中图法分类号:TP391.41;TP18　　**文献标识码**:A　　**文章编号**:1006-8961(2026)10-3523-22

论文引用格式:Li J Y, Zhang M Q, Huang S Y, Feng P and Liu L. 2026. Dual learning paradigms for AI-generated image detection and forensics: a survey. Journal of Image and Graphics, 31(10):3523-3544(李佳业, 张敏情, 黄思远, 冯佩, 刘浪. 2026. 双学习范式AI生成图像检测取证技术综述. 中国图象图形学报, 31(10):3523-3544)[DOI:10.11834/jig.250624]

双学习范式AI生成图像检测取证技术综述

李佳业[1,2],张敏情[1,2*],黄思远[1,2],冯佩[3],刘浪[1,2]

1. 武警工程大学密码工程学院,西安　710086; 2. 网络与信息安全武警部队重点实验室,西安　710086;
3. 武警工程大学装备管理与保障学院,西安　710086

摘　要:人工智能(artificial intelligence,AI)生成内容技术的爆发增长引发了严重的信任危机,AI生成图像检测已成为多媒体取证领域的核心议题。为破解当前检测技术研究视角碎片化、理论体系割裂的困境,本文提出一种从"学习范式"切入的统一分析框架,对AI生成图像检测技术进行系统性综述。首先,本文构建了包含"证伪"与"存真"两大根本性类别的双学习范式分析体系,并对各类方法的基本思想进行全面阐述。其中,"证伪"范式侧重于挖掘生成过程中残留的频域异常、空间不一致等伪造痕迹;而"存真"范式旨在通过单类分类或物理一致性约束,对真实图像的固有分布进行建模以识别异常。其次,通过梳理理论分析与技术演进路径,本文揭示了检测技术由被动识别特定痕迹的"证伪"向主动建模真实分布的"存真"演进的趋势,并深入剖析了"证伪"范式在跨域泛化能力上的局限性,以及"存真"范式在建模高维复杂分布时的核心挑战。此外,本文构建了包含库内性能、跨模态泛化性、鲁棒性及可解释性等多维度的评估体系,对比了代表性算法的优劣。最后,本文探讨了范式融合、基于基础模型的检测及反取证防御等未来研究方向,为构建高鲁棒、可解释的AI生成图像检测体系提供理论参考。本文将相关开源论文及代码链接汇总至GitHub仓库(https://github.com/qqqqwddw/wesome-image-Detection),并将整理后的数据集与分析数据上传至ScienceDB社区(https://www.scidb.cn/detail?dataSetId=d6427185d9ff4c43bb6b190a117d723b),以供后续研究参考。

关键词:多媒体取证;AI生成图像检测;深度伪造;表征学习;异常检测;泛化性

Dual learning paradigms for AI-generated image detection and forensics: a survey

Li Jiaye[1,2], Zhang Minqing[1,2*], Huang Siyuan[1,2], Feng Pei[3], Liu Lang[1,2]

1. College of Cryptography Engineering, Engineering University of PAP, Xi'an　710086, China; 2. Key Laboratory of Network and Information Security Under the Chinese People's Armed Police Force (Engineering University of PAP), Xi'an　710086, China; 3. College of Equipment Management and Support, Engineering University of PAP, Xi'an　710086, China

Abstract: AI-generated image detection technology aims to distinguish synthetic content from authentic imagery to safeguard media integrity and facilitate the subsequent forensic tasks further. In recent years, it has been a concern in the field of information security, especially in artificial intelligence-related industries such as judicial evidence verification, social media moderation, identity authentication, news broadcasting, and political security surveillance. Moreover, the growth of

收稿日期:2025-12-18;**修回日期**:2026-01-21;**预印本日期**:2026-01-28

*通信作者:张敏情　api_zmq@126.com

基金项目:国家自然科学基金项目(62272478)

Supported by: National Natural Science Foundation of China (62272478)

generative models has been promoting deep learning-based detection algorithms. In particular, the emergence of advanced techniques, such as the variational autoencoder, generative adversarial network, and denoising diffusion probabilistic model, has led to a qualitative leap in synthesis quality. However, a comprehensive review and analysis of state-of-the-art deep learning-based detection algorithms for different generative attacks are required to be realized to address the current theoretical fragmentation. Thus, we develop a systematic and critical review to explore the developments of AI-generated image detection in recent years. First, a comprehensive and systematic introduction of the detection field is presented from the following three aspects: 1) the evolution of forensic technology from handcrafted features to deep representations, 2) the prevailing deepfake datasets, and 3) the common evaluation metrics. Then, more extensive qualitative experiments, quantitative experiments, and generalization evaluations of representative detection methods are conducted on the public datasets to compare their performance. Finally, the summary and challenges in the forensic community are highlighted. In particular, some prospects are recommended further in the field of digital forensics. First, from the perspective of learning paradigms, the existing image detection methods can be divided into two fundamental categories, i. e. , the disproof paradigm aimed at identifying falsity, and the truth-preserving paradigm aimed at modeling reality. Specifically, the disproof paradigm focuses on extracting specific forgery traces including frequency domain artifacts and spatial inconsistencies, while the truth-preserving paradigm targets the modeling of intrinsic real image distributions to detect anomalies. In addition, the domain of deep learning-based detection algorithms can be classified into the convolutional neural network-based fusion framework and vision transformer-based framework from the aspect of network architectures. The disproof-based framework achieves the feature extraction of manipulation signatures by supervised binary classification and accomplishes deep feature discrimination via learning distinguishable boundaries between real and fake domains. To clarify the transition of methodologies, the truth-preserving framework is originated from one-class classification and reconstruction principles. From the perspective of the supervision paradigm, the deep learning detection methods can also be categorized into three classes, i. e. , supervised learning, unsupervised learning, and self-supervised learning. The supervised methods leverage labeled synthetic data to guide the training processes, and the unsupervised approaches construct loss functions via constraining the similarity between the input and the learned manifold of natural images. The self-supervised algorithms are associated with the masked image modeling framework in common. Our critical review is focused on the main concepts and discussions of the characteristics of each method for different forensic scenarios from the perspectives of the network architecture and learning paradigm. First, we summarize the limitations of different algorithms, such as the generalization failure of disproof methods and the modeling complexity of truth-preserving methods, and provide some recommendations for further research. Second, we briefly introduce the popular public datasets such as FaceForensics++ and Wild-deepfake and provide the interfaces related to download them for each specific detection scenario. Then, we present the common evaluation metrics in the forensic field from two aspects: classification-based evaluation metrics and robustness-based metrics designed for cross-generator settings. The generic metrics can be utilized to evaluate the precision and sensitivity of algorithms of those are accuracy-based, error rate-based, receiver operating characteristic-based, and area under the curve-based metrics in total. Some of the generic metrics, such as accuracy (ACC), area under the curve (AUC), equal error rate (EER), and true positive rate (TPR), are also used for the quantitative assessment of generalization. The specific metrics designed for localization consist of pixel-level intersection over union and bounding box precision that employ the pixel-wise masks as the reference ground truth. Third, we present the qualitative and quantitative results, and average inference times of representative alternatives for various forensic missions. Finally, this review critically analyzes the conclusion, highlights the challenges in the detection community, and conducts forecasting analysis, such as cross-domain image detection, high-level semantic consistency-driven detection, antiforensic robust detection, real-time mobile detection, explainable forensics based on physical imaging principles, detection under extreme compression, and comprehensive evaluation benchmarks. The methods, datasets, and evaluation metrics mentioned are linked at https://github.com/qqqqwddw/wesome-image-Detection and https://www.scidb.cn/detail?dataSetId=d6427185d9ff4c43bb6b190a117d723b.

Key words: multimedia forensics; AI-generated image detection; deepfake; representation learning; anomaly detection; generalization performance

0 引 言

近年来，人工智能（artificial intelligence，AI）在图像生成领域引发了深刻的范式变革，推动了生成模型从早期的生成对抗网络（generative adversarial network，GAN）（Goodfellow 等，2020）向更复杂、更逼真的扩散模型（diffusion model）（Ho 等，2020）演进，并进一步与基于 Transformer 架构的文生图（text-to-image）模型（Vaswani 等，2017）深度融合。以 Midjourney 和 Stable Diffusion（Esser 等，2024）为代表的先进生成模型，已能生成在视觉细节、语义连贯性和真实感方面与真实照片高度相似的图像，显著模糊了真实与合成内容之间的界限。上述技术突破在推动创意产业与数字内容生产效率提升的同时，也引发了关于新闻真实性、社会信任机制、身份伪造以及知识产权归属等多重社会风险担忧（Roose，2022）。

随着生成技术的不断进步，伪造内容的传播速度和隐蔽性显著增强，传统的检测手段正面临着严峻挑战。鉴于生成技术的进步，发展高效、鲁棒且可泛化的 AI 生成图像检测技术已成为保障数字内容可信度、维护信息生态安全的迫切需求。然而，在面对新型生成模型时，尤其是扩散模型与大型文生图模型，普遍存在泛化能力不足、特征提取不稳定等问题，亟需从理论层面构建系统性分析框架，以揭示检测技术演进的内在逻辑与核心瓶颈。基于此，本研究旨在从“学习范式”的视角切入，对 AI 生成图像检测技术进行系统梳理与理论重构，为该领域的进一步发展提供理论支撑与方向指引。

面对这一挑战，学术界已涌现出若干综述文献以梳理应对之策。这些工作普遍停留在对检测方法技术实现的梳理，未能揭示驱动技术演进的深层逻辑。具体而言，现有综述主要陷入了 3 种“表象式”分类的局限：1）“枚举式”分类，通过罗列基于频率、生物信号等具体伪影的检测技术，导致分类体系庞杂且缺乏前瞻性（王任颖 等，2022；谢天圻 等，2024；丁峰 等，2024；姚文达 等，2025；Mahara 和 Rishe，2025）。2）“架构驱动式”分类，聚焦于 MesoNet（Afchar 等，2018）等特定网络模型，过度关注实现“工具”而忽略了共通的检测原理，使其内容极易随着新架构的出现而过时（Khan 等，2024；Cogranne，2025）。3）“宽泛标签式”分类，采用“传统”与“先进”或按模态划分，这种宏观划分掩盖了方法论的内在共性（Sharma 等，2024；Alrashoud，2025；Liu 等，2024b；Sunil 等，2025）。

上述分类方式的共同缺陷在于均停留在对检测方法“外在表现”的描述层面，而未能深入到驱动其发展的“内核思想”——即学习范式。例如，最新综述（李美玲 等，2026）虽然系统覆盖了像素域、频域以及预训练模型特征等多类检测依据，并对代表性方法的泛化性与鲁棒性进行了横向比较，但其分类主线仍以特征来源为中心，尚未上升到方法论层面的范式抽象。学习范式作为方法论层面的抽象概括，能够揭示检测策略的根本目标与认知路径，即“证伪”与“存真”的对立统一，从而为技术演进提供统一的分析框架。为系统性地凸显本文研究的独特性与理论深度，如表 1 所示，本文从分类维度、理论深度、时效性与可扩展性等关键指标，对现有代表性综述进行了定量对比分析。结果表明，本文在理论框架构建与技术洞察力方面具有显著优势。

可以看出，现有综述文献难以回答一个关键的理论问题：在生成模型从生成对抗网络向扩散模型乃至未来更先进架构快速演进的背景下，AI 生成图像检测领域是否存在一条贯穿技术发展始终、可指导未来研究的内在技术主脉络？这一问题的解答，不仅关乎对现有技术的系统性理解，更关系到检测技术能否在面对不断涌现的新型生成手段时保持持续有效性。

表 1　与其他综述对比

Table 1　Compared with other existing literature reviews

参考文献	学习范式分析	算法比较	局限分析	未来方向
王任颖等人（2022）	×	√	×	√
Mahara 和 Rishe（2025）	×	√	√	√
谢天圻等人（2024）	×	√	×	×
Khan 等人（2024）	×	√	√	√
Cogranne（2025）	×	√	√	√
Sharm 等人（2024）	×	√	√	√
Alrashoud（2025）	×	×	√	√
Liu 等人（2024b）	×	×	√	√
Sunil 等人（2025）	×	√	×	×
李美玲等人（2026）	×	√	√	√
本文	√	√	√	√

