方式和尺寸一致，在此基础上对部分检测方法的性能进行了测试，发现在无偏数据集上训练的检测器展现出更高的鲁棒性和泛化性。为降低现有数据集中真实图像和生成图像分布差异对检测结果的影响，Rajan 等人（2025）使用自编码器重建真实图像得到生成图像，以对齐真实图像和生成图像特征，提高检测性能和计算效率。类似地，Guillaro 等人（2025）提出 B-Free 无偏训练框架，通过自条件重建和图像修复等方式构造对齐数据集，以降低两类图像的格式和语义差异对检测结果的影响，然后微调 ViT 预训练模型得到检测器。

此外，Lu 等人（2023）提出 Fake2M 数据集，并分别对人类和模型的检测能力进行了评估。Moeßner 和 Adel（2024）发现不同详细程度的提示词对于检测结果有影响：越详细，生成的图像越容易被检测，为此构建了 COCOXGEN 数据集。Li 等人（2025d）提出 FakeBench 数据集，实现对生成图像的检测、解释、推理和细粒度伪造分析。Huang 等人（2025）针对社交媒体图像，构建了 SID-Set 数据集，旨在评估检测方法对于生成图像的检测、篡改定位与解释能力。为应对开放世界环境下的检测和篡改定位挑战，Wang 等人（2025）构建 OpenSDID 数据集，充分考虑了开放世界环境下识别人工智能生成内容的 3 项核心需求：用户多样性、模型创新性以及操作与篡改范围的广度。为解决 AI 生成图像检测模型可解释性不足与泛化能力有限的核心挑战，Zhou 等人（2025）构建了首个大规模、富含语义解释并包含人类偏好对齐标注的 Holmes-Set 基准数据集。该数据集突破传统二元标注模式，采用“多专家评审团”标注机制，在低层视觉与高层语义层面提供了可验证的人类解释，为训练具备可解释性与强泛化能力的检测模型奠定了数据基础。

3.2 面向检测方法的综合评测

基于上述基准数据集，一些工作致力于对检测方法进行综合评测。Gragnaniello 等人（2021）对 2021 年之前的 7 种方法进行了横向比较。Corvi 等人（2023a）分析发现，生成模型无法模拟真实图像的中高频信号，并且径向频谱、角度频谱和 Fisher 判别比也存在一定差异，Corvi 等人（2023b）进一步对部分检测方法进行了评估，发现仅在 GAN 生成图像上训练的检测器不能较好地泛化到 DM 生成图像检测，且检测器对架构相近的生成模型具备较高的检测性能。进一步地，Ricker 等人（2024a）发现，基于 DM 生成图像训练的检测器能够更好地泛化到 GAN 生成图像检测，反映出 DM 生成图像相较于 GAN 生成图像含有更少可用于检测的痕迹。Epstein 等人（2023）对检测器在模拟在线场景下的检测能力进行了评估，发现当生成模型的网络架构相似时，检测器展现出更好的泛化性。Schinas 和 Papadopoulos（2024）提出 SIDBench，使用 Forensic Synthetics、UniversalFakeDetect 和 Synthbuster 组合数据集实现了对 11 种检测方法的比较。Zhong 等人（2024）提出 AIGCBenchmark，对 8 种通用 AIGI 检测方法进行准确性、鲁棒性和泛化性测试，测试集包括 12 种模型的生成图像，涵盖 GAN 和 SD（Stable Diffusion）模型。Guo 等人（2025）提出 ImageDetectBench，包含 8 种主被动检测方法的横向比较。Ha 等人（2024）聚焦艺术图像，对当前检测方法和人类对艺术作品真实性的鉴别能力进行了综合评估。为评估检测方法在真实场景下的鲁棒性，Li 等人（2025a）构建了首个面向现实鲁棒性评测的数据集 RRDataset，从多场景泛化、网络传输鲁棒性和重数字化鲁棒性 3 个维度出发，系统模拟真实世界中的复杂干扰。基于该数据集，作者进一步建立了 RRBench，对 17 种检测器和 10 种视觉语言模型进行了系统评估。结果显示，现有方法在传输和重数字化场景下性能显著下降，揭示了其在现实条件中的脆弱性。同时，大规模人类实验表明，人类观察者虽受图像劣化影响，但仍具备较强的适应性与小样本学习能力。该研

究强调了在评估框架中引入现实干扰因素的必要性，为AI生成图像检测算法的实际应用提供了重要参考。

4 评估维度

实际应用中，检测方法面临三大挑战。1）检测精度难以保证；2）待测图像可能由训练集以外的未知模型生成或图像语义内容有较大差异；3）待测图像可能遭受过不同程度的后处理。这三大挑战分别对应3个评估维度，下面分别进行介绍。

4.1 域内（In-Domain）准确性

域内准确性旨在衡量检测方法在训练集涵盖的模型生成内容上的检测性能，主要的评价指标包括准确率（accuracy）、精确率（precision）、召回率（recall）以及平均精度（average precision，AP），各指标分别计算为

$$Acc = (TP+TN)/(P+N) \tag{1}$$

$$Pre = TP/(TP+FP) \tag{2}$$

$$Rec = TP/(TP+FN) \tag{3}$$

$$AP = \int_0^1 Pre(Rec)\mathrm{d}Rec \tag{4}$$

式中，TP和TN分别表示预测正确的正样本和负样本数量，P和N分别表示正样本和负样本总数量。FP表示真实为负但被错误预测为正的样本数量。FN表示真实为正但被错误预测为负的样本数量。AP表示精确率—召回率（PR）曲线下的面积，在实际计算中通常通过对离散PR曲线进行数值积分近似获得。

4.2 域外（Cross-Domain）泛化性

如今，生成模型发展态势迅猛，新型网络架构和训练手段不断涌现，很大程度上增加了模型生成图像的多样性。为有效提升检测性能，检测器需要基于大量真实图像和生成图像进行训练。然而实际场景下，待测图像可能由训练集以外的模型生成。泛化性旨在衡量检测方法的通用性，即检测方法是否能够在训练集之外的未知模型和未知类别等方面仍然达到很好的检测性能。

当前，跨生成模型（cross-generator）泛化性测试的常规模式是“train-on-one and test-on-many”，即使用单个模型生成的图像集训练检测器，并在多个未知模型的生成图像集上测试。

表3总结了该模式下常用的训练和测试模型，其中训练模型主要选取ProGAN（Karras等，2018）。由此可见，跨生成模型泛化性，不仅包括GAN架构下不同模型的泛化（如ProGAN生成图像到StyleGAN的生成图像的泛化），而且包括GAN架构到DM架构的泛化（如ProGAN生成图像到Stable Diffusion生成图像的泛化）。不同于上述模式，还有些研究工作（Yang等，2024）使用“train-on-many and test-on-many”的泛化性测试方法，即将检测器在多个模型生成的图像集上测试，并测试其在多个未知模型的生成图像集上的检测性能。

表3 跨生成模型泛化性测试常用模型汇总

Table 3 Common generators for cross-generator generalization evalution

检测方法	训练模型	测试模型
CNNSpot（Wang等，2020）	ProGAN	GAN
LNP（Liu等，2022）	ProGAN	GAN、Glow等
FingerprintNet（Jeong等，2022b）	ProGAN	AutoGAN等
UniversalFakeDetect（Ojha等，2023）	ProGAN	GAN、deepfakes、DM
LGrad（Tan等，2023）	ProGAN	GAN、deepfakes
Khan和Dang-Nguyen（2024）	ProGAN	GAN、DM、商业模型
DRCT（Chen等，2024a）	SDv1.4	GAN、DM
FatFormer（Liu等，2024a）	ProGAN	GAN、deepfakes、DM
SAGNet（Tao等，2025）	ProGAN/DM	GAN、DM

跨类别（cross-category）泛化性测试，指将检测器在固定类别的图像集上训练，然后在其他类别图像上的检测效果。例如，Jeong等人（2022b）选取“马”相关的图像集训练检测器，然后在“苹果”相关的图像集上进行测试；Tan等人（2023）选取“马”相关的图像集训练检测器，然后在其余类别的图像集上进行测试。

Tao等人(2025)在“马”和“椅子”两类图像上训练检测器,然后在20个类别图像上进行测试。

此外,还有一些工作对检测方法的跨颜色泛化性(Jeong等,2022b)、跨数据集(cross-dataset)泛化性(Sha等,2023)、跨概念(cross-concept)泛化性(Dogoulis等,2023)和跨场景(cross-scene)泛化性(Zheng等,2024)进行了测试,以更加全面地评估检测方法的性能。

4.3　鲁棒性

鲁棒性旨在衡量检测方法是否能够抵抗图像后处理等操作,其评估方法是处理后的图像作为输入查看检测结果。常见的图像处理手段包括JPEG压缩、高斯模糊、裁剪、尺寸变换以及高斯噪声,还有一些工作使用WEBP压缩、中值滤波、Gamma校正和对抗样本攻击等后处理操作评估检测方法的鲁棒性,如表4所示。