注：“√”表示涉及，“×”表示未涉及。

为应对上述挑战并突破现有综述在分类维度上的表层局限，本文提出一个全新的、基于“学习范式”的统一分析框架。该框架超越了对具体伪影特征或网络架构的依赖，转而从方法论的本源出发，如图1所示，将现有的AI生成图像检测技术系统性地归纳为两大根本范式：1）发现足以否定真实性的异常证据，其学习对象可以来源于生成过程引入的人工伪影，也可以来源于真实对象在几何结构、生理规律或时序一致性等层面被破坏后所显现的不可成立性特征的“证伪”范式，在此意义下，即便部分方法（如Face X-ray、头姿或生理一致性检测）并未直接对生成模型进行建模，其检测逻辑若仍以“发现违背真实人脸应有约束的证据”为目标，仍应归属于“证伪”范式。2）以显式或隐式建模真实图像分布本身为目标，其核心任务并非定位具体伪造痕迹，而是判定样本是否偏离真实数据所构成的潜在流形。

基于该框架，本文主要贡献可概括为以下3个方面：1）构建了一个统一且具有深层洞察力的分类体系。该体系首次以“学习范式”为理论主线，对检测方法进行本质性归类，为AI生成图像检测领域提供了更具系统性、前瞻性和理论深度的知识图谱。2）揭示了检测技术演进的内在逻辑与发展趋势。该框架不仅实现了对深度伪造（deepfake）检测与通用合成图像检测两大子领域的统一分析，更清晰地勾勒出一条由被动识别特定伪造痕迹向主动建模真实图像分布演进的技术路径，体现了检测范式从“症状识别”到“本质建模”的必然趋势。3）提出了面向未来研究的系统性展望。通过深入剖析“证伪”范式在泛化能力与跨模型适应性上的局限，以及“存真”范式在建模复杂性与计算效率方面的挑战，本文指出了当前检测技术在鲁棒性、效率与可扩展性等方面的核心瓶颈，并为如范式融合、对抗性防御机制设计、自适应检测模型等未来研究方向提供了明确的理论指引与实践路径。本文所提出的双范式框架不仅为理解AI生成图像检测技术的演进逻辑提供了新的理论视角，也为该领域的可持续发展奠定了理论基础。

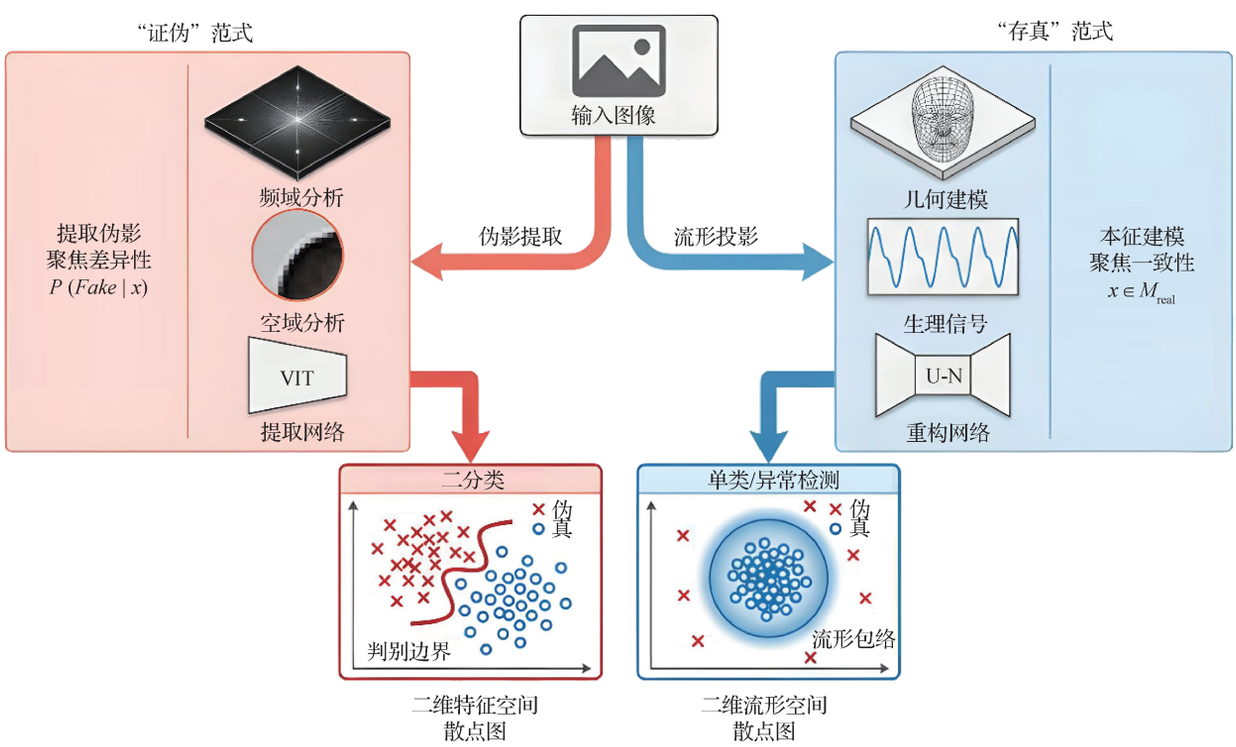


图1　双学习范式AI生成图像检测原理框架图

Fig. 1　Theoretical framework of AI-generated image detection based on dual learning paradigms

1　图像生成技术演进及检测挑战

AI生成图像指由生成式模型自主创造或对真实图像进行篡改所生成的视觉内容。基于上述“自主创造”与“篡改”的核心区别，本文将人工智能生成图像（AI-generated images，AIGI）主要划分为deepfake与通用合成图像两大类别。一方面，deepfake特指以身份伪造为核心目标的定向图像篡改技术，其典型应用包括人脸替换、表情操控和语音同步等，常

用于身份欺诈、虚假信息传播及社会操纵等高风险场景。其技术实现依赖于特定主体生物的面部结构、皮肤纹理和动态表情等特征的精确建模与迁移，因此在生成过程中往往保留特定的伪造痕迹，如面部边缘不自然、光影不一致或微表情失真等。另一方面，通用合成图像则指从文本、草图或其他模态输入出发，生成无特定身份指向的全新视觉内容，其核心目标在于创造性内容生成。该类图像广泛应用于艺术创作、虚拟场景构建与数字内容生产等领域，但其大规模应用也诱发了虚假信息泛滥、版权归属争议及内容真实性危机等社会风险。

尽管上述两类图像在生成目的与应用场景上存在显著差异，但其底层均依赖于以生成对抗网络（GAN）、扩散模型等为代表的深度生成模型。上述模型在图像生成过程中，由于其特定的架构设计、训练机制及优化目标，往往会在合成图像中引入可被识别的“生成痕迹”，如频域异常、噪声分布偏差、纹理不一致性或局部结构失真等，从而构成了当前 AI 生成图像检测技术的理论基础与关键突破口。

值得注意的是，此类生成痕迹并非静态不变，而是随着底层生成算法的持续演进而动态演化。生成技术与检测技术之间呈现出一种持续的“技术博弈”关系，即随着生成模型从早期简单对抗结构向更复杂的扩散模型、多模态融合及自回归生成机制演进，合成图像的视觉逼真度与语义连贯性显著提升，其残留的伪造特征也日益隐蔽、多样化且具有更强的对抗性，从而对检测技术提出了更高的要求，这进而推动了检测范式的持续升级。为系统性地呈现这一技术迭代背景及其对检测任务的影响，本文将 2014—2024 年通用图像生成技术的发展脉络梳理如图 2 所示，该演进过程可划分为 4 个关键阶段：第 1 阶段为以深度卷积生成对抗网络（deep convolutional generative adversarial network，DCGAN）（Radford 等，2016）、基于风格的生成对抗网络（style-based generative adversarial network，StyleGAN）（Karras 等，2019）为代表的 GAN 主导期，该阶段生成图像在局部纹理与结构上表现出明显伪影，为早期检测方法提供了相对明确的特征空间；第 2 阶段为基于 Transformer 的视觉生成模型（Chen 等，2020）萌芽期，生成质量提升的同时，伪影分布趋于平滑与全局化；第 3 阶段为以 Stable Diffusion 为代表的扩散模型爆发期，生成图像在高分辨率与语义一致性上取得突破，伪造痕迹显著稀疏化，这使得依赖局部伪影统计的传统检测器性能急剧下降；第 4 阶段为以 Sora 文生视频模型（以下简称 Sora）为代表的多模态与高保真视频生成成熟期，生成内容扩展至动态场景与长时序语义，检测任务面临从静态图像向动态内容的范式迁移挑战。

上述技术演进不仅重塑了生成内容的质量边界，也深刻影响了检测策略的设计逻辑与有效性。因此，理解生成技术的演进规律，对于构建具有前瞻性与泛化能力的检测框架至关重要。本文所提出的“学习范式”分析框架，正是在这一动态博弈背景下，试图揭示检测技术演进的内在逻辑，为未来研究提

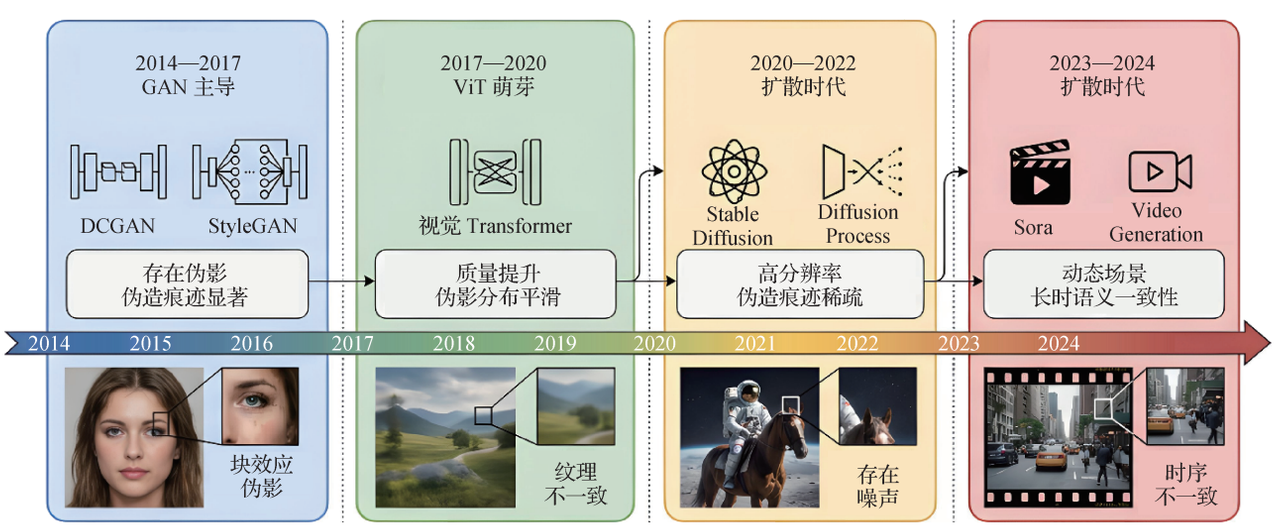


图 2　图像生成技术发展历程

Fig. 2　The evolutionary course of image generation techniques

供理论支撑与方向指引。

2 文献检索与筛选方法

为确保本研究在理论视野上的全面性与技术探讨上的前沿性，本文构建了一套系统且严谨的文献检索与筛选方法论体系。检索工作主要依托于Google Scholar、IEEE Xplore、ScienceDirect及ACM Digital Library等权威学术数据库展开，以全面覆盖计算机视觉与模式识别领域的核心知识库。在文献来源的选择上，本文采取了明确的质量导向策略，优先纳入发表于顶级国际会议（如CVPR、ICCV、ECCV）及权威期刊（如TPAMI、IJCV、TIFS）上的研究成果，以确保分析素材的高学术水准。

为精准定位研究主题，检索策略采用了组合式关键词方案，涵盖了从宏观的“AI-generated image detection”、“synthetic image detection”到更为聚焦的“deepfake detection”、“image forensics”及“GAN detection”等多维视角。考虑到生成与检测技术的快速迭代特性，将检索时间窗口限定为2020—2025年，确保所综述文献的前沿性，精准捕捉技术发展的最新态势。基于上述策略，本文最终遴选出超过200篇具有代表性的核心论文构筑文献库，如图3所示，从而为后续提出的双范式分析框架奠定了坚实的数据支撑与理论基石。



(a) 研究热点　　(b) 技术热点

图3　词云分析图

Fig. 3　Word cloud analysis chart ((a) research hotspots; (b) technology hotspots)

3 双学习范式检测技术体系分析

基于前文提出的“证伪”与“存真”双重范式框架，本节遵循由特定场景（deepfake）向普适化场景（通用合成图像）拓展的演进逻辑，揭示不同技术路线在动态攻防博弈中的演化脉络、核心优势及其固有的理论局限。

具体而言，本节首先聚焦于具有高度身份敏感性的deepfake检测领域，深入剖析“证伪”与“存真”两大范式如何利用人脸特有的生物特征与物理属性约束，构建有效的检测机制；随后，将分析视野拓展至更为复杂的通用合成图像领域，探讨在面对语义内容无限开放且生成痕迹日益隐蔽的挑战下，上述两大检测范式如何进行适应性的演化与策略调整。图4展示了基于检测对象约束强度与语义开放程度的差异进行的双维度展开。一方面，deepfake检测针对具有明确身份与强生物约束的特定对象，其检测策略可充分利用人脸在几何结构、生理节律及身份一致性等方面的强先验；另一方面，通用AI合成图像检测面向语义开放、身份不确定的生成内容，其检测必须依赖更具普适性的统计规律或物理一致性约束。

因此，在两类应用场景中，“证伪”与“存真”并非简单复现，而是在不同对象约束条件下的范式实例化，从而揭示同一学习范式在不同检测语境中的适应性演化。

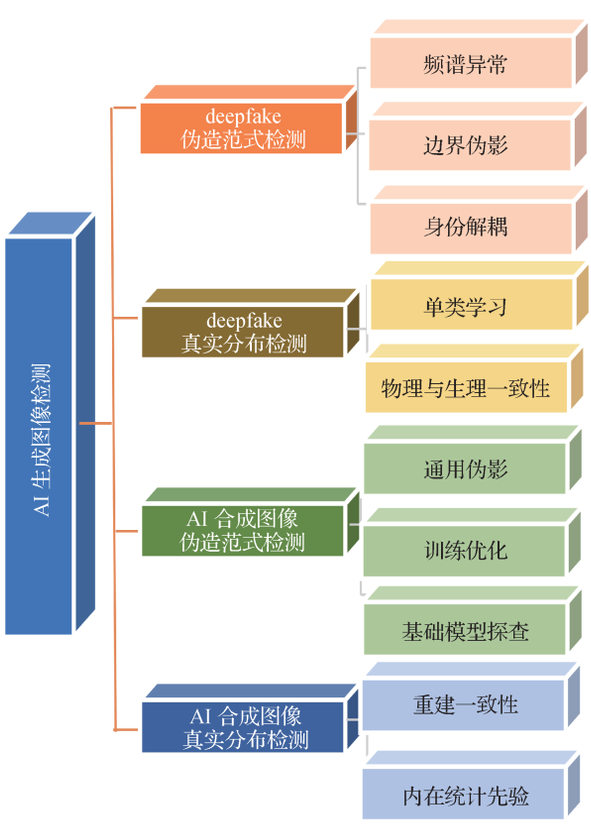


图4　技术分类框架

Fig. 4　Technical classification framework

3.1 面向特定身份的 deepfake 检测

当前的人脸合成技术虽然在视觉真实感上已逼近真实影像，但在维持面部复杂的生理一致性与物理属性约束方面，现有生成模型仍存在难以克服的内在缺陷。相关检测算法正是利用“人脸”这一特定对象所具备的强拓扑结构与物理一致性先验，构建了鲁棒的判别机制。本节将沿着“证伪”与“存真”两条技术范式，对 deepfake 检测方法的演进脉络进行系统梳理。为直观呈现该领域关键文献的发展轨迹及其技术归属，表 2 首先对本节重点讨论的代表性工作进行了系统整理。

3.1.1 基于特定域生成伪影的伪造表征学习

1）频域统计异常分析。如图 5 所示，由于生成模型普遍采用上采样（upsampling）操作，不可避免地破坏了自然图像的频谱统计规律，从而在频率空间内遗留了可被鉴别的伪造指纹（Durall 等，2020）。

Qian 等人（2020）提出了 F^3-Net，主要聚焦于捕捉显性的幅度谱异常。然而，幅度信息往往对 JPEG（joint photographic experts group）压缩等后处理操作高度敏感，限制了其在现实场景中的鲁棒性。针对这一局限，Liu 等人（2021）提出了空间相位浅层学习网络 SPSL（spatial-phase shallow learning）。该方法通过抑制语义信息的干扰，转而探索更具不变性的相位谱特征，显著增强了模型对后处理扰动的泛化能力。

尽管上述方法取得了进展，但依赖预设固定频谱的策略仍存在不足。Luo 等人（2021）提出的 HFF（high-frequency features）与 Kashiani 等人（2025）提出的 FreqDebias 则代表了自适应频谱学习的趋势。它们分别引入注意力机制与频域正则化策略，使模型能自主定位鉴别性最强的频带并校正学习偏差。总体而言，从幅度估计到相位建模，再到自适应学习的演进，本质上是检测技术对抗后处理伪影、提升鲁棒性的系统性优化过程。除了利用频域统计异常，另一类基于“证伪”范式的方法则聚焦于空间域中，因篡改操作直接引发的物理不连续性即篡改边界伪影。

2）篡改边界伪影检测。与频域异常不同，另一类方法直接针对人脸置换过程中的物理不连续性。典型的换脸算法遵循“生成—拼接—融合”的流程，这种操作往往在面部轮廓与背景的交界处留下物理伪影。早期研究多采用监督学习策略显式定位融合边界，但像素级标注的高昂成本限制了其扩展性。鉴于此，研究重心逐渐转向自监督学习。如图 6 所示，Shiohara 和 Yamasaki（2022）提出了自混合影像（self-blended images，SBI）方法，通过在真实图像内混合合成“模拟伪造”样本，迫使模型学习通用的融合边界特征而非特定生成器的偏见，在跨数据集测试中实现了 86.83% 的高准确率。

随后，Chen 等人（2022）提出的 SLADD（self-supervised learning of adversarial example for deepfake detection）与 Jiang 等人（2024）提出的 Bi-LIG（bi-level inconsistency generator）进一步引入了对抗扰动机制，通过生成更高难度的负样本提升判别边界的稳定性。此类方法利用了篡改过程固有的物理缺陷，实现了较强的跨生成器泛化能力。然而，其有效性高度依赖于显式物理拼接操作的存在。因此，在面对全图合成等不具备物理边界的伪造方法时，其检测机制会基本失效。

3）身份解耦与特征重构。现有的检测模型常面临“捷径学习”（Dong 等，2023b）的挑战，即模型倾向于捕捉训练数据中特定身份与真伪标签之间的非因果关联，而非学习本质的伪造特征。

为了解决伪造特征与身份信息纠缠的问题（Guan 等，2024），研究者提出了以身份解耦为核心的方法。这一转变标志着研究重点从隐式表征学习向显式解耦约束的转移。IID（implicit identity driven）（Huang 等，2023）作为早期实例，展示了如何利用对比学习初步实现这种特征分离。结果表明，该方法在多源混合训练下能保持最高性能，尤其在低质量数据上 AUC（area under the curve）提升显著。

在此基础上，后续研究进一步发展出更严格的正交分解策略。如图 7 所示，UCF（uncovering common features）（Yan 等，2023）引入基于多任务学习的正交约束，引导分类器聚焦于与身份正交的通用伪造特征子空间，从而显著提升了模型在跨生成器和跨身份场景下的鲁棒性。

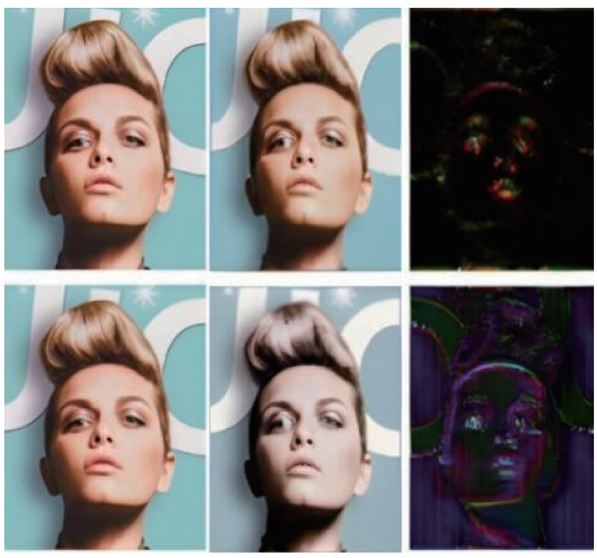
身份解耦方法的关键在于通过分离混淆因子优化特征表示空间，从而缓解导致泛化能力受限的根本问题。然而，解耦策略的有效性依赖于约束强度。过强的约束可能导致与面部生物结构紧密相关的关键伪造线索丢失，从而损害判别能力。因此，如何在抑制捷径学习与保留有效判别信息之间实现最佳权衡，成为当前的一个关键挑战。

表2 deepfake主要检测技术文献

Table2 Key literature on deepfake detection technologies

年份	参考文献	原理	训练数据集	是否开源	开源/论文链接
2020	Durall等人(2020)	频谱伪影	FF++	√	https://github.com/cc-hpc-itwm/UpConv
	Li等人(2020a)	单类学习	FF++	×	https://arxiv.org/pdf/1912.13458v1.pdf
	Qian等人(2020)	频谱伪影	FF++	×	https://arxiv.org/pdf/2007.09355v2.pdf
2021	Liu等人(2021)	频谱伪影	FF++	×	https://arxiv.org/pdf/2103.01856×.pdf
	Luo等人(2021)	频谱伪影	FF++	×	https://arxiv.org/pdf/2103.12376.pdf
	Li等人(2021)	频谱伪影	FF++	×	https://arxiv.org/pdf/2103.09096.pdf
2022	Chen等人(2022)	特征解耦、边界伪影	FF++	√	https://github.com/liangchen527/SLADD
	Shiohara和Yamasaki(2022)	边界伪影	FF++	√	https://github.com/mapooon/SelfBlendedImages
	Cao等人(2022)	单类学习	FF++	×	https://ieeexplore.ieee.org/document/9878441
	Han等人(2023)	单类学习	FF++	×	https://dlnext.acm.org/doi/pdf/10.1145/3591106.3592282
	Korshunov和Marcel(2022)	特征解耦	FF++	√	https://gitlab.idiap.ch/bob/bob.paper.deepfakes_generalization
2023	Yan等人(2023)	特征解耦	FF++	×	https://arxiv.org/pdf/2304.13949.pdf
	Larue等人(2023)	单类学习	FF++	√	https://github.com/anonymous-author-sub/seeable
	Huang等人(2023)	单类学习	FF++	×	https://ieeexplore.ieee.org/document/10204862
	Dong等人(2023b)	特征解耦	FF++	√	https://github.com/megvii-research/CADDM
	Ke和Wang(2023)	特征解耦	Celeb-DF等	×	sciencedirect.com/science/article/abs/pii/S0893608023000011
	Dong等人(2023a)	特征解耦、频谱伪影	FF++	×	https://dlnext.acm.org/doi/10.1016/j.jksuci.2023.03.005
2024	Usmani等人(2024)	特征解耦	DFDC等	×	https://dl.acm.org/doi/10.1145/3702250.3702258
	Jiang等人(2024)	特征解耦、边界伪影	FF++	×	https://ieeexplore.ieee.org/document/10565921
	Yin等人(2024)	物生一致	Celeb-DF DFD等	√	https://github.com/Yzx835/InvariantDomainorientedDeepfakeDetection
	Guan等人(2024)	特征解耦	FF++	×	https://ieeexplore.ieee.org/document/10516609
	Xiao等人(2024)	物生一致	Celeb-DF	×	https://ieeexplore.ieee.org/document/10141892
	Gao等人(2024)	边界伪影	FF++	×	sciencedirect.com/science/article/pii/S0952197624006080
2025	Fu等人(2025b)	特征解耦	FF++	×	https://ojs.aaai.org/index.php/AAAI/article/view/32312
	Kashiani等人(2025)	频谱伪影	FF++	×	https://arxiv.org/pdf/2509.22412.pdf
	Wang等人(2025c)	频谱伪影	FF++	√	https://github.com/wangzhiyuan120/idcnet
	Wang等人(2025a)	物生一致	VGG-Face2	√	https://github.com/Mark-Dou/Forensics
	Li等人(2025a)	频谱伪影	FF++	×	sciencedirect.com/science/article/abs/pii/S0031320325002821
	Shi等人(2025)	物生一致	FF++	×	sciencedirect.com/science/article/abs/pii/S0031320324010501
	Gao等人(2025)	特征解耦	FF++	×	sciencedirect.com/science/article/abs/pii/S0031320325001888
	Batagelj等人(2025)	单类分类	FF++	×	sciencedirect.com/science/article/pii/S240595952500027X
	Bai等人(2025)	物生一致	FF++	×	sciencedirect.com/science/article/abs/pii/S0893608024008384