表4　鲁棒性测试常用操作及代表性工作汇总

Table 4　Common operations and representative works for robustness evalution

图像处理操作	代表性研究工作
JPEG压缩	Wang等人(2020)、Frank等人(2020)、Liu等人(2022)、Tan等人(2023)、Liu等人(2024a)、Khan和Dang-Nguyen(2024)、Chen等人(2024a)、Ricker等人(2024b)、Yan等人(2025)、Li等人(2025c)、Chu等人(2025)、Karageorgiou等人(2025)
高斯模糊	Wang等人(2020)、Frank等人(2020)、Liu等人(2022)、Tan等人(2023)、Liu等人(2024a)、Khan和Dang-Nguyen(2024)、Ricker等人(2024b)、Yan等人(2025)、Li等人(2025c)、Chu等人(2025)、Karageorgiou等人(2025)
裁剪	Frank等人(2020)、Liu等人(2022)、Tan等人(2023)、Liu等人(2024a)、Ricker等人(2024b)、Koutlis和Papadopoulos(2024)、Chu等人(2025)
尺寸变换	Liu等人(2022)、Tan等人(2023)、Chen等人(2024a)、Karageorgiou等人(2025)
高斯噪声	Frank等人(2020)、Liu等人(2024a)、Ricker等人(2024b)、Chu等人(2025)、Karageorgiou等人(2025)
WEBP压缩	Li等人(2024)、Karageorgiou等人(2025)
对抗样本攻击	Sha等人(2023,2024)、de Rosa等人(2024)、Zhou等人(2024)

鉴于不同的后处理操作程度对图像质量的影响不同,一些工作在评估检测方法的鲁棒性时,考虑多种后处理操作程度,如不同压缩因子的JPEG压缩、不同模糊程度的高斯模糊以及不同裁剪尺寸的裁剪操作。

5　方法比较与分析

5.1　实验设置

为实现对各类检测方法性能的综合横向评估,本文选取基于像素特征的CNNSpot(Wang等,2020)、基于图像频域特征的FreDect(Frank等,2020)、基于预训练模型特征的UnivFD(Ojha等,2023)、基于图像梯度图特征的LGrad(Tan等,2023)、基于图像内部相关性的NPR(neighboring pixel relationships)(Tan等,2024a)、基于图像频域特征的FreqNet(Tan等,2024c)、基于图像像素域与频域融合特征的FatFormer(Liu等,2024a)、基于图像局部特征的SSP(Chen等,2024f)、基于图像重建特征的DRCT(diffusion reconstruction contrastive training)(Chen等,2024a)、基于预训练模型特征的C2P-CLIP(Tan等,2025)、基于图像空间与语义融合特征的AIDE(Yan等,2025)和基于图像伪影特征与局部相关性增强的SAFE(Li等,2025c)共12种检测方法进行比较,衡量上述检测方法在AIGCBenchmark基准数据集的各测试集上的检测性能。

5.2　性能分析

表5展示了所选检测方法在不同生成图像测试集的检测准确率结果。数据来源https://github.com/graydove/Awesome-AIGC-Image-Detection。由此可见,不同检测方法的泛化性有较大差异。其中,CNNSpot在ProGAN和WFIR测试集上能够达到最优的检测结果,Fatformer、DRCT和AIDE在多种测试模型生成图像集上能够达到较好的检测结果,SAFE的综合检测结果最优。

不同骨干网络的选取对于检测结果也有较大影

响，例如，基于CNN的DRCT测试性能优于基于CLIP预训练模型的DRCT测试性能。此外，基于CNN的SAFE综合检测性能较高，说明卷积神经网络在检测领域仍然具有一定优势，且由于模型规模相对较小，对实时性检测更加友好。

不同训练集的选取也会对检测结果产生影响，例如，AIDE在ProGAN生成图像集上训练得到检测效果要优于在SDv1.4生成图像集上训练得到的检

表5　主流人工智能生成图像检测方法性能对比

Table 5　Comparison on mainstream AI-generated image detection methods

/%

方法	ProGAN	StyleGan	BigGAN	CycleGAN	StarGAN	GauGAN	Stylegan2	WFIR	ADM
CNNSpot	**100.00**	90.17	71.17	87.59	94.60	81.44	86.91	**95.10**	60.38
FreDect	99.36	78.02	81.97	78.77	94.62	80.57	66.17	46.45	64.67
UnivFD	99.81	84.93	95.08	98.33	95.75	**99.47**	74.96	87.20	66.94
LGrad	99.86	89.72	84.58	87.81	99.30	82.26	85.30	52.95	64.12
NPR	99.79	97.71	84.35	96.10	99.35	82.50	98.39	65.75	69.72
FreqNet	99.58	90.24	90.45	95.84	85.69	93.41	87.95	49.20	83.27
FatFormer	99.89	97.13	**99.50**	**99.36**	99.75	99.43	**98.80**	88.16	78.44
SSP	84.01	83.14	74.35	65.37	89.59	58.82	86.23	70.30	93.15
DRCT-C	54.79	57.09	55.55	49.89	51.13	49.13	52.05	50.05	61.78
DRCT-U	76.85	71.58	82.67	94.51	63.08	78.39	67.73	63.10	79.43
C2P-CLIP	99.98	96.44	99.12	97.31	99.60	99.17	95.59	94.80	77.12
AIDE-G	99.99	**99.65**	83.95	98.49	**99.90**	73.24	98.01	94.20	**93.46**
AIDE-D	69.28	71.05	77.20	74.38	80.24	64.36	72.53	72.50	78.54
SAFE	99.86	98.04	89.72	98.87	**99.90**	91.52	98.57	51.95	82.05

方法	Glide	Midjourney	SDv1.4	SDv1.5	VQDM	Wukong	DALLE2	平均
CNNSpot	58.03	51.40	50.58	50.52	56.45	51.03	51.25	71.04
FreDect	55.43	46.89	40.02	40.39	78.95	41.54	34.65	64.28
UnivFD	62.53	56.24	63.75	63.57	85.42	71.06	50.75	78.49
LGrad	70.30	67.09	63.78	65.08	72.62	60.14	68.55	75.84
NPR	78.35	77.83	78.62	78.88	78.13	76.10	64.90	82.90
FreqNet	81.66	69.83	64.22	64.91	81.65	57.72	55.30	78.18
FatFormer	88.03	56.08	67.83	68.06	86.87	73.06	69.67	85.63
SSP	**98.77**	84.21	99.17	99.08	96.09	98.55	96.30	86.07
DRCT-C	65.92	94.63	**99.88**	**99.82**	74.88	**99.91**	73.20	68.11
DRCT-U	89.12	91.51	95.03	94.40	90.02	94.65	92.60	82.79
C2P-CLIP	88.48	59.12	82.76	82.73	87.16	80.12	65.10	87.79
AIDE-G	95.10	77.21	92.99	92.85	95.16	93.53	**96.60**	92.77
AIDE-D	91.83	79.38	99.74	99.76	80.26	98.66	95.00	81.54
SAFE	96.29	**95.27**	99.41	99.27	**96.29**	98.21	95.30	**93.16**

注：加粗字体表示各列最优结果。DRCT-C表示DRCT-Conv-B，DRCT-U表示DRCT-UnivFD；AIDE-G表示AIDE_ProGAN，AIDE-D表示AIDE_SDv1.4。

测效果。整体而言，基于融合特征的检测方法能够取得较高的检测准确性和泛化性。

6 总结和展望

6.1 总结

为使相关研究人员快速掌握通用人工智能生成图像检测领域研究现状，把握相关研究发展脉络，本文对该领域的研究方法进行全面总结与分析。

具体而言，首先通过深度解构生成模型的演进脉络，从技术原理层面剖析生成对抗网络、自回归模型与扩散模型等主流生成范式的架构特征与生成机理，从而为检测方法的设计确立明确的技术锚点。之后，从监督范式、检测依据、骨干网络和技术手段等多个维度构建检测技术体系，并将检测依据作为核心分类标准，详细阐述了基于像素域特征、频域特征、预训练模型特征、融合特征和特定规则等各类检测方法的核心思想与特点。最后，为便于同行开展研究，本文构建了完整的评估体系，首先对当前主流基准数据集进行汇总，并从图像类型、数据规模和生成模型覆盖度等维度对各基准数据集进行详细介绍与分类。然后，给出该研究领域常用评估维度，并对十余种代表性检测方法进行系统性评测与分析。

本综述不仅为领域内学者提供了全景式技术路线图，更为构建下一代高鲁棒、可扩展和可解释的人工智能生成图像检测系统奠定了理论基础。

6.2 展望

尽管当前检测方法在生成图像检测领域取得了巨大成就，但仍存在一些值得研究的挑战和难题。针对如何进一步提升检测性能，本文对未来的研究课题进行展望。

1）用于训练检测模型的大规模无偏数据集构建。当前研究主要聚焦检测方法设计，然而现有研究表明，训练数据自身的偏差会使得检测模型的鲁棒性和泛化性降低，因此，如何充分挖掘和利用现有训练数据，以最大程度降低训练数据噪声对检测结果的影响，是检测领域十分具有挑战性的关键问题。