注:“√”表示代码开源,并提供其仓库链接;“×”表示未开源,所附链接为论文原文地址。



(a) 原始　(b) 输出　(c) 差异

图 5　上采样频谱差异(Durall 等,2020)

Fig. 5　Upsampling spectral difference (Durall et al. ,2020) ((a) original; (b) output; (c) residual)

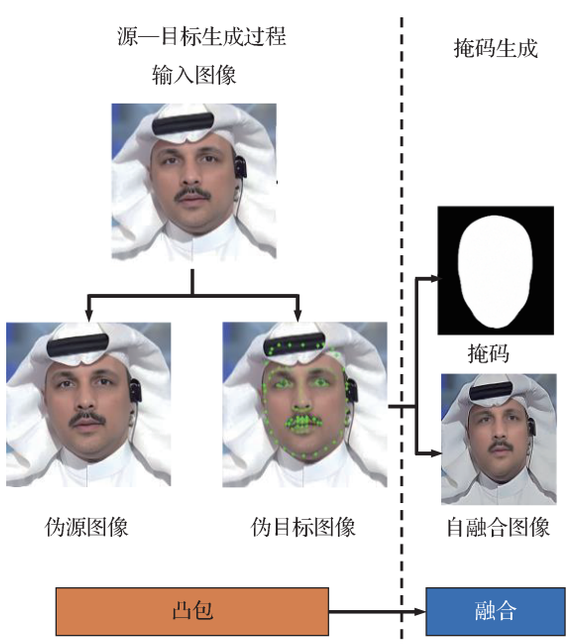


图 6　模拟伪影学习检测

Fig. 6　Learning detection of simulated artifacts

3.1.2　基于人脸流形分布与生物物理属性的真实建模学习

1)基于单类分类的异常检测机制。在“存真”范式理念的指引下,一种重要的方法是将 deepfake 检测重构为单类分类(one-class classification,OCC)或异常检测问题。该范式基于核心假设是:真实人脸数据通常位于一个紧凑的潜在流形上,而任何伪造样本均可视为偏离该分布的离群点。

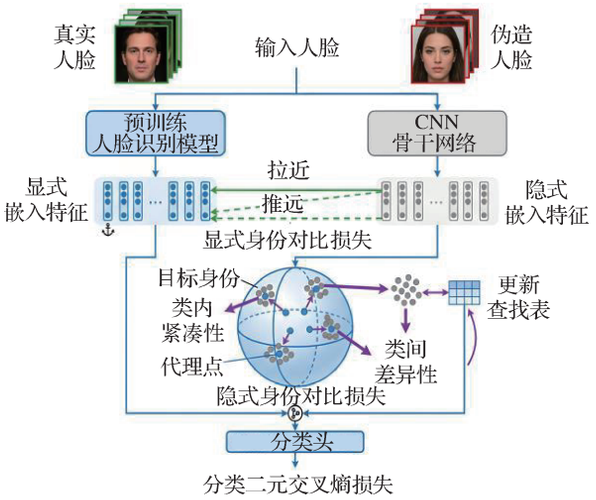


图 7　基于对比学习的身份特征解耦示意图

Fig. 7　Schematic diagram of identity feature disentanglement based on contrastive learning

基于这一逻辑,初始阶段的探索倾向于采用基于生成的异常检测路径。这些方法在仅暴露于真实数据的情况下训练生成模型,如变分自编码器(variational autoencoder, VAE)(Kingma 和 Welling, 2022)、GAN(Goodfellow 等,2020),并利用固定的重建阈值作为决策依据(Shi 等,2025)。虽然这种对伪造数据不可知的设定带来了一定的泛化优势,但缺乏弹性的决策机制往往难以有效表征高质量伪造中存在的微弱统计偏离。为了提升检测的灵敏度与鲁棒性,学术界开始探索自适应度量和连续评分方案。一方面,以 RECCE(reconstruction-classification learning)(Cao 等,2022)为代表的方法引入了联合优化框架,将重建残差内化为一种注意力机制,从而动态地引导分类器关注潜在的异常区域;另一方面,SeeABLE (soft discrepancies and bounded contrastive learning for exposing deepfakes)(Larue 等,2023)等研究则超越了硬阈值划分,转而建立连续的真实性评分体系,有效地提升了对细微伪造瑕疵的辨识能力。基于同源训练条件下,两类方法在 FF++(FaceForensics++)数据集上的 AUC 平均值分别可达 97% 和 98.5%。然而,在更具挑战性的 WildDeepfake 数据集上,RECCE 等方法的性能出现显著下降,表明其泛化能力有待提升。

异常检测方法的主要优势在于其不依赖伪造样本训练,从而具备应对未知攻击的泛化能力。然而,其实际性能高度依赖于对真实数据分布建模的准确性,即模型的分布拟合偏差直接决定了其检测边界。

因此，学习更具鲁棒性和判别力的人脸特征表示，以减少对非典型真实样本的误检，成为当前的一个关键研究方向。Batagelj 等人（2025）提出的多任务自监督框架 M-Task-SS 是这一方向的典型代表。

如图 8 所示，该方法引入了辅助性的自监督学习任务，利用多任务协同优化的策略，引导模型在缺乏伪造样本监督的情况下，仍能敏锐捕捉面部微观纹理与宏观结构的本质一致性。实验表明，通过强化对真实特征的解构能力，该方法显著降低了复杂非受控环境下的误判率，验证了在“存真”范式中引入自监督辅助信号的有效性。

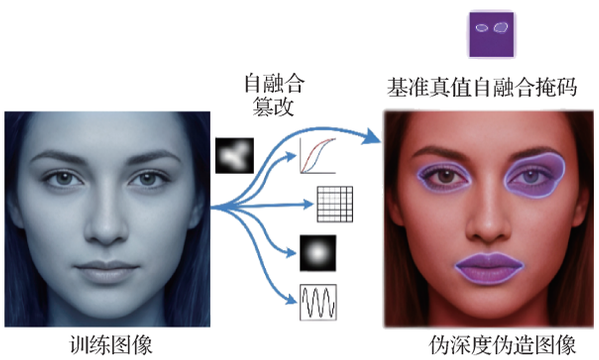


图 8　多任务学习框架

Fig. 8　Multi-task learning framework

2）生物信号与物理几何约束验证。区别于纯数据驱动的异常检测，另一种基于“存真”理念的途径是利用基本的物理或生理原理进行检测。如图 9 所示，该方法的核心假设是：尽管生成模型能够合成逼真的纹理，但在复现复杂的人体生理节律与物理几何约束方面仍存在本质缺陷。这一研究方向经历了一个从“静态几何拓扑”向“动态生理机理”的演进过程。Xie 等人（2025）提出的 GrDT（geometric representation, distribution and texture）通过提取面部关键点的几何拓扑结构，构建了以真实人脸分布为基准的支持向量机（support vector machines，SVM）。该方法利用了真实人脸在解剖学上的结构恒常性，有效识别了生成模型在五官定位与轮廓合成中产生的非自然扭曲。随着生成质量的提升，研究重心逐渐转向更为隐秘的动态生理信号，Ciftci 等人（2022）提出的 FakeCatcher，基于伪造人脸通常难以维持这种时间同步的血流动力学特征，巧妙地利用远程光体积描记技术，捕捉人脸因血流引起的周期性微弱颜色变化，来检测伪造人脸，后者往往表现出生物信号的频域异常或时序断裂。

这种从静态“物理结构”审查向动态“生理存活”验证的技术迭代，为对抗高度逼真的深度伪造开辟了基于生物本质的防御维度。然而，其实用性边界也受到明显制约，特别是在特定场景环境方面设有较高门槛。为此，当前前沿研究正致力于将此类强先验知识与深度神经网络进行更深层次的耦合表达，将几何或生理异常区域转化为神经网络的强引导信号，从而在保留先验知识可解释性的同时，显著增强模型在复杂非受控环境下的鲁棒性。

为了客观评估不同检测范式在应对多样化伪造攻击时的效能边界，本节选取了 FF++（FaceForensics++）、Celeb-DF（Celeb-DeepFakes）（Li 等，2020b）及 DFDC（deepfake detection challenge）等主流基准数据集进行定量对比。在深入对比各算法性能之前，有必要明确检测任务的标准化测试流水线，这构成了理解各文献实验数据的基础。尽管不同研究的细节设置存在差异，但该领域公认的评估流程通常遵循 4 个关键阶段：1）预处理，利用检测器从原始照片定位并提取特征，随后进行对齐与归一化裁剪；2）扰动模拟，为评估鲁棒性，原作者常在输入前引入不同因子的压缩、模糊或噪声干扰；3）模型推理，将处理后的张量输入检测网络，输出该样本为伪造图像的概率分数；4）指标计算，基于全测试集的预测分数与真实标签，计算 ACC（accuracy）、AUC 等核心指标。

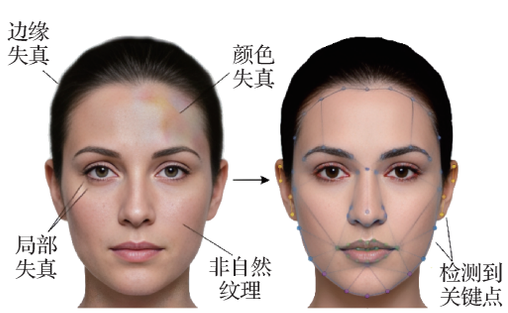


图 9　面部关键点拓扑结构

Fig. 9　Topological structure of facial key point set

需要指出的是，直接横向比较不同文献报告的实验数据存在固有的缺陷，因为各研究在数据预处理及实验设置上可能存在差异。为了最大程度地确保比较的公平性与科学有效性，本节遵循“标准基准优先”原则。表 3 汇总了在相同或高度相似实验协议下的代表性算法性能。基于此，相关对比并非旨在构建严格的算法排名，而是为了揭示不同学习范式在同源测试与跨域泛化场景下的整体演化规律。

首先，“证伪”范式在已知分布下表现出极致的判别精度，但面临严重的过拟合风险。如表3所示，基于边界伪影如Shiohara和Yamasaki（2022）或特征解耦如Yan等人（2023）的方法在与训练集同源的FF++测试中，AUC均突破了99.6%。然而，这种对特定生成痕迹的强依赖导致其泛化鲁棒性脆弱。以Liu等人（2021）频谱分析方法为例，当测试环境迁移至未知的Celeb-DF或DFDC（deepfake detection challenge）时，其性能出现了断崖式下跌，AUC衰减幅度超过20%。这一数据趋势有力地佐证了对特定生成器指纹的过拟合，本质上制约了模型在未知攻击场景下的泛化表现。

其次，“存真”范式在复杂与未知场景下展现出更优的性能韧性。尽管基于单类学习的方法在FF++上的绝对精度略低于SOTA（state-of-the-art）的证伪类方法，但其在更具挑战性的Celeb-DF数据集上却实现了性能反超，同时在跨库测试中保持了更低的衰减率。这表明，通过生理信号（Ciftci等，2022）或重构一致性建立的“真实性准则”，虽然在简单场景下不如暴力特征提取敏锐，但其捕捉的特征更接近人脸的本质属性，从而在异构数据源上具有更强的生存能力。

综上所述，当前的检测技术正处于一个关键的权衡点：是继续在“证伪”的道路上追求特定数据集的刷榜，还是转向“存真”范式以换取更长远的泛化鲁棒性？这一基于数据的实证分析，也直接指引了后文关于“范式融合”这一未来方向的提出。

因此，如何摆脱对特定训练数据的统计依赖，开发能够提取域不变特征并有效泛化至未知分布的鲁棒检测器，已成为该领域亟待解决的问题。

表3　代表性deepfake检测算法在同源与跨域场景下的性能对比评估

Table 3　Performance evaluation of representative deepfake detection algorithms in intra-domain and cross-domain scenarios

范式	技术子类	参考文献	AUC/%			
			FF++	CelebDF	DFDC	平均跨库
“证伪”范式	频率伪影	Kashiani等人（2025）	97.5	83.6	74.1	82.88
	边界伪影	Chen等人（2022）	98.4	79.7	76.05	76.9
	边界伪影	Shiohara和Yamasaki（2022）	**99.64**	93.18	72.42	**87.33**
	频率伪影	Liu等人（2021）	95.32	76.88	66.16	–
	特征解耦	Yan等人（2023）	99.6	82.4	80.5	82.8
“存真”范式	生理信号	Ciftci等人（2022）	94.65（ACC）	91.50（ACC）	–	86.48（ACC）
	单类学习	Larue等人（2023）	98.5	87.3	75.9	83.2
	单类学习	Cao等人（2022）	95.02	**99.94**	**91.33**	74.47

注：加粗字体表示各列最优结果。“–”表示该方法暂无对应指标数据，（ACC）为论文仅提供相关准确率。

3.2　面向通用生成内容的检测

随着生成技术从人脸特定域拓展至任意语义的通用内容演进，检测任务面临的核心挑战转变为跨域泛化。从检测对象的角度看，面向特定身份的deepfake检测可被视为一种强先验约束下的特殊检测场景，其方法设计高度依赖人脸所具备的生物与物理一致性假设；而面向通用生成内容的检测则进一步放宽了对象约束，检测任务由“身份真实性判别”转向“内容来源真实性判别”。这种领域的拓展促使现有的检测方法进行相应的适应性演进。因此，通用生成内容检测并非对deepfake检测的简单替代，而是在前者方法论基础上的泛化延展与约束松弛。两者在检测目标、可利用先验以及方法有效性边界上存在内在关联，这也促使“证伪”与“存真”两种学习范式在不同场景下呈现出差异化的技术实现路径。鉴于此，本文将应对策略归纳为两类：1）基于二分类的“证伪”范式，其核心关注点从挖掘特定领域的伪造痕迹转向提取具有普适性的模型指纹；2）基于异常检测的“存真”范式，重点从特定对象属性建模扩展至自然图像普适物理与统计规律的挖掘。为了概述通用内容检测领域的研究进展，表4归纳了该领域的代表性工作。基于上述背景，下文

表4 通用生成图像主要检测技术

Table 4 Key detection technologies for general-purpose generated images

年份	参考文献	原理	训练数据集	是否开源	开源/论文链接
2020	Dzanic等人(2020)	通用伪影	FFHQ等	×	https://arxiv.org/pdf/1911.06465v2.pdf
	Mi等人(2020)	通用伪影	PGGAN	×	https://ieeexplore.ieee.org/abstract/document/9093190
2022	Jeong等人(2022)	通用伪影	ProGAN等	√	https://github.com/SamsungSDS-Team9/BiHPF
	Dong等人(2022)	通用伪影	BigGAN	×	www.comp.polyu.edu.hk/~csajaykr/deepdeepfake.htm
	Chandrasegaran等人(2022)	通用伪影	ForenSynths	√	https://arxiv.org/pdf/2208.11342.pdf
2023	Epstein等人(2023)	训练优化	Midjourney等	×	https://arxiv.org/pdf/2310.15150.pdf
	Ojha等人(2023)	基础模型	ProGAN	√	https://github.com/Yuheng-Li/UniversalFakeDetect
	Ma等人(2023)	统计先验	CelebA等	×	https://icml.cc/virtual/2023/26129
	Sha等人(2023)	基础模型	MSCOCO	×	https://dl.acm.org/doi/abs/10.1145/3576915.3616588
2024	He等人(2024)	统计先验	无训练	√	https://arxiv.org/pdf/2405.20112.pdf
	Yu等人(2024)	训练优化	Midjourney等	×	https://dl.acm.org/doi/10.1145/3664647.3680776
	Sarkar等人(2024)	统计先验	Kandinsky-v3	√	https://projective-geometry.github.io/
	Sinitsa和Fried(2024)	通用伪影	Midjourney等	√	https://sergo2020.github.io/DIF/
	Alam等人(2024)	通用伪影	Diffusion	×	https://dl.acm.org/doi/abs/10.1145/3627673.3680085
	Ricker等人(2024)	重建一致	Kandinsky 2.1等	√	https://github.com/jonasricker/aeroblade
	Liu等人(2024a)	基础模型	ProGAN	×	https://ieeexplore.ieee.org/document/10657197
	Li等人(2024b)	通用伪影	PolarDiffShield	√	https://github.com/li-yanhao/masksim
	Tan等人(2024)	通用伪影	ForenSynths	√	github.com/chuangchuangtan/NPR-DeepfakeDetection
	Li等人(2024a)	通用伪影	GenImage等	×	https://dl.acm.org/doi/10.1145/3664647.3689003
	Wang和Kalogeiton(2025)	重建一致	GenImage	√	lix.polytechnique.fr/vista/projects/2024_detector_wang
	Fu(2024)	通用伪影、训练优化	GenImage等	×	https://dl.acm.org/doi/10.1145/3664647.3689002
2025	Shen等人(2025)	重建一致	GenImage	×	https://ieeexplore.ieee.org/document/10890455
	Huang等人(2025)	基础模型	SID-Set等	√	https://hzlsaber.github.io/projects/SIDA/
	Wang等人(2025b)	基础模型	SD1.5等	√	https://github.com/iamwangyabin/OpenSDI
	Fu等人(2025a)	统计先验	SDv1.4等	×	https://icml.cc/virtual/2025/poster/44442
	Karageorgiou等人(2025)	统计先验	LatentDiffusion	√	https://mever-team.github.io/spai
	Li等人(2025b)	训练优化	ProGAN	√	https://github.com/Ouxiang-Li/SAFE
	Chu等人(2025)	重建一致	Diffusion、Forensics	×	https://arxiv.org/pdf/2412.07140.pdf
	Yang(2025)	通用伪影	COCO	×	https://arxiv.org/pdf/2507.02995.pdf
	Choi等人(2025)	训练优化、基础模型	A-GenImage	×	https://arxiv.org/pdf/2507.02995.pdf
	Guillaro等人(2025)	重建一致	SD1.5	√	github.com/grip-unina/B-Free
	Tao等人(2025)	训练优化	ForenSynths等	√	https://github.com/rstao-bjtu/SAGNet/
	Jia等人(2025)	通用伪影	ProGAN等	×	ieeexplore.ieee.org/document/11093000/media#media
	Deng等人(2025)	通用伪影	CASIA等	×	https://doi.org/10.1016/j.engappai.2025.111325
	Tan等人(2025)	通用伪影	ForenSynths	×	https://doi.org/10.1016/j.asoc.2025.113873

注:“√”表示代码开源,并提供其仓库链接;“×”表示未开源,所附链接为论文原文地址。

将详细阐述这些方法在应对高泛化场景时的技术策略与贡献。

3.2.1　基于深度生成架构固有缺陷溯源的伪造表征学习

1)跨语义通用生成指纹提取。为有效追溯跨语义的通用生成指纹，学术界主要从频域特性、机理溯源和统计规律3个互补维度展开了深入探索。

具体而言，在频域分析方面，研究者致力于利用深度生成模型普遍存在“高频频谱衰减”这一固有缺陷(Mi等，2020)，来鉴别真伪图像。早期工作Synthbuster(Bammey，2024)主要依赖预设的固定滤波器提取高频残差，而MaskSim(Li等，2024b)则逐步演进为利用可训练掩码实现自适应精细化频带学习的动态策略，以提升检测的灵活性和准确性。在生成机理溯源方面，研究逐步深入网络架构引发伪影的内在本质。研究视角从最初关注表层的像素共生矩阵分析(Nataraj等，2019)，转移到LGrad(learning on gradients)(Tan等，2023)中在梯度空间放大细微异常信号，并最终发展为NPR(neighboring pixel relationships)(Tan等，2024)等通过建模局部像素相关性从根本上揭示上采样等核心操作引入的结构性伪影，实现了对未知生成模型架构优异的泛化检测能力。此外，在普适统计特征方面，相关研究指出，与自然图像相比，生成图像的DCT(discrete cosine transform)系数分布往往违背本福特定律(Bonettini等，2021)，并且在不同颜色通道间表现出显著的系统性偏差(Jia等，2025)，这些统计异常为检测提供了有效线索。

该类方法因不依赖特定语义特征，在理论上具备对任意生成内容进行鉴别的潜力。虽然上述基于通用指纹的方法因其所捕获痕迹的普适性而展现出广阔的泛化前景，然而，在生成与检测持续演进的动态博弈中，任何被显式定义或参数化的特定痕迹，理论上都面临着被新一代生成模型通过对抗训练等手段进行针对性规避或“修复”的系统性风险。

鉴于此，为了构建更持久的防御能力，研究重点已转向两个互补的方向：(1)建立特征融合的集成防御机制(Alam等，2024；Meng等，2024)，如图10所示，待检测样本在多模态特征构建过程中，势必经历频域的径向积分与空域的象限位移变换，而在随后的特征联合阶段，就通过融合不同视域下的频谱衰减差异与空间相关性异常，实现了检测鲁棒性的显著提升；(2)向更本质的生成机理溯源，聚焦于LGrad(Tan等，2023)和NPR(Tan等，2024)等工作所揭示的架构性固有缺陷，这些深层痕迹是难以通过对抗训练完全消除的，因此提供了更为鲁棒的检测依据。

2)去偏见学习与数据源净化的训练优化。除了挖掘更鲁棒的伪造痕迹，从训练过程本身入手，消除模型学习中的偏见，是提升其泛化能力的另一条途径。其核心目标在于引导模型剥离语义内容等“虚