2）抵御物理层面攻击的高鲁棒检测方法设计。当前，对检测方法的鲁棒性测试仍停留在数字层面，即JPEG压缩、高斯模糊和尺寸变换等对图像的常规后处理操作。然而，检测方法是否能够抵抗打印、社交网络传播等现实场景下的后处理操作是值得探索的研究方向。

3）面向多种生成模型的可扩展检测方法设计。随着生成模型不断涌现，对检测方法泛化性的需求日益提高，在生成式人工智能时代，如何使得检测方法跟上生成模型快速发展的步伐，并在此基础上保持良好的检测性能，具有重要的研究价值和意义。

4）依托大语言模型生成的可解释检测方法设计。当前人工智能生成图像检测任务多被视为一个二分类问题，如何进一步给出检测依据，提高检测结果的可解释性具有重要的探索价值。这不仅能够促进图像生成模型的技术发展，而且能够反向推动人工智能生成图像检测的技术进步。

5）适配于实际应用场景的实用检测方法设计。人工智能生成图像遍布社交网络与新闻媒体，为司法取证等实际场景下的图像真实性鉴别需求带来了重大挑战。因此，如何提高检测方法在实际应用场景下的适用性至关重要。此外，在大规模计算（如数据中心）和嵌入式计算（如移动终端）等对检测效率有较高需求的场景下，如何设计轻量级检测方法，以保证检测技术的实时响应能力是有待解决的关键问题。

6）针对检测方法的反取证攻击研究。一些研究陆续关注针对检测方法的反取证攻击。与检测方法相反，反取证攻击旨在使得生成图像逃避当前检测方法的识别，取证与反取证是两个相辅相成的研究方向，二者在对抗博弈中共同发展，因此，针对检测方法的反取证攻击也是值得持续关注的研究问题。

参考文献（References）

Aha D W. 1997. Lazy Learning. Dordrecht：Springer［DOI：10.1007/978-94-017-2053-3］

Amoroso R，Morelli D，Cornia M，Baraldi L，Del Bimbo A and Cucchiara R. 2024. Parents and children：distinguishing multimodal deepfakes from natural images. ACM Transactions on Multimedia Computing，Communications and Applications，21（1）：#11［DOI：10.1145/3665497］

Bammey Q. 2024. Synthbuster：towards detection of diffusion model generated images. IEEE Open Journal of Signal Processing，5：1-9［DOI：10.1109/OJSP.2023.3337714］

Baraldi L，Cocchi F，Cornia M，Baraldi L，Nicolosi A and Cucchiara R. 2024. Contrasting deepfakes diffusion via contrastive learning and global-local similarities//Proceedings of the 18th European Conference on Computer Vision. Milan，Italy：Springer：199-216

[DOI: 10.1007/978-3-031-73036-8_12]

Betker J, Goh G, Jing L, Brooks T, Wang J F, Li L J, et al. 2023. Improving image generation with better captions. Computer Science, 2(3): #8

Bi X L, Liu B, Yang F, Xiao B, Li W S, Huang G, et al. 2023. Detecting generated images by real images only [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2311.00962.pdf

Bird J J and Lotfi A. 2024. CIFAKE: image classification and explainable identification of AI-generated synthetic images. IEEE Access, 12: 15642-15650 [DOI: 10.1109/ACCESS.2024.3356122]

Bonettini N, Bestagini P, Milani S and Tubaro S. 2020. On the use of Benford's law to detect GAN-generated images//Proceedings of the 25th International Conference on Pattern Recognition. Milan, Italy: IEEE: 5495-5502 [DOI: 10.1109/ICPR48806.2021.9412944]

Boychev D and Cholakov R. 2025. ImagiNet: a multi-content benchmark for synthetic image detection [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2407.20020v3.pdf

Brock A, Donahue J and Simonyan K. 2019. Large scale GAN training for high fidelity natural image synthesis//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA

Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. 2020. Language models are few-shot learners//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 1877-1901 [DOI: 10.5555/3495724.3495883]

Cao S H, Liu X H, Mao X Q and Zou Q. 2022. A review of human face forgery and forgery-detection technologies. Journal of Image and Graphics, 27(4): 1023-1038 (曹申豪, 刘晓辉, 毛秀青, 邹勤. 2022. 人脸伪造及检测技术综述. 中国图象图形学报, 27(4): 1023-1038) [DOI: 10.11834/jig.200466]

Cavia B, Horwitz E, Reiss T and Hoshen Y. 2024. Real-time deepfake detection in the real-world [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2406.09398.pdf

Cazenavette G, Sud A, Leung T and Usman B. 2024. FakeInversion: learning to detect images from unseen text-to-image models by inverting stable diffusion//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10759-10769 [DOI: 10.1109/CVPR52733.2024.01023]

Chandrasegaran K, Tran N T, Binder A and Cheung N M. 2022. Discovering transferable forensic features for CNN-generated images detection//Proceedings of the 16th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 671-689 [DOI: 10.1007/978-3-031-19784-0_39]

Chandrasegaran K, Tran N T and Cheung N M. 2021. A closer look at Fourier spectrum discrepancies for CNN-generated images detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 7196-7205 [DOI: 10.1109/CVPR46437.2021.00712]

Chang H W, Zhang H, Barber J, Maschinot A J, Lezama J, Jiang L, et al. 2023. Muse: text-to-image generation via masked generative transformers//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: 4055-4075 [DOI: 10.5555/3618408.3618570]

Chang Y M, Yeh C, Chiu W C and Yu N. 2024. AntifakePrompt: prompt-tuned vision-language models are fake image detectors [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2310.17419v3.pdf

Chen B Y, Zeng J S, Yang J Q and Yang R. 2024a. DRCT: diffusion reconstruction contrastive training towards universal detection of diffusion generated images//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: JMLR.org: 7621-7639 [DOI: 10.5555/3692070.3692367]

Chen J H, Lo M, Liao H L and Huang T L. 2024b. IPD-Net: detecting AI-generated images via inter-patch dependencies. International Journal of Advanced Computer Science and Applications, 15(7): 1372-1379 [DOI: 10.14569/IJACSA.2024.01507133]

Chen J S, Ge C J, Xie E Z, Wu Y, Yao L W, Ren X Z, et al. 2024c. PIXART-Σ: weak-to-strong training of diffusion transformer for 4K text-to-image generation//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 74-91 [DOI: 10.1007/978-3-031-73411-3_5]

Chen J S, Wu Y, Luo S M, Xie E Z, Paul S, Luo P, et al. 2024d. PIXART-δ: fast and controllable image generation with latent consistency models [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2401.05252.pdf

Chen J S, Yu J C, Ge C J, Yao L W, Xie E Z, Wu Y, et al. 2024e. PixArt-α: fast training of diffusion transformer for photorealistic text-to-image synthesis//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria

Chen J X, Yao J T and Niu L. 2024f. A single simple patch is all you need for AI-generated image detection [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2402.01123v2.pdf

Chen M, Radford A, Child R, Wu J, Jun H, Luan D, et al. 2020. Generative pretraining from pixels//Proceedings of the 37th International Conference on Machine Learning. Virtual: JMLR.org: 1691-1703 [DOI: 10.5555/3524938.3525096]

Chen Y M and Yashtini M. 2024. Detecting AI generated images through texture and frequency analysis of patches//Proceedings of the 4th International Conference on Artificial Intelligence, Virtual Reality and Visualization. Nanjing, China: IEEE: 103-110 [DOI: 10.1109/AIVRV63595.2024.10860248]

Chen Y T and Zou J. 2023. TWIGMA: a dataset of AI-generated images with metadata from twitter//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 37748-37760 [DOI: 10.5555/3666122.3667764]

Choi S, Park S, Lee J, Kim S, Choi S J and Lee M. 2024. HFI: a unified framework for training-free detection and implicit watermarking

of latent diffusion model generated images [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2412.20704.pdf

Choi Y, Choi M, Kim M, Ha J W, Kim S and Choo J. 2018. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8789-8797 [DOI: 10.1109/CVPR.2018.00916]

Chu B L, Xu X, Wang X, Zhang Y F, You W K and Zhou L N. 2025. FIRE: robust detection of diffusion-generated images via frequency-guided reconstruction error//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 12830-12839 [DOI: 10.1109/CVPR52734.2025.01197]

Coccomini D A, Caldelli R, Gennaro C, Fiameni G, Amato G and Falchi F. 2024. Deepfake detection without deepfakes: generalization via synthetic frequency patterns injection [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2403.13479.pdf

Corvi R, Cozzolino D, Poggi G, Nagano K and Verdoliva L. 2023a. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver, Canada: IEEE: 973-982 [DOI: 10.1109/CVPRW59228.2023.00104]

Corvi R, Cozzolino D, Zingarini G, Poggi G, Nagano K and Verdoliva L. 2023b. On the detection of synthetic images generated by diffusion models//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island, Greece: IEEE: 1-5 [DOI: 10.1109/ICASSP49357.2023.10095167]

Cozzolino D, Poggi G, Corvi R, Nießner M and Verdoliva L. 2024a. Raising the bar of AI-generated image detection with CLIP//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE: 4356-4366 [DOI: 10.1109/CVPRW63382.2024.00439]