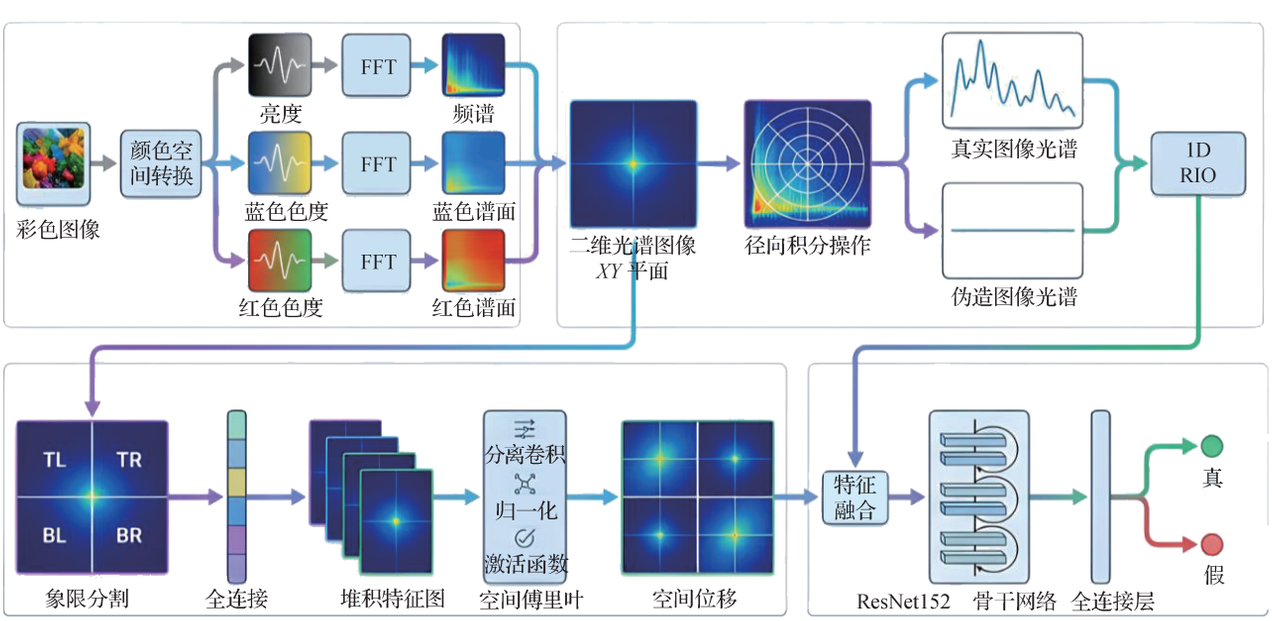


图10　多源特征融合原理图

Fig. 10　Schematic diagram of multi-source feature fusion

假相关性”。如图11所示，这一领域的检测研究可大致分为特征解耦、数据增强和数据净化3个层面。针对内容偏差导致的泛化难题，研究者在不同层面探索了解决方案。

首先是在特征表示层面进行显式解耦尝试，CADDM（common artifact deepfake detection model）（Dong等，2023b）利用专用注意力模块引导模型聚焦于局部伪影区域，而SAGNet（semantic-agnostic artifact network）（Tao等，2025）则采用对抗性训练策略，通过潜在空间的正交分解机制分离内容信息与伪造信号，增强了跨库检测的稳定性。进一步，策略转向通过构建高多样性的样本分布以打破特定的虚假相关性。LSDA（latent space data augmentation）（Yan等，2024）利用潜在空间的流形变换生成多样化的伪造特征分布，DRCT（diffusion reconstruction contrastive training）（Chen等，2024）则结合扩散模型生成的“难负例”与对比学习框架，以掌握更鲁棒的判别边界。这些策略共同缓解了因训练数据分布单一而导致的过拟合，显著提升了对未知攻击的泛化能力。更为根本的策略则是直接在数据源层面进行净化，B-Free（Guillaro等，2025）通过构建与真实图像内容完全匹配的合成样本对，消除了数据集中的固有内容偏差，从而创建一个迫使模型专注于通用伪影模式的理想化训练环境，从根本上提升了检测特征的纯净度。

综上所述，这类去偏见优化方法共同面临着“内容与伪影”的权衡困境。这种矛盾推动了解决方案从最初在特征空间的后处理解耦，发展到数据层面的增强与对源头数据的净化的迈进，以期从源头上解决模型的泛化瓶颈。



图11　训练优化

Fig. 11　Training optimization

3）基础模型的迁移与适配检测。近年来，以CLIP（contrastive language-image pre-training）（Radford等，2021）为代表的视觉—语言基础模型（vision-language models，VLM）为“证伪”范式提供了全新思路。凭借在大规模图文数据集上预训练所获取的丰富视觉先验，VLM能够捕捉传统卷积神经网络（convolutional neural network，CNN）（Lecun等，1998）难以察觉的细微语义冲突与纹理异常，从而显著提升了检测效能，针对VLM在检测中的应用，研究者探索了不同层次的适配策略。如图12所示，初期探索集中于直接利用VLM的冻结特征进行监督学习。UniFD（Ojha等，2023）的工作表明，仅凭VLM的通用表征配合简单线性分类器；即可构建出具备优异跨域泛化性能的检测器，证明了通用视觉先验在伪造检测中的有效性。

为进一步释放模型潜力，策略转向在保留预训练知识的同时进行轻量级适配（Houlsby等，2019）。FatFormer（Liu等，2024a）通过插入轻量级适配器进行微调，实现了在不破坏预训练知识结构的前提下，引导模型专攻特定领域的伪造特征，从而在训练效率与检测精度之间取得了平衡。

更为深入的策略则是将VLM与显式的伪影分析有机融合，以规避单纯监督学习可能导致的捷径学习风险；D^3（discrepancy deepfake detector）（Yang等，2025）即采用此方法，通过融合CLIP特征与结构伪影信号并学习二者差异，强制VLM关注更为底层

的伪造痕迹，增强了模型在面对高逼真生成图像时的判别鲁棒性。总体而言，VLM 所蕴含的通用知识在提升检测性能与泛化能力方面展现出了巨大潜力。这表明，将基础模型的通用视觉知识与“证伪”范式对细微痕迹的专注相结合，是催生新一代高性能、高泛化检测器的有效路径。

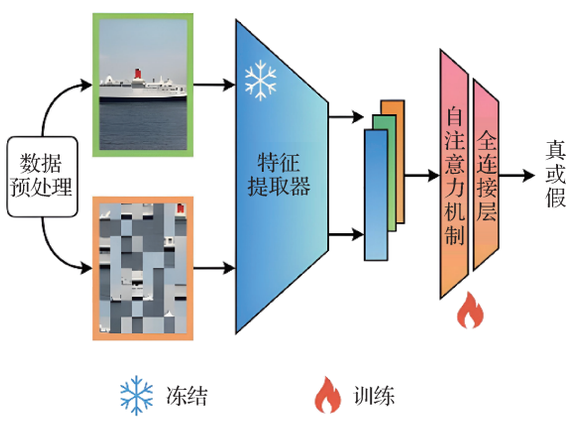


图 12　基础模型检测架构图

Fig. 12　Foundation model-based detection

3.2.2　基于自然图像统计特性与物理先验重构的真实分布建模

1）基于生成先验的重构误差分析。在“存真”体系内，将大规模预训练生成模型逆向应用为检测工具，已发展成为极具影响力的研究分支。该类方法的核心思想在于利用预训练模型对真实数据分布的近似能力，通过分析输入图像在重构过程中的响应差异，建立区分真伪的判别准则。

如图 13 所示，DIRE（diffusion reconstruction error）（Wang 等，2023）率先提出了基于扩散重构误差的检测框架，通过分析图像经“加噪—去噪”过程后的残差图，有效捕获了生成图像在复原过程中的语义丢失现象。虽然该方法取得了优异的检测性能，但其面临着高昂的计算开销及对特定高斯扰动过于敏感的局限。

随后的研究围绕解决这些问题展开。在效率方面，DistilDIRE（Lim 等，2024）利用知识蒸馏技术构建了 DIRE 的轻量化版本，大幅降低了推理延时，提升了实际部署的可行性。在鲁棒性方面，FIRE（frequency-guided reconstruction error）（Chu 等，2025）提出频率引导的主动误差注入策略，通过强化高频细节的重构差异，显著增强了对高质量伪造图像的检测敏感度。此外，在检测机制创新方面，AEROBLADE（Ricker 等，2024）挖掘了潜在扩散模型（latent diffusion models，LDM）自编码器（Rombach 等，2022）的固有属性，实现了无需额外训练的零样本检测；而 Wang 等人（2025a）则探索了重构误差之外的潜在检测信号，进一步拓展了逆向检测的特征维度。

此类逆向应用策略虽然凭借生成模型对真实分布的强大拟合能力，取得了优异的跨域泛化性能，但这本质上使其检测能力“寄生”于特定生成模型之上。其密集的计算需求、对特定预训练权重的依赖，构成了该路径难以规避的根本瓶颈。此外，在生成模型与检测模型架构日益趋同的背景下，两者同源性是否会存在对抗性盲点，仍是一个亟待阐明的安全问题。

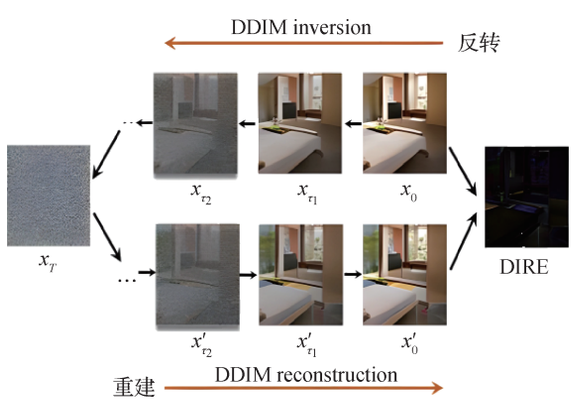


图 13　重建对比检测（Wang 等，2023）

Fig. 13　Reconstruction-contrastive detection (Wang et al, 2023)

2）底层统计规律与宏观物理一致性建模。在“存真”框架下，另一条研究分支致力于直接挖掘自然图像固有的统计与物理规律。利用真实物理世界的底层普适规律构成了生成模型难以完全捕捉的固有约束，可作为区分真伪的可靠基准。

对此类固有规律的探索沿着一条从微观统计到宏观物理、再到抽象信息论的清晰脉络展开。在微观统计层面，LNP（learned noise pattern）（Bi 等，2023）等工作致力于对真实图像固有的噪声模式进行精细建模，借助单类支持向量机（one-class SVM）等异常检测器，实现了对像素级统计异常的有效辨识。在宏观物理层面，研究视角转向全局性的几何与物理一致性线索。Sarkar 等人（2024）证实了生成模型在维持投影几何一致性方面存在系统性缺陷；RIGID（robust AI-generated image detection）（He 等，

2024)则揭示了真实图像对噪声扰动的鲁棒性显著高于生成图像这一统计物理特性,利用该物理感知特征实现了对高逼真生成内容的精准拦截。从统计表征的视角层面,“法医自描述(forensic self-discription)”(Nguyen等,2025)方法通过对预测滤波器的残差进行建模,从信息论视角捕捉真实图像的统计本质,为检测提供了更具解释性的理论支撑。

该方法的发展前景不在于寻求某种“通用”的单一真实性先验,而在于不断拓展“真实性”的维度边界,通过寻找那些比生成模型自身更恒久、更基础的判别依据,构建融合多模态先验的完备自然图像统计模型,以应对不断演进的伪造技术。

为了对上述多元技术路径进行系统性的量化评估,本文构建了涵盖库内性能、跨模态泛化性、鲁棒性、计算效率、训练成本及可解释性六大维度的多维性能评估体系。为了确保评估的科学公正性与可复现性,本研究采用了“实测数据与开源审计”相结合的双重验证策略。

针对库内性能、泛化性与鲁棒性等定量指标,在保证相同或高度相似训练条件下直接采用原文献在标准基准上报告的归一化数值;而针对计算效率、训练成本与可解释性等定性指标,则基于其官方开源代码,通过严格审计网络架构深度、参数规模及推理逻辑复杂度建立分级评分标准,从而将非结构化的架构特性转化为0～100的量化得分。这一评估体系旨在通过标准化的度量衡,客观揭示“证伪”与“存真”两大范式在不同维度上的性能拓扑差异。