Cozzolino D, Poggi G, Nießner M and Verdoliva L. 2024b. Zero-shot detection of AI-generated images//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 54-72 [DOI: 10.1007/978-3-031-72649-1_4]

Dai W L, Li J N, Li D X, Tiong A M H, Zhao J Q, Wang W S, et al. 2023. InstructBLIP: towards general-purpose vision-language models with instruction tuning//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 49250-49267 [DOI: 10.5555/3666122.3668264]

Deng J Y, Lin C H, Zhao Z Y, Liu S, Wang Q and Shen C. 2025. A survey of defenses against AI-generated visual media: detection, disruption, and authentication. ACM Computing Surveys [DOI: 10.1145/3770916]

de Rosa V, Guillaro F, Poggi G, Cozzolino D and Verdoliva L. 2024. Exploring the adversarial robustness of CLIP for AI-generated image detection//Proceedings of 2024 IEEE International Workshop on Information Forensics and Security. Rome, Italy: IEEE: 1-6 [DOI: 10.1109/WIFS61860.2024.10810719]

Dhariwal P and Nichol A. 2021. Diffusion models beat GANs on image synthesis//Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual: Curran Associates Inc.: 8780-8794 [DOI: 10.5555/3540261.3540933]

Ding M, Yang Z Y, Hong W Y, Zheng W D, Zhou C, Yin D, et al. 2021. CogView: mastering text-to-image generation via transformers//Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual: Curran Associates Inc.: 19822-19835 [DOI: 10.5555/3540261.3541777]

Ding M, Zheng W D, Hong W Y and Tang J. 2022. CogView2: faster and better text-to-image generation via hierarchical transformers//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 16890-16902 [DOI: 10.5555/3600270.3601499]

Dinh L, Krueger D and Bengi Y. 2014. NICE: non-linear independent components estimation//Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada

Dogoulis P, Kordopatis-Zilos G, Kompatsiaris I and Papadopoulos S. 2023. Improving synthetically generated image detection in cross-concept settings//Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation. Thessaloniki, Greece: ACM: 28-35 [DOI: 10.1145/3592572.3592846]

Doloriel C T and Cheung N M. 2024. Frequency masking for universal deepfake detection//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Seoul, Korea: IEEE: 13466-13470 [DOI: 10.1109/ICASSP48485.2024.10446290]

Dong C D, Kumar A and Liu E Y. 2022. Think twice before detecting GAN-generated fake images from their spectral domain imprints//Proceedings of 2022 IEEE/CVF conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 7855-7864 [DOI: 10.1109/CVPR52688.2022.00771]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. Virtual

Durall R, Keuper M and Keuper J. 2020. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions//Proceedings of 2020 IEEE/CVF conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7887-7896 [DOI: 10.1109/CVPR42600.2020.00791]

Dzanic T, Shah K and Witherden F D. 2020. Fourier spectrum discrepancies in deep network generated images//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 3022-3032 [DOI: 10.5555/3495724.3495978]

Epstein D C, Jain I, Wang O and Zhang R. 2023. Online detection of

AI-generated images//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision Workshops. Paris, France: IEEE: 382-392 [DOI: 10.1109/ICCVW60793.2023.00045]

Esser P, Rombach R and Ommer B. 2021. Taming transformers for high-resolution image synthesis//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 12868-12878 [DOI: 10.1109/CVPR46437.2021.01268]

Fan Y M, Yang D M, Zhang J G, Yang B and Zou Y X. 2024. FakeGPT: detecting fake image via large language model//Proceedings of the 7th Chinese Conference on Pattern Recognition and Computer Vision. Urumqi, China: Springer: 122-136 [DOI: 10.1007/978-981-97-8685-5_9]

Frank J, Eisenhofer T, Schönherr L, Fischer A, Kolossa D and Holz T. 2020. Leveraging frequency analysis for deep fake image recognition//Proceedings of the 37th International Conference on Machine Learning. Virtual: JMLR.org: 3247-3258 [DOI: 10.5555/3524938.3525242]

Gaintseva T, Kushnareva L, Magai G, Piontkovskaya I, Nikolenko S, Benning M, et al. 2024. Improving interpretability and robustness for the detection of AI-generated Images [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2406.15035.pdf

Gallagher J and Pugsley W. 2024. Development of a dual-input neural model for detecting AI-generated imagery [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2406.13688.pdf

Goebel M, Nataraj L, Nanjundaswamy T, Mohammed T M, Chandrasekaran S and Manjunath B S. 2021. Detection, attribution and localization of GAN generated images. Electronic Imaging, 33(4): 276-1-276-11 [DOI: 10.2352/ISSN.2470-1173.2021.4.MWSF-276]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. 2014. Generative adversarial nets//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2672-2680 [DOI: 10.5555/2969033.2969125].

Gragnaniello D, Cozzolino D, Marra F, Poggi G and Verdoliva L. 2021. Are GAN generated images easy to detect? A critical analysis of the state-of-the-art//Proceedings of 2021 IEEE International Conference on Multimedia and Expo. Shenzhen, China: IEEE: 1-6 [DOI: 10.1109/ICME51207.2021.9428429]

Grommelt P, Weiss L, Pfreundt F J and Keuper J. 2024. Fake or JPEG? Revealing common biases in generated image detection datasets//Proceedings of the 18th European Conference on Computer Vision Workshops. Milan, Italy: Springer: 80-95 [DOI: 10.1007/978-3-031-92089-9_6]

Gu S Y, Chen D, Bao J M, Wen F, Zhang B, Chen D D, et al. 2022. Vector quantized diffusion model for text-to-image synthesis//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10686-10696 [DOI: 10.1109/CVPR52688.2022.01043]

Guarnera L, Giudice O and Battiato S. 2024. Mastering deepfake detection: a cutting-edge approach to distinguish GAN and diffusion-model images. ACM Transactions on Multimedia Computing, Communications and Applications, 20(11): #343 [DOI: 10.1145/3652027]

Guillaro F, Zingarini G, Usman B, Sud A, Cozzolino D and Verdoliva L. 2025. A bias-free training paradigm for more general AI-generated image detection//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 18685-18694 [DOI: 10.1109/CVPR52734.2025.01741]

Guo M Y, Hu Y P, Jiang Z Y, Li Z Y, Sadovnik A, Daw A, et al. 2025. AI-generated image detection: passive or watermark [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2411.13553v2.pdf

Guo N H, Bethge J, Meinel C and Yang H J. 2022. Join the high accuracy club on ImageNet with a binary neural network ticket [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2211.12933v3.pdf

Gye S, Ko J, Shon H, Kwon M and Kim J. 2025. Reducing the content bias for AI-generated image detection//Proceedings of 2025 IEEE/CVF Winter Conference on Applications of Computer Vision. Tucson, USA: IEEE: 399-408 [DOI: 10.1109/WACV61041.2025.00049]

Ha A Y J, Passananti J, Bhaskar R, Shan S, Southen R, Zheng H T, et al. 2024. Organic or diffused: can we distinguish human art from AI-generated images//Proceedings of 2024 on ACM SIGSAC Conference on Computer and Communications Security. Salt Lake City, USA: ACM: 4822-4836 [DOI: 10.1145/3658644.3670306]

He K M, Chen X L, Xie S N, Li Y H, Dollár P and Girshick R. 2022. Masked autoencoders are scalable vision learners//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 15979-15988 [DOI: 10.1109/CVPR52688.2022.01553]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

He P S, Li W C, Zhang J Y, Wang H X and Jiang X H. 2022. Overview of passive forensics and anti-forensics techniques for GAN-generated image. Journal of Image and Graphics, 27(1): 88-110 (何沛松, 李伟创, 张婧媛, 王宏霞, 蒋兴浩. 2022. 面向GAN生成图像的被动取证及反取证技术综述. 中国图象图形学报, 27(1): 88-110) [DOI: 10.11834/jig.210430]

He Y, Yu N, Keuper M and Fritz M. 2021. Beyond the spectrum: detecting deepfakes via re-synthesis//Proceedings of the 30th International Joint Conference on Artificial Intelligence. Montreal, Canada: Morgan Kaufmann: 2534-2541 [DOI: 10.24963/ijcai.2021/349]

He Z Y, Chen P Y and Ho T Y. 2024. RIGID: a training-free and model-agnostic framework for robust AI-generated image detection [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2405.20112.pdf

Herur A N, Santhosh V, Shetty N and Seelamantula C S. 2024. Addressing diffusion model based counter-forensic image manipulation for synthetic image detection//Proceedings of the 15th Indian Conference on Computer Vision Graphics and Image Processing. Bengaluru, India: ACM: #46 [DOI: 10.1145/3702250.3702296]

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 6840-6851 [DOI: 10.5555/3495724.3496298]

Hong Y, Feng J M, Chen H X, Lan J, Zhu H J, Wang W Q, et al. 2025. WildFake: a large-scale and hierarchical dataset for AI-generated images detection//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI: 3500-3508 [DOI: 10.1609/aaai.v39i4.32363]

Hooda A, Mangaokar N, Feng R, Fawaz K, Jha S and Prakash A. 2024. D4: Detection of adversarial diffusion deepfakes using disjoint ensembles//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 3800-3810 [DOI: 10.1109/WACV57701.2024.00377]

Hu E J, Shen Y L, Wallis P, Allen-Zhu Z Y, Li Y Z, Wang S A, et al. 2022. LoRA: low-rank adaptation of large language models//Proceedings of the 10th International Conference on Learning Representations. Virtual

Huang Z L, Hu J W, Li X T, He Y W, Zhao X Y, Peng B, et al. 2025. SIDA: social media image deepfake detection, localization and explanation with large multimodal model//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 28831-28841 [DOI: 10.1109/CVPR52734.2025.02685]

Javaheri A H, Motamednia H and Mahmoudi-Azanveh A. 2024. Enhancing the generalization of synthetic image detection models through the exploration of features in deep detection models//Proceedings of the 13th Iranian/the 3rd International Machine Vision and Image Processing Conference. Tehran, Islamic Republic of Iran: IEEE: 1-6 [DOI: 10.1109/MVIP62238.2024.10491148]

Jeon H, Bang Y, Kim J and Woo S S. 2020. T-GD: transferable GAN-generated images detection framework//Proceedings of the 37th International Conference on Machine Learning. Virtual: JMLR.org: 4746-4761 [DOI: 10.5555/3524938.3525379]

Jeong Y, Kim D, Min S, Joe S, Gwon Y and Choi J. 2022a. BiHPF: bilateral high-pass filters for robust deepfake detection//Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 2878-2887 [DOI: 10.1109/WACV51458.2022.00293]

Jeong Y, Kim D, Ro Y, Kim P and Choi J. 2022b. FingerprintNet: synthesized fingerprints for generated image detection//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 76-94 [DOI: 10.1007/978-3-031-19781-9_5]

Jia Z X, Huang C W, Zhu Y S, Fei H Y, Duan X Y, Yuan Z Q, et al. 2025. Secret lies in color: enhancing AI-generated images detection with color distribution analysis//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 13445-13454 [DOI: 10.1109/CVPR52734.2025.01255]

Ju Y, Jia S, Cai J L, Guan H Y and Lyu S W. 2024. GLFF: global and local feature fusion for AI-synthesized image detection. IEEE Transactions on Multimedia, 26: 4073-4085 [DOI: 10.1109/TMM.2023.3313503]

Ju Y, Jia S, Ke L P, Xue H F, Nagano K and Lyu S W. 2022. Fusing global and local features for generalized AI-synthesized image detection//Proceedings of 2022 IEEE International Conference on Image Processing. Bordeaux, France: IEEE: 3465-3469 [DOI: 10.1109/ICIP46576.2022.9897820]

Kang M, Zhu J Y, Zhang R, Park J, Shechtman E, Paris S, et al. 2023. Scaling up GANs for text-to-image synthesis//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 10124-10134 [DOI: 10.1109/CVPR52729.2023.00976]

Karageorgiou D, Bammey Q, Porcellini V, Goupil B, Teyssou D and Papadopoulos S. 2024. Evolution of detection performance throughout the online lifespan of synthetic images//Proceedings of 2024 European Conference on Computer Vision 2024——Trust What You learN Workshop. Milan, Italy: Springer: 1-18 [DOI: 10.5281/zenodo.13648239]

Karageorgiou D, Papadopoulos S, Kompatsiaris I and Gavves E. 2025. Any-resolution AI-generated image detection by spectral learning//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 18706-18717 [DOI: 10.1109/CVPR52734.2025.01743]

Karras T, Aila T, Laine S and Lehtinen J. 2018. Progressive growing of GANs for improved quality, stability, and variation//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada

Karras T, Laine S and Aila T. 2019. A style-based generator architecture for generative adversarial networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4396-4405 [DOI: 10.1109/CVPR.2019.00453]

Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J and Aila T. 2020. Analyzing and improving the image quality of StyleGAN//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8107-8116 [DOI: 10.1109/CVPR42600.2020.00813]

Keita M, Hamidouche W, Bougueffa H, Hadid A and Taleb-Ahmed A. 2024a. Harnessing the power of large vision language models for synthetic image detection [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2404.02726.pdf

Keita M, Hamidouche W, Eutamene H B, Taleb-Ahmed A, Camacho

D and Hadid A. 2025. Bi-LORA: a vision-language approach for synthetic image detection. Expert Systems, 42(2): #e13829 [DOI: 10.1111/exsy.13829]

Keita M, Hamidouche W, Eutamene H B, Taleb-Ahmed A and Hadid A. 2024b. FIDAVL: fake image detection and attribution using vision-language model//Proceedings of the 27th International Conference on Pattern Recognition. Kolkata, India: Springer: 160-176 [DOI: 10.1007/978-3-031-78305-0_11]

Khan S A and Dang-Nguyen D T. 2024. CLIPping the deception: adapting vision-language models for universal deepfake detection//Proceedings of 2024 International Conference on Multimedia Retrieval. Phuket, Thailand: ACM: 1006-1015 [DOI: 10.1145/3652583.3658035]

Kingma D P and Dhariwal P. 2018. Glow: generative flow with invertible 1 × 1 convolutions//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc.: 10236-10245 [DOI: 10.5555/3327546.3327685]

Kingma D P and Welling M. 2022. Auto-encoding variational bayes [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/1312.6114.pdf

Konstantinidou D, Koutlis C and Papadopoulos S. 2025. TextureCrop: enhancing synthetic image detection through texture-based cropping//Proceedings of 2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. Tucson, USA: IEEE: 1369-1378 [DOI: 10.1109/WACVW65960.2025.00160]

Koutlis C and Papadopoulos S. 2024. Leveraging representations from intermediate encoder-blocks for synthetic image detection//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 394-411 [DOI: 10.1007/978-3-031-73220-1_23]

Lamichhane D. 2025. Advanced detection of AI-generated images through vision transformers. IEEE Access, 13: 3644-3652 [DOI: 10.1109/ACCESS.2024.3522759]

Lanzino R, Fontana F, Diko A, Marini M R and Cinque L. 2024. Faster than lies: real-time deepfake detection using binary neural networks//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE: 3771-3780: [DOI: 10.1109/CVPRW63382.2024.00381]

Lempitsky V, Vedaldi A and Ulyanov D. 2018. Deep image prior//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 9446-9454 [DOI: 10.1109/CVPR.2018.00984]

Leporoni G, Maiano L, Papa L and Amerini I. 2024. A guided-based approach for deepfake detection: RGB-depth integration via features fusion. Pattern Recognition Letters, 181: 99-105 [DOI: 10.1016/j.patrec.2024.03.025]

Li C X, Wang X X, Li M L, Miao B M, Sun P, Zhang Y J, et al. 2025a. Bridging the gap between ideal and real-world evaluation: benchmarking ai-generated image detection in challenging scenarios [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2509.09172.pdf

Li J, Jiang W T, Shen L Y and Ren Y W. 2025b. Optimized frequency collaborative strategy drives AI image detection. IEEE Internet of Things Journal, 12(11): 16192-16203 [DOI: 10.1109/JIOT.2025.3531053]

Li J and Wang K. 2024. Detecting computer-generated images by using only real images//Proceedings of the 17th International Conference on Machine Vision. Edinburgh, UK: SPIE: 202-214 [DOI: 10.1117/12.3055092]

Li J N, Li D X, Savarese S and Hoi S. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: 19730-19742 [DOI: 10.5555/3618408.3619222]

Li J N, Li D X, Xiong C M and Hoi S. 2022a. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR: 12888-12900

Li O X, Cai J Y, Hao Y B, Jiang X L, Hu Y and Feng F L. 2025c. Improving synthetic image detection towards generalization: an image transformation perspective//Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Toronto, Canada: ACM: 2405-2414 [DOI: 10.1145/3690624.3709392]

Li W C, He P S, Li H L, Wang H X and Zhang R M. 2022b. Detection of GAN-generated images by estimating artifact similarity. IEEE Signal Processing Letters, 29: 862-866 [DOI: 10.1109/LSP.2021.3130525]

Li Y H, Bammey Q, Gardella M, Nikoukhah T, Morel J M, Colom M, et al. 2024. MaskSim: detection of synthetic images by masked spectrum similarity analysis//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE: 3855-3865 [DOI: 10.1109/CVPRW63382.2024.00390]

Li Y X, Liu X L, Wang X Y, Lee B S, Wang S Q, Rocha A, et al. 2025d. FakeBench: probing explainable fake image detection via large multimodal models. IEEE Transactions on Information Forensics and Security, 20: 8730-8745 [DOI: 10.1109/TIFS.2025.3597211]

Liang Z Y, Liu W F, Wang R, Wu M J, Li B H, Zhang Y Y, e tal. 2025. Transfer learning of real image features with soft contrastive loss for fake image detection//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI: 26281-26289 [DOI: 10.1609/aaai.v39i25.34826]

Lim Y, Lee C, Kim A and Etzioni O. 2024. DistilDIRE: a small, fast, cheap and lightweight diffusion synthesized deepfake detection//Proceedings of the 41st International Conference on Machine Learning Workshop on Foundation Models in the Wild. Vienna, Austria: JMLR.org

Lin L, Gupta N, Zhang Y, Ren H N, Liu C H, Ding F, et al. 2024.