如图14所示,两大范式的性能雷达图呈现出截然不同的拓扑形态,揭示了当前技术路线在性能权衡上的深层逻辑。其中,上半区“证伪”范式的雷达图多呈尖锐的放射状,表明该类方法往往在单一指标上具有优异性能,但在抗压缩鲁棒性或计算效率等实用维度上存在显著短板,呈现出鲜明的“专精化”特征;相比之下,下半区“存真”范式则呈现出更饱满的几何形态,验证了其通过建模物理一致性或统计规律,在维持较高精度的同时能更有效地平衡训练开销与鲁棒性,具备更强的落地潜力。值得注意的是,以基础模型为代表的检测方法虽然在泛化性维度得分极高,但往往伴随着较低的计算效率得分,这表明当前领域正面临“高泛化能力”与“实时推理速度”难以兼得的瓶颈。基于此,未来研究可向两个逻辑方向延展:(1)探索混合式模型架构,旨在整合“存真”范式的稳健均衡与“证伪”范式的精准定位能力,利用范式间的特性互补打破性能短板,这种融合并非诞生了第3种范式,而是如何设计高效的协同机制,利用“真”的物理约束来校准“假”的判别边界;(2)研发参数高效适配与知识蒸馏技术,通过优化资源开销缓解大模型在检测任务中的效率瓶颈,

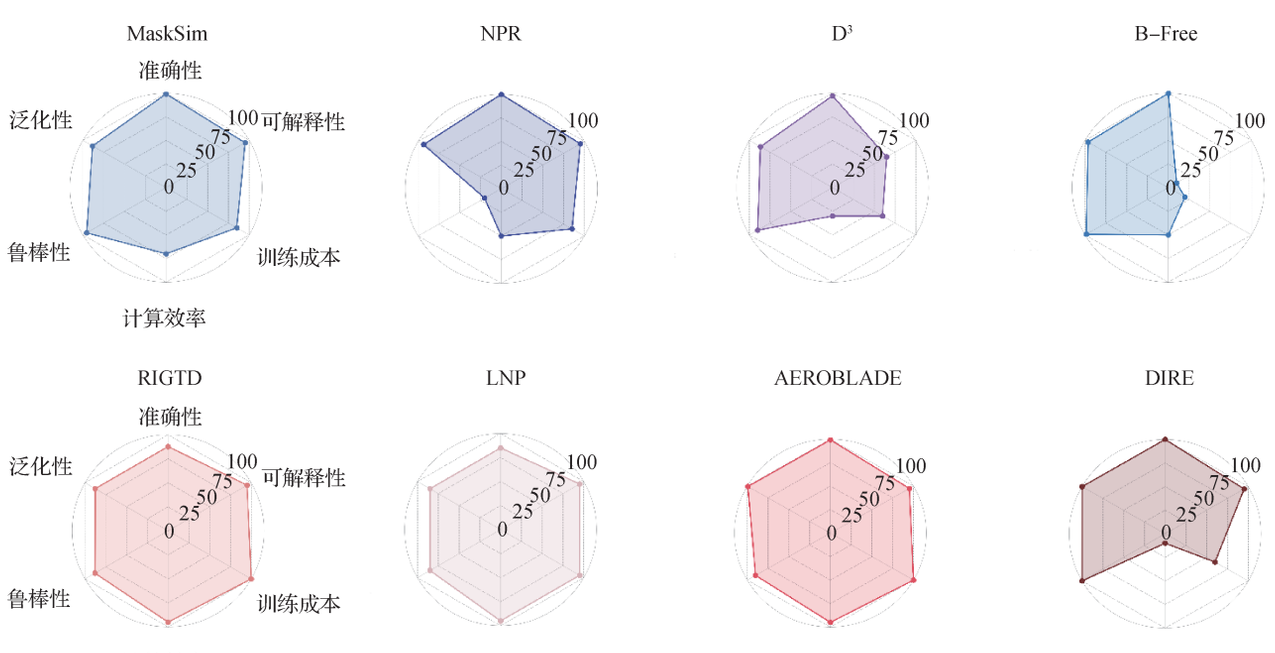


图14　代表性算法的性能雷达图

Fig. 14　Performance radar chart of representative algorithms

推动该领域技术向实用化落地迈进。

4 结 语

为破解当前AI生成图像检测研究视角碎片化、理论体系割裂的困境，本文构建了一个基于“证伪”与“存真”这一认识论范式的统一分析框架。区别于传统综述侧重于对具体算法或模型架构的表层梳理，本文尝试从方法论的深层维度出发，将现有检测技术体系化地归纳为“寻找伪造证据”与“定义真实准则”两大根本性逻辑基准。这一理论重构工作不仅确立了一个逻辑自洽的理论坐标系，更为梳理技术演进脉络、评估方法优劣与未来研究路径提供了一套全新可借鉴的价值标尺，从而构建出更具前瞻洞察力的领域知识图谱。

基于本文框架的系统性分析深刻揭示了AI生成图像检测技术演进呈现出一条从被动追溯特定生成器瑕疵，向主动建模自然图像普适规律演进的清晰轨迹的内在逻辑。这一趋势折射出学术界在与生成技术持续博弈的过程中，正致力于摆脱对单一伪造模式的滞后响应，转而寻求具备更强跨域泛化能力的本质解法。然而，分析同样阐明了当前两大范式均面临着固有的理论瓶颈。

1)“证伪”范式的泛化受限。尽管该类方法在已知特定场景下表现优异，然而其对特定伪影特征的强路径依赖，导致模型在面对未知生成算法或跨域攻击时，往往因特征分布偏移而陷入泛化失效的困境。

2)“存真”范式的建模困难。虽然该范式在应对未知攻击方面具备理论上的先验优势，但受限于真实世界数据分布的高维复杂性与多样性，如何构建精确且计算高效的普适性真值模型仍是当前面临的巨大挑战。

立足于本文构建的双范式框架及其揭示的深层矛盾，未来的研究亟需在以下具有战略意义的方向上寻求突破：

1)超越单一范式，迈向辩证融合。鉴于当前“证伪”与“存真”相对独立的发展状态已难以应对日益复杂的伪造挑战，未来的核心突破口在于实现两大范式的有机协同与统一。(1)“以真鉴伪”，即利用“存真”范式习得的物理一致性与统计规律作为先验知识，指导“证伪”模型剥离内容干扰，聚焦于违背物理规律的本质性破坏；(2)“以伪修真”，即将经实证检验的高频伪影特征或对抗性指纹作为正则化约束，融入“存真”模型的流形学习过程，以提升其对微弱异常的敏感度。通过这种深度的范式互补，有望打破当前泛化性与精确性难以兼顾的博弈僵局，从而构建出既能敏锐捕捉特定伪造痕迹，又能对未知攻击保持稳健防御能力的复合型检测体系。

2)探寻具备恒常性的底层检测原理。为有效应对生成技术的快速迭代与演进，研究重心应逐步从追踪表层伪影转向挖掘独立于特定生成算法的底层核心机理。该方向致力于探索在物理规律、统计定律或信息论视角下具有高度恒常性的判别依据。借此构建的防御体系将不再依赖于易变的生成器指纹，而是锚定于真实世界不可磨灭的固有属性，从而从根本上提升检测技术在对抗生成模型跨代际演进过程中的长期鲁棒性与泛化生存能力。

3)重构高效可信的基础模型应用生态。尽管基础模型(foundation model)已在检测任务中展现出强大的表征能力，然而其庞大的参数规模伴随的高昂计算开销以及决策过程的“黑盒”特性，构成了制约其实际应用的关键瓶颈。鉴于此，未来的研究重心需聚焦于提升模型适配过程中的参数效率，并致力于增强其决策机制的透明度与安全性。这一演进路径旨在攻克当前大模型面临的部署资源受限与信任危机，从而推动其从实验室环境下的“性能标杆”转化为现实场景中兼具经济高效性与透明可信度的实用化检测基础设施。

综上所述，本文构建的“证伪—存真”双范式分析框架，不仅为系统梳理AI生成图像检测技术的演进脉络提供了统一的理论视角，更为关键的是，它为该领域突破现有的泛化瓶颈、实现下一代检测技术的跨越式发展指明了潜在路径。未来研究应致力于实现从单一范式向辩证融合的转变、从表层特征向底层原理的深挖，以及从独立模型向高效基础模型的生态重构。解决上述核心挑战，将是推动该领域构建高鲁棒、可解释且具备持续演进能力的防御体系的关键所在。

参考文献(References)

Afchar D, Nozick V, Yamagishi J and Echizen I. 2018. MesoNet: a compact facial video forgery detection network//Proceedings of

2018 IEEE International Workshop on Information Forensics and Security (WIFS). Hong Kong, China: IEEE: 1-7 [DOI: 10.1109/WIFS.2018.8630761]

Alam I, Muneer M S and Woo S S. 2024. UGAD: universal generative AI detector utilizing frequency fingerprints//Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. Boise, USA: ACM: 4332-4340 [DOI: 10.1145/3627673.3680085]

Alrashoud M. 2025. Deepfake video detection methods, approaches, and challenges. Alexandria Engineering Journal, 125: 265-277 [DOI: 10.1016/j.aej.2025.04.007]

Bai N N, Wang X F, Han R D, Hou J P, Wang Q and Pang S M. 2025. Towards generalizable face forgery detection via mitigating spurious correlation. Neural Networks, 182: #106909 [DOI: 10.1016/j.neunet.2024.106909]

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]

Batagelj B, Kronovšek A, Štruc V and Peer P. 2025. Robust cross-dataset deepfake detection with multitask self-supervised learning. ICT Express, 11(5): 858-862 [DOI: 10.1016/j.icte.2025.02.011]

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]. [2025-12-01]. http://arxiv.org/pdf/2311.00962.pdf

Bonettini N, Bestagini P, Milani S and Tubaro S. 2021. On the use of benford's law to detect GAN-generated images//Proceedings of the 25th International Conference on Pattern Recognition (ICPR). Milan, Italy: IEEE: 5495-5502 [DOI: 10.1109/ICPR48806.2021.9412944]

Cao J Y, Ma C, Yao T P, Chen S, Ding S H and Yang X K. 2022. End-to-end reconstruction-classification learning for face forgery detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 4103-4112 [DOI: 10.1109/CVPR52688.2022.00408]

Chandrasegaran K, Tran N T, Binder A and Cheung N M. 2022. Discovering transferable forensic features for CNN-generated images detection//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 671-689 [DOI: 10.1007/978-3-031-19784-0_39]

Chen B Y, Zeng J S, Yang J Q and Yang R. 2024. DRCT: diffusion reconstruction contrastive training towards universal detection of diffusion generated images//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: JMLR: #297 [DOI: 10.5555/3692070.3692367]

Chen L, Zhang Y, Song Y B, Liu L Q and Wang J. 2022. Self-supervised learning of adversarial example: towards good generalizations for deepfake detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18689-18698 [DOI: 10.1109/CVPR52688.2022.01815]

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 Event: JMLR: #158 [DOI: 10.5555/3524938.3525096]

Choi H, Jung I and Woo S S. 2025. Combating dataset misalignment for robust AI-generated image detection in the real world//Proceedings of the 4th Workshop on Security Implications of Deepfakes and Cheapfakes. Hanoi, Vietnam: ACM: 15-20 [DOI: 10.1145/3709022.3736541]

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]

Ciftci U A, Demir I and Yin L J. 2022. FakeCatcher: detection of synthetic portrait videos using biological signals. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(11): 6984-6997. [DOI:10.1109/TPAMI.2020.3009287]

Cogranne R. 2025. A comparative review of deep learning models for deepfake detection//Proceedings of 2025 International Conference on Advanced Machine Learning and Data Science. Tokyo, Japan: IEEE: 228-235 [DOI: 10.1109/AMLDS63918.2025.11159448]

Deng F, Liu T, Yu H B and Yang R T. 2025. A multi-scale feature and cross-domain fusion network for image tampering localization. Engineering Applications of Artificial Intelligence, 157: #111325 [DOI: 10.1016/j.engappai.2025.111325]

Ding F, Kuang R S, Zhou Y, Sun L, Zhu X G and Zhu G P. 2024. A survey of deepfake and related digital forensics. Journal of Image and Graphics, 29(2): 295-317(丁峰, 匡仁盛, 周越, 孙珑, 朱小刚, 朱国普. 2024. 深度伪造及其取证技术综述. 中国图象图形学报, 29(2): 295-317) [DOI:10.11834/jig.230088]

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]