Detecting multimedia generated by large AI models: a survey [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2402.00045v7.pdf

Lin M Q, Shang L and Gao X Y. 2023. Enhancing interpretability in AI-generated image detection with genetic programming//Proceedings of 2023 IEEE International Conference on Data Mining Workshops. Shanghai, China: IEEE: 371-378 [DOI: 10.1109/ICDMW60847.2023.00053]

Liu B, Yang F, Bi X L, Xiao B, Li W S and Gao X B. 2022. Detecting generated images by real images//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 95-110 [DOI: 10.1007/978-3-031-19781-9_6]

Liu H, Tan Z C, Tan C C, Wei Y C, Wang J D and Zhao Y. 2024a. Forgery-aware adaptive transformer for generalizable synthetic image detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10770-10780 [DOI: 10.1109/CVPR52733.2024.01024]

Liu H T, Li C Y, Wu Q Y and Lee Y J. 2023. Visual instruction tuning//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates, Inc.: 34892-34916 [DOI: 10.5555/3666122.3667638]

Liu R, Zhang S C, Xu Y, Xu W D and He X L. 2025. High-resolution network-based multi-feature fusion for generalized forgery detection. Multimedia Systems, 31: #35 [DOI: 10.1007/s00530-024-01626-z]

Liu Z H, Wang H Y, Kang Y Y and Wang S L. 2024b. Mixture of low-rank experts for transferable AI-generated image detection [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2404.04883.pdf

Lorenz P, Durall R L and Keuper J. 2023. Detecting images generated by deep diffusion models using their local intrinsic dimensionality//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision Workshops. Paris, France: IEEE: 448-459 [DOI: 10.1109/ICCVW60793.2023.00051]

Lu Z Y, Huang D, Bai L, Qu J J, Wu C Y, Liu X H, et al. 2023. Seeing is not always believing: benchmarking human and model perception of AI-generated images//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates, Inc.: 25435-25447 [DOI: 10.5555/3666122.3667227]

Luo Y P, Du J L, Yan K and Ding S H. 2024. LaRE2: latent reconstruction error based method for diffusion-generated image detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 17006-17015 [DOI: 10.1109/CVPR52733.2024.01609]

Lyu Z H, Xiao J, Zhang C and Lam K M. 2024. AI-generated image detection with Wasserstein distance compression and dynamic aggregation//Proceedings of 2024 IEEE International Conference on Image Processing. Abu Dhabi, United Arab Emirates: IEEE: 3827-3833 [DOI: 10.1109/ICIP51287.2024.10648186]

Ma R P, Duan J H, Kong F, Shi X S and Xu K D. 2023. Exposing the fake: effective diffusion-generated images detection [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2307.06272.pdf

Mandelli S, Bonettini N, Bestagini P and Tubaro S. 2020. Training CNNs in presence of JPEG compression: multimedia forensics vs computer vision//Proceedings of 2020 IEEE International Workshop on Information Forensics and Security. New York, USA: IEEE: 1-6 [DOI: 10.1109/WIFS49906.2020.9360903]

Marra F, Gragnaniello D, Cozzolino D and Verdoliva L. 2018. Detection of GAN-generated fake images over social networks//Proceedings of 2018 IEEE Conference on Multimedia Information Processing and Retrieval. Miami, USA: IEEE: 384-389 [DOI: 10.1109/MIPR.2018.00084]

Marra F, Saltori C, Boato G and Verdoliva L. 2019. Incremental learning for the detection and classification of GAN-generated images//Proceedings of 2019 IEEE International Workshop on Information Forensics and Security. Delft, Netherlands: IEEE: 1-6 [DOI: 10.1109/WIFS47025.2019.9035099]

Mehta A, McArthur B, Kolloju N and Tu Z Z. 2025. HFMF: hierarchical fusion meets multi-stream models for deepfake detection [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2501.05631.pdf

Meng Z L, Peng B, Dong J, Tan T N and Cheng H N. 2024. Artifact feature purification for cross-domain detection of AI-generated images. Computer Vision and Image Understanding, 247: #104078 [DOI: 10.1016/j.cviu.2024.104078]

Mi Z J, Jiang X H, Sun T F and Xu K. 2020. GAN-generated image detection with self-attention mechanism against GAN generator defect. IEEE Journal of Selected Topics in Signal Processing, 14(5): 969-981 [DOI: 10.1109/JSTSP.2020.2994523]

Mink J, Luo L C, Barbosa N M, Figueira O, Wang Y and Wang G. 2022. DeepPhish: understanding user trust towards artificially generated profiles in online social networks//Proceedings of the 31st USENIX Security Symposium. Boston, USA: USENIX Association: 1669-1686

Mirza M and Osindero S. 2014. Conditional generative adversarial nets [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/1411.1784.pdf

Moeßner P and Adel H. 2024. Human vs. AI: a novel benchmark and a comparative study on the detection of generated images and the impact of prompts [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2412.09715.pdf

Moskowitz A, Gaona T and Peterson J. 2024. Detecting AI-generated images via CLIP [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2404.08788.pdf

Nataraj L, Mohammed T M, Manjunath B S, Chandrasekaran S, Flenner A, Bappy J H, et al. 2019. Detecting GAN generated fake images using co-occurrence matrices. Electronic Imaging, 31(5): 532-1-532-7 [DOI: 10.2352/ISSN.2470-1173.2019.5.MWSF-532]

Nguyen T D, Azizpour A and Stamm M C. 2025. Forensic self-descriptions are all you need for zero-shot detection, open-set source attribution, and clustering of AI-generated images//Proceed-

ings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 3040-3050 [DOI: 10.1109/CVPR52734.2025.00289]

Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, et al. 2022. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR: 16784-16804

Ojha U, Li Y H and Lee Y J. 2023. Towards universal fake image detectors that generalize across generative models//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 24480-24489 [DOI: 10.1109/CVPR52729.2023.02345]

Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. 2024. DINOv2: learning robust visual features without supervision [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2304.07193v2.pdf

Park J and Owens A. 2025. Community forensics: using thousands of generators to train fake image detectors//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 8245-8257 [DOI: 10.1109/CVPR52734.2025.00772]

Park T, Liu M Y, Wang T C and Zhu J Y. 2019. Semantic image synthesis with spatially-adaptive normalization//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2332-2341 [DOI: 10.1109/CVPR.2019.00244]

Podell D, English Z, Lacey K, Blattmann A, Dockhorn T, Müller J, et al. 2024. SDXL: improving latent diffusion models for high-resolution image synthesis//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria

Pontorno O, Guarnera L and Battiato S. 2024. On the exploitation of DCT-traces in the generative-AI domain//Proceedings of 2024 IEEE International Conference on Image Processing. Abu Dhabi, United Arab Emirates: IEEE: 3806-3812 [DOI: 10.1109/ICIP51287.2024.10648013]

Qiao T, Chen Y X, Zhou X F, Shi R, Shao H, Shen K Y, et al. 2024a. CSC-Net: cross-color spatial co-occurrence matrix network for detecting synthesized fake images. IEEE Transactions on Cognitive and Developmental Systems, 16(1): 369-379 [DOI: 10.1109/TCDS.2023.3274450]

Qiao T, Shao H, Xie S C and Shi R. 2024b. Unsupervised generative fake image detector. IEEE Transactions on Circuits and Systems for Video Technology, 34(9): 8442-8455 [DOI: 10.1109/TCSVT.2024.3383833]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. Virtual: PMLR: 8748-8763

Radford A, Narasimhan K, Salimans T and Sutskever I. 2018. Improving language understanding by generative pre-training. OpenAI [EB/OL]. [2025-02-01]. https://cdn. openai. com/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford A, Wu J, Child R, Luan D, Amodei D and Sutskever I. 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8): #9

Rahman M A, Paul B, Sarker N H, Hakim Z I A and Fattah S A. 2023. Artifact: a large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection//Proceedings of 2023 IEEE International Conference on Image Processing. Kuala Lumpur, Malaysia: IEEE: 2200-2204 [DOI: 10.1109/ICIP49359.2023.10222083]

Rajan A S, Ojha U, Schloesser J and Lee Y J. 2025. Aligned Datasets Improve Detection of Latent Diffusion-Generated Images//Proceedings of the 13th International Conference on Learning Representations. Sinapore

Ramesh A, Dhariwal P, Nichol A, Chu C and Chen M. 2022. Hierarchical text-conditional image generation with CLIP latents [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2204.06125.pdf

Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, et al. 2021. Zero-shot text-to-image generation//Proceedings of the 38th International Conference on Machine Learning. Virtual: PMLR: 8821-8831

Raza S A, Habib U, Usman M, Cheema A A and Khan M S. 2024. MMGANGuard: a robust approach for detecting fake images generated by GANs using multi-model techniques. IEEE Access, 12: 104153-104164 [DOI: 10.1109/ACCESS.2024.3393842]

Razzhigaev A, Shakhmatov A, Maltseva A, Arkhipkin V, Pavlov I, Ryabov I, et al. 2023. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion//Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Singapore: ACL: 286-295 [DOI: 10.18653/v1/2023.emnlp-demo.25]

Ricker J, Damm S, Holz T and Fischer A. 2024a. Towards the detection of diffusion model deepfakes//Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. Roma, Italy: SciTePress: 446-457 [DOI: 10.5220/0012422000003660]

Ricker J, Lukovnikov D and Fischer A. 2024b. AEROBLADE: training-free detection of latent diffusion images using autoencoder reconstruction error//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9130-9140 [DOI: 10.1109/CVPR52733.2024.00872]

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10674-10685

［DOI：10.1109/CVPR52688.2022.01042］

Ruiz N，Li Y Z，Jampani V，Pritch Y，Rubinstein M and Aberman K. 2023. DreamBooth：fine tuning text-to-image diffusion models for subject-driven generation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：22500-22510［DOI：10.1109/CVPR52729.2023.02155］

Saharia C，Chan W，Saxena S，Lit L，Whang J，Denton E，et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans，USA：Curran Associates Inc.：36479-36494［DOI：10.5555/3600270.3602913］

Santosh，Lin L，Amerini I，Wang X and Hu S. 2024. Robust CLIP-based detector for exposing diffusion model-generated images［EB/OL］.［2025-02-01］. https://arxiv.org/pdf/2404.12908.pdf

Sarkar A，Mai H，Mahapatra A，Lazebnik S，Forsyth D A and Bhattad A. 2024. Shadows don't lie and lines can't bend! Generative models don't know projective geometry... for now//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：28140-28149［DOI：10.1109/CVPR52733.2024.02658］

Sauer A，Karras T，Laine S，Geiger A and Aila T. 2023. StyleGAN-T：unlocking the power of GANs for fast large-scale text-to-image synthesis//Proceedings of the 40th International Conference on Machine Learning. Honolulu，USA：PMLR：30105-30118［DOI：10.5555/3618408.3619658］

Schinas M and Papadopoulos S. 2024. SIDBench：a Python framework for reliably assessing synthetic image detection methods//Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation. Phuket，Thailand：ACM：55-64［DOI：10.1145/3643491.3660277］

Schuhmann C，Beaumont R，Vencu R，Gordon C，Wightman R，Cherti M，et al. 2022. LAION-5B：an open large-scale dataset for training next generation image-text models//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans，USA：Curran Associates Inc.：25278-25294［DOI：10.5555/3600270.3602103］

Selvaraju R R，Cogswell M，Das A，Vedantam R，Parikh D and Batra D. 2017. Grad-CAM：visual explanations from deep networks via gradient-based localization//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice，Italy：IEEE：618-626［DOI：10.1109/ICCV.2017.74］

Sha Z Y，Li Z，Yu N and Zhang Y. 2023. De-fake：detection and attribution of fake images generated by text-to-image generation models//Proceedings of 2023 ACM SIGSAC Conference on Computer and Communications Security. Copenhagen，Denmark：ACM：3418-3432［DOI：10.1145/3576915.3616588］

Sha Z Y，Tan Y C，Li M J，Backes M and Zhang Y. 2024. ZeroFake：zero-shot detection of fake images generated and edited by text-to-image generation models//Proceedings of 2024 ACM SIGSAC Conference on Computer and Communications Security. Salt Lake City，USA：ACM：4852-4866［DOI：10.1145/3658644.3690297］

Shen C J，Liu Z J，Chen K X，Lei J，Song M L and Feng Z L. 2025. Spatial-temporal reconstruction error for AIGC-based forgery image detection//Proceedings of 2025 IEEE International Conference on Acoustics，Speech and Signal Processing. Hyderabad，India：IEEE：1-5［DOI：10.1109/ICASSP49660.2025.10890455］

Sinitsa S and Fried O. 2024. Deep image fingerprint：towards low budget synthetic image detection and model lineage analysis//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa，USA：IEEE：4055-4064［DOI：10.1109/WACV57701.2024.00402］

Sohl-Dickstein J，Weiss E A，Maheswaranathan N and Ganguli S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics//Proceedings of the 32nd International Conference on Machine Learning. Lille，France：PMLR：2256-2265［DOI：10.5555/3045118.3045358］

Sun K，Xiao B，Liu D and Wang J D. 2019. Deep high-resolution representation learning for human pose estimation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach，USA：IEEE：5686-5696［DOI：10.1109/CVPR.2019.00584］

Tan C C，Liu H，Zhao Y，Wei S K，Gu G H，Liu P and Wei Y C. 2024a. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：28130-28139［DOI：10.1109/CVPR52733.2024.02657］

Tan C C，Liu P，Tao R S，Liu H，Zhao Y，Wu B Y，et al. 2024b. Data-independent operator：a training-free artifact representation extractor for generalizable deepfake detection［EB/OL］.［2025-02-01］. https://arxiv.org/pdf/2403.06803.pdf

Tan C C，Tao R S，Liu H，Gu G H，Wu B Y，Zhao Y，et al. 2025. C2P-CLIP：injecting category common prompt in CLIP to enhance generalization in deepfake detection//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia，USA：AAAI：7184-7192［DOI：10.1609/aaai.v39i7.32772］

Tan C C，Zhao Y，Wei S K，Gu G H，Liu P，et al. 2024c. Frequency-aware deepfake detection：improving generalizability through Frequency space domain learning//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver，Canada：AAAI：5052-5060［DOI：10.1609/aaai.v38i5.28310］

Tan C C，Zhao Y，Wei S K，Gu G H and Wei Y C. 2023. Learning on gradients：generalized artifacts representation for GAN-generated images detection//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：12105-12114［DOI：10.1109/CVPR52729.2023.01165］

Tao M, Bao B K, Tang H and Xu C S. 2023. GALIP: generative adversarial CLIPs for text-to-image synthesis//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 14214-14223 [DOI: 10.1109/CVPR52729.2023.01366]

Tao R S, Tan C C, Liu H, Wang J K, Qin H T, Chang Y K, et al. 2025. SAGNet: decoupling semantic-agnostic artifacts from limited training data for robust generalization in deepfake detection. IEEE Transactions on Information Forensics and Security, 20: 6429-6442 [DOI: 10.1109/TIFS.2025.3581726]

Tariang D, Corvi R, Cozzolino D, Poggi G, Nagano K and Verdoliva L. 2024. Synthetic image verification in the era of generative artificial intelligence: what works and what isn't there yet. IEEE Security and Privacy, 22(3): 37-49 [DOI: 10.1109/MSEC.2024.3376637]

Tsai C T, Ko C Y, Chung I, Wang Y C F and Chen P Y. 2024. Understanding and improving training-free AI-generated image detections with vision foundation models [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2411.19117.pdf

Uhlenbrock L, Cozzolino D, Moussa D, Verdoliva L and Riess C. 2024. Did you note my palette? Unveiling synthetic images through color statistics//Proceedings of 2024 ACM Workshop on Information Hiding and Multimedia Security. Baiona, Spain: ACM: 47-52 [DOI: 10.1145/3658664.3659652]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA. Curran Associates Inc.: 6000-6010 [DOI: 10.5555/3295222.3295349]

Verdoliva L, Cozzolino D and Nagano K. 2022. IEEE Image and Video Processing Cup Synthetic Image Detection. IEEE

Vladimir A, Vasilev V, Filatov A, Pavlov I, Agafonova J, Gerasimenko N, et al. 2024. Kandinsky 3: text-to-image synthesis for multifunctional generative framework//Proceedings of 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Miami, USA: ACL: 475-485 [DOI: 10.18653/v1/2024.emnlp-demo.48]

Wang S Y, Wang O, Zhang R, Owens A and Efros A A. 2020. CNN-generated images are surprisingly easy to spot... for now//Proceedings of 2020 IEEE/CVF conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8692-8701 [DOI: 10.1109/cvpr42600.2020.00872]

Wang T, Zhang Y S, Qi S R, Zhao R Y, Xia Z H and Weng J. 2024a. Security and privacy on generative data in AIGC: a survey. ACM Computing Surveys, 57(4): #82 [DOI: 10.1145/3703626]

Wang T Y, Liao X, Chow K P, Lin X D and Wang Y L. 2024b. Deepfake detection: a comprehensive survey from the reliability perspective. ACM Computing Surveys, 57(3): #58 [DOI: 10.1145/3699710]

Wang W T, Huang X Y, Wang T Y and Roy S K. 2023a. DeepArt: a benchmark to advance fidelity research in AI-generated content [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2312.10407.pdf

Wang Y B, Huang Z W and Hong X P. 2023b. Benchmarking deepart detection [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2302.14475.pdf

Wang Y B, Huang Z W and Hong X P. 2025. OpenSDI: spotting diffusion-generated images in the open world//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 4291-4301 [DOI: 10.1109/CVPR52734.2025.00405]

Wang Z D, Bao J M, Zhou W G, Wang W L, Hu H Z, Chen H, et al. 2023c. DIRE for diffusion-generated image detection//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22388-22398 [DOI: 10.1109/ICCV51070.2023.02051]

Wang Z J, Montoya E, Munechika D, Yang H Y, Hoover B and Chau D H. 2023d. DiffusionDB: a large-scale prompt gallery dataset for text-to-image generative models//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: ACL: 893-911 [DOI: 10.18653/v1/2023.acl-long.51]

Wißmann A, Zeiler S, Nickel R M and Kolossa D. 2024. Whodunit: detection and attribution of synthetic images by leveraging model-specific fingerprints//Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation. Phuket, Thailand: ACM: 65-72 [DOI: 10.1145/3643491.3660280]

Wu H W, Zhou J T and Zhang S L. 2025a. Generalizable synthetic image detection via language-guided contrastive learning [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2305.13800v2.pdf

Wu S Y, Liu J, Li J and Wang Y Q. 2025b. Few-shot learner generalizes across AI-generated image detection [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2501.08763v2.pdf

Xi Z Y, Huang W M, Wei K K, Luo W Q and Zheng P J. 2023. AI-generated image detection using a cross-attention enhanced dual-stream network//Proceedings of 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference. Taipei, China: IEEE: 1463-1470 [DOI: 10.1109/APSIPAASC58517.2023.10317126]

Xiao Y, Yang B B, Chen W Y, Chen J H, Cao Z J, Dong Z Y, et al. 2025. Are high-quality AI-generated images more difficult for models to detect?//Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: JMLR.org

Xie C Y, Ye D P, Zhang Y M, Tang L, Lv Y N, Deng J C, et al. 2024. Take fake as real: realistic-like robust black-box adversarial attack to evade AIGC detection [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2412.06727v2.pdf

Xie T Q, Wu Y Y, Jing C and Sun W H. 2024. Survey of image detection methods generated by GAN models. Computer Engineering and Applications, 60(22): 74-86 (谢天圻, 吴媛媛, 敬超, 孙伟恒.

2024. GAN模型生成图像检测方法综述. 计算机工程与应用，60(22)：74-86)［DOI：10.3778/j.issn.1002-8331.2405-0346］

Xu J C，Yang Y，Fang H，Liu H G and Zhang W M. 2025a. FAMSeC：a few-shot-sample-based general AI-generated image detection method. IEEE Signal Processing Letters，32：226-230［DOI：10.1109/LSP.2024.3511421］

Xu K，Zhang L Z and Shi J B. 2025b. Detecting origin attribution for text-to-image diffusion models//Proceedings of 2025 IEEE/CVF Winter Conference on Applications of Computer Vision. Tucson，USA：IEEE：8775-8785［DOI：10.1109/WACV61041.2025.00850］

Xu Q，Jia S，Jiang X H，Sun T F，Wang Z and Yan H. 2024a. MDTL-NET：computer-generated image detection based on multi-scale deep texture learning. Expert Systems with Applications，248：#123368［DOI：10.1016/j.eswa.2024.123368］

Xu Q，Jiang X H，Sun T F，Wang H，Meng L J and Yan H. 2024b. Detecting artificial intelligence-generated images via deep trace representations and interactive feature fusion. Information Fusion，112：#102578［DOI：10.1016/j.inffus.2024.102578］

Xu Q，Wang H，Meng L J，Mi Z J，Yuan J Y and Yan H. 2023. Exposing fake images generated by text-to-image diffusion models. Pattern Recognition Letters，176：76-82［DOI：10.1016/j. patrec. 2023.10.021］

Xue L，Shu M L，Awadalla A，Wang J，Yan A，Purushwalkam S，et al. 2025. BLIP-3：a family of open large multimodal models［EB/OL］.［2025-02-01］. https://arxiv.org/pdf/2408.08872v4.pdf

Yan S L，Li O X，Cai J Y，Hao Y B，Jiang X L，Hu Y，et al. 2025. A sanity check for AI-generated image detection//Proceedings of the 13th International Conference on Learning Representations. Sinapore，Sinapore

Yang D，Huang Y H，Guo Q，Juefei-Xu F，Jia X F，Wang R，et al. 2024. Text modality oriented image feature extraction for detecting diffusion-based deepfake［EB/OL］.［2025-02-01］. https://arxiv.org/pdf/2405.18071.pdf

Yang Y Q，Qian Z H，Zhu Y，Russakovsky O and Wu Y. 2025. D3：scaling up deepfake detection by learning from discrepancy//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville，USA：IEEE：23850-23859［DOI：10.1109/CVPR52734.2025.02221］-03-23. v2

Yu J H，Xu Y Z，Koh J Y，Luong T，Baid G，Wang Z R，et al. 2022. Scaling autoregressive models for content-rich text-to-image generation［EB/OL］.［2025-02-01］. https://arxiv.org/pdf/2206.10789.pdf

Yu X，Chen K J，Zeng K，Fang H，Yang Z J，Shang X W，et al. 2024. SemGIR：semantic-guided image regeneration based method for AI-generated image detection and attribution//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne，Australia：ACM：8480-8488［DOI：10.1145/3664647.3680776］

Zamir S W，Arora A，Khan S，Hayat M，Khan F S，Yang M H，et al. 2020. CycleISP：real image restoration via improved data synthesis//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：2696-2705［DOI：10.1109/CVPR42600.2020.00277］

Zhang K，Zuo W M，Chen Y J，Meng D Y and Zhang L. 2017. Beyond a Gaussian denoiser：residual learning of deep CNN for image denoising. IEEE Transactions on Image Processing，26(7)：3142-3155［DOI：10.1109/TIP.2017.2662206］

Zhang L，Chen H，Hu S，Zhu B，Lin C S，Wu X，et al. 2024. X-Transfer：a transfer learning-based framework for GAN-generated fake image detection//Proceedings of 2024 International Joint Conference on Neural Networks. Yokohama，Japan：IEEE：1-8［DOI：10.1109/IJCNN60899.2024.10650566］

Zhang T. 2022. Deepfake generation and detection，a survey. Multimedia Tools and Applications，81(5)：6259-6276［DOI：10.1007/s11042-021-11733-y］

Zhang X，Karaman S and Chang S F. 2019. Detecting and simulating artifacts in GAN fake images//Proceedings of 2019 IEEE International Workshop on Information Forensics and Security. Delft，Netherlands：IEEE：1-6［DOI：10.1109/WIFS47025.2019.9035107］

Zhang H F，He Q H，Bi X L，Li W S，Liu B and Xiao B. 2025. Towards universal AI-generated Image detection by variational information bottleneck network//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville，USA：IEEE：23828-23837［DOI：10.1109/CVPR52734.2025.02219］

Zhang Y C and Xu X G. 2025. Diffusion noise feature：accurate and fast generated image detection［EB/OL］.［2025-02-01］. https://arxiv.org/pdf/2312.02625v3.pdf

Zheng C D，Lin C H，Zhao Z Y，Wang H，Guo X，Liu S，et al. 2024. Breaking semantic artifacts for generalized AI-generated image detection//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver，Canada：Curran Associates Inc.：59570-59596［DOI：10.5555/3737916.3739819］

Zhong N，Chen H Y，Xu Y R，Qian Z X and Zhang X P. 2025. Beyond generation：a diffusion-based low-level feature extractor for detecting AI-generated images//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville，USA：IEEE：8258-8268［DOI：10.1109/CVPR52734.2025.00773］

Zhong N，Xu Y R，Li S，Qian Z X and Zhang X P. 2024. PatchCraft：exploring texture patch for efficient AI-generated image detection［EB/OL］.［2025-02-01］. https://arxiv.org/pdf/2311.12397v3.pdf

Zhou Z Y，Luo Y P，Wu Y C，Sun K，Ji J Y，Yan K，et al. 2025. AIGI-holmes：towards explainable and generalizable AI-generated image detection via multimodal large language models［EB/OL］.［2025-02-01］. https://arxiv.org/pdf/2507.02664v2.pdf

Zhou Z Y，Sun K，Chen Z X，Kuang H F，Sun X S and Ji R R. 2024. StealthDiffusion：towards evading diffusion forensic detection through diffusion model//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne，Australia：ACM：3627-3636［DOI：10.1145/3664647.3681535］

Zhu J Y，Park T，Isola P and Efros A A. 2017. Unpaired image-to-image

translation using cycle-consistent adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2242-2251 [DOI: 10.1109/ICCV.2017.244]

Zhu M J, Chen H T, Huang M X, Li W, Hu H L, Hu J, et al. 2023a. GenDet: towards good generalizations for AI-generated image detection [EB/OL]. [2025-02-01]. https://arxiv.org/pdf/2312.08880.pdf

Zhu M J, Chen H T, Yan Q Y, Huang X D, Lin G Y, Li W, et al. 2023b. GenImage: a million-scale benchmark for detecting AI-generated image//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 77771-77782 [DOI: 10.5555/3666122.3669520]

作者简介

李美玲，女，博士研究生，主要研究方向为人工智能生成内容的检测与溯源。E-mail：mlli20@fudan.edu.cn

钱振兴，通信作者，男，教授，主要研究方向为信息隐藏、多媒体与人工智能安全。E-mail：zxqian@fudan.edu.cn

张新鹏，男，教授，主要研究方向为信息隐藏、多媒体与人工智能安全。E-mail：zhangxinpeng@fudan.edu.cn