Dong F K, Zou X Q, Wang J H and Liu X Y. 2023a. Contrastive learning-based general deepfake detection with multi-scale RGB frequency clues. Journal of King Saud University-Computer and Information Sciences, 35(4): 90-99 [DOI: 10.1016/j.jksuci.2023.03.005]

Dong S C, Wang J, Ji R H, Liang J J, Fan H Q and Ge Z. 2023b. Implicit identity leakage: the stumbling block to improving deepfake detection generalization//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 3994-4004 [DOI: 10.1109/CVPR52729.2023.00389]

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 Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates: 3022-3032

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, Kulal S, Blattmann A, Entezari R, Müller J, Saini H, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: JMLR: 12606-12633

Fu H H. 2024. Optimizing AIGC image detection: strategies in data augmentation and model architecture//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM: 11472-11474 [DOI: 10.1145/3664647. 3689002]

Fu X H, Yan Z Y, Yang Z, Yao T P, Zhao Y D, Ding S H, et al. 2025a. PiD: generalized AI-generated images detection with pixel-wise decomposition residuals//Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: JMLR.org: #689 [DOI: 10.5555/3780338.3781027]

Fu X H, Yan Z Y, Yao T P, Chen S and Li X. 2025b. Exploring unbiased deepfake detection via token-level shuffling and mixing//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press: 3040-3048 [DOI: 10.1609/aaai.v39i3.32312]

Gao J, Micheletto M, Orrù G, Concas S, Feng X Y, Marcialis G L, et al. 2024. Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection. Engineering Applications of Artificial Intelligence, 133: #108450 [DOI: 10.1016/j.engappai.2024.108450]

Gao Q, Zhang B P, Wu J H, Luo W X, Teng Z and Fan J P. 2025. Leveraging facial landmarks improves generalization ability for deepfake detection. Pattern Recognition, 164: #111528 [DOI: 10.1016/j.patcog.2025.111528]

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. 2020. Generative adversarial networks. Communications of the ACM, 63(11): 139-144 [DOI: 10.1145/3422622]

Guan W N, Wang W, Dong J and Peng B. 2024. Improving generalization of deepfake detectors by imposing gradient regularization. IEEE Transactions on Information Forensics and Security, 19: 5345-5356 [DOI: 10.1109/TIFS.2024.3396064]

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]

Han B, Li J S, Ren W Q, Luo M, Liu J and Cao X C. 2023. SIGMA-DF: single-side guided meta-learning for deepfake detection//Proceedings of 2023 ACM International Conference on Multimedia Retrieval. Thessaloniki, Greece: ACM: 153-161 [DOI: 10.1145/3591106.3592282]

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]. [2025-12-01]. http://arxiv.org/pdf/2405.20112.pdf

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc.: #574 [DOI: 10.5555/3495724.3496298]

Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, De Laroussilhe Q, Gesmundo A, et al. 2019. Parameter-efficient transfer learning for NLP//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 2790-2799

Huang B J, Wang Z Y, Yang J F, Ai J X, Zou Q, Wang Q, et al. 2023. Implicit identity driven deepfake face swapping detection//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 4490-4499 [DOI: 10.1109/CVPR52729.2023.00436]

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]

Jeong Y, Kim D, Min S, Joe S, Gwon Y and Choi J. 2022. 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]

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]

Jiang P Q, Xie H T, Yu L Y, Jin G Q and Zhang Y D. 2024. Exploring bi-level inconsistency via blended images for generalizable face forgery detection. IEEE Transactions on Information Forensics and Security, 19: 6573-6588 [DOI: 10.1109/TIFS.2024.3417266]

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, 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]

Kashiani H, Talemi N A and Afghah F. 2025. FreqDebias: towards generalizable deepfake detection via consistency-driven frequency debiasing//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 8775-8785 [DOI: 10.1109/CVPR52734.2025.00820]

Ke J P and Wang L N. 2023. DF-udetector: an effective method towards robust deepfake detection via feature restoration. Neural Networks, 160: 216-226 [DOI: 10.1016/j.neunet.2023.01.001]

Khan R, Sohail M, Usman I, Sandhu M, Raza M, Yaqub M A, et al. 2024. Comparative study of deep learning techniques for deepfake video detection. ICT Express, 10(6): 1226-1239 [DOI: 10.1016/j.icte.2024.09.018]

Kingma D P and Welling M. 2022. Auto-encoding variational bayes [EB/OL]. [2025-12-01]. http://arxiv.org/pdf/1312.6114.pdf

Korshunov P and Marcel S. 2022. Improving generalization of deepfake detection with data farming and few-shot learning. IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(3): 386-397 [DOI: 10.1109/TBIOM.2022.3143404]

Larue N, Vu N S, Struc V, Peer P and Christophides V. 2023. SeeABLE: soft discrepancies and bounded contrastive learning for exposing deepfakes//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 20954-20964 [DOI: 10.1109/ICCV51070.2023.01921]

Lecun Y, Bottou L, Bengio Y and Haffner P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11): 2278-2324 [DOI: 10.1109/5.726791]

Li J C, Hu Y J, Liu B B, She H M and Li C T. 2025a. Deepfake detection with domain generalization and mask-guided supervision. Pattern Recognition, 165: #111622 [DOI: 10.1016/j.patcog.2025.111622]

Li J M, Xie H T, Li J H, Wang Z Y and Zhang Y D. 2021. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 6454-6463 [DOI: 10.1109/CVPR46437.2021.00639]

Li L Z, Bao J M, Zhang T, Yang H, Chen D, Wen F, et al. 2020a. Face X-ray for more general face forgery detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5000-5009 [DOI: 10.1109/CVPR42600.2020.00505]

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): 13-44(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): 13-44) [DOI: 10.11834/jig.250053]

Li O X, Cai J Y, Hao Y B, Jiang X L, Hu Y and Feng F L. 2025b. Improving synthetic image detection towards generalization: an image transformation perspective//Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1. Toronto, Canada: ACM: 2405-2414 [DOI: 10.1145/3690624.3709392]

Li S H, Wu H S and Wang B. 2024a. A solution to ACMMM 2024 on artificial intelligence generated image detection//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM: 11475-11477 [DOI: 10.1145/3664647.3689003]

Li Y H, Bammey Q, Gardella M, Nikoukhah T, Morel J M, Colom M, et al. 2024b. 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 Z, Yang X, Sun P, Qi H G and Lyu S W. 2020b. Celeb-DF: a large-scale challenging dataset for DeepFake forensics//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3204-3213 [DOI: 10.1109/CVPR42600.2020.00327]

Lim Y, Lee C, Kim A and Etzioni O. 2024. DistilDIRE: a small, fast, cheap and lightweight diffusion synthesized deepfake detection [EB/OL]. [2025-12-01]. http://arxiv.org/pdf/2406.00856.pdf

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 G, Li X D, Zhou W B, Chen Y F, He Y, Xue H, et al. 2021. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 772-781 [DOI: 10.1109/CVPR46437.2021.00083]

Liu P, Tao Q Q and Zhou J. 2024b. Robust deepfake detection by addressing generalization and trustworthiness challenges: a short survey//Proceedings of the 1st ACM Multimedia Workshop on Multimodal Misinformation Governance in the Era of Foundation Models. Melbourne, Australia: ACM: 3-11 [DOI: 10.1145/3689090.3689386]

Luo Y C, Zhang Y, Yan J C and Liu W. 2021. Generalizing face forgery detection with high-frequency features//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 16312-16321 [DOI: 10.1109/CVPR46437.2021.01605]

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]. [2025-12-01]. http://arxiv.org/pdf/2307.06272.pdf

Mahara A and Rishe N. 2025. Methods and trends in detecting AI-generated images: a comprehensive review [EB/OL]. [2025-12-01]. http://arxiv.org/pdf/2502.15176.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]

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): #art00008 [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//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 3040-3050 [DOI: 10.1109/CVPR52734.2025.00289]

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]

Qian Y Y, Yin G J, Sheng L, Chen Z X and Shao J. 2020. Thinking in frequency: face forgery detection by mining frequency-aware clues//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 86-103 [DOI: 10.1007/978-3-030-58610-2_6]

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 Event: PMLR: 8748-8763

Radford A, Metz L and Chintala S. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks//Proceedings of the 4th International Conference on Learning Representations. San Juan, Puerto Rico: #ICLR

Ricker J, Lukovnikov D and Fischer A. 2024. 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]

Roose K. 2022. An A.I.-generated picture won an art prize. artists aren't happy [EB/OL]. [2025-12-01]. https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html

Sarkar A, Mai H L, 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]

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]

Sharma V K, Garg R and Caudron Q. 2024. A systematic literature review on deepfake detection techniques. Multimedia Tools and Applications, 84(20): 22187-22229 [DOI: 10.1007/s11042-024-19906-1]

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]

Shi L, Zhang J, Ji Z L, Bai J F and Shan S G. 2025. Real face foundation representation learning for generalized deepfake detection. Pattern Recognition, 161: #111299 [DOI: 10.1016/j. patcog. 2024. 111299]

Shiohara K and Yamasaki T. 2022. Detecting deepfakes with self-blended images//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18699-18708 [DOI: 10.1109/CVPR52688.2022.01816]

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]

Sunil R, Mer P, Diwan A, Mahadeva R and Sharma A. 2025. Exploring autonomous methods for deepfake detection: a detailed survey on techniques and evaluation. Heliyon, 11(3): #e42273 [DOI: 10.1016/j.heliyon.2025.e42273]

Tan C C, Liu H, Zhao Y, Wei S K, Gu G H, Liu P, et al. 2024. 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, 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]

Tan D T, Yang D Y, Tan B A, Niu C Y, Yang Y and Li S C. 2025. AOT-PixelNet: lightweight and interpretable detection of forged images via adaptive orthogonal transform. Applied Soft Computing, 185: #113873 [DOI: 10.1016/j.asoc.2025.113873]

Tao R S, Tan C C, Liu H, Wang J K, Qin H T, Chang Y, 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]

Usmani S, Kumar S and Sadhya D. 2024. Enhancing generalization ability in deepfake detection via continual learning//Proceedings of the 15th Indian Conference on Computer Vision Graphics and Image Processing. Bengaluru, India: ACM: #18 [DOI: 10.1145/3702250.3702258]

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. Red Hook, USA: Curran Associates Inc.: 6000-6010 [DOI: 10.5555/3295222.3295349]

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

Wang H Y, Liu Z H and Wang S L. 2025a. MIFAE-forensics: masked image-frequency AutoEncoder for deepfake detection//Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing. Hyderabad, India: IEEE: 1-5 [DOI: 10.1109/ICASSP49660.2025.10889125]

Wang R Y, Chu B L, Yang Z and Zhou L N. 2022. An overview of visual DeepFake detection techniques. Journal of Image and Graphics, 27(1): 43-62 (王任颖, 储贝林, 杨震, 周琳娜. 2022. 视觉深度伪造检测技术综述. 中国图象图形学报, 27(1): 43-62) [DOI: 10.11834/jig.210410]

Wang X and Kalogeiton V. 2025. Your diffusion model is an implicit synthetic image detector//Del Bue A, Canton C, Pont-Tuset J, Tommasi T, eds. Computer Vision——ECCV 2024 Workshops. Milan, Italy: Springer: 418-434 [DOI: 10.1007/978-3-031-92648-8_25]

Wang Y B, Huang Z W and Hong X P. 2025b. 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 Y, Chen Y X, Yao Y Z, Han M, Xing W P and Li M. 2025c. IDCNet: image decomposition and cross-view distillation for generalizable deepfake detection. IEEE Transactions on Information Forensics and Security, 20: 8373-8386 [DOI: 10.1109/TIFS.2025.3593353]

Xiao S, Lan G P, Yang J C, Lu W, Meng Q G and Gao X B. 2024. MCS-GAN: a different understanding for generalization of deep forgery detection. IEEE Transactions on Multimedia, 26: 1333-1345 [DOI: 10.1109/TMM.2023.3279993]

Xie H Y, He H Y, Fu B Y and Sanchez V. 2025. GrDT: towards robust deepfake detection using geometric representation distribution and texture//Proceedings of 2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. Tucson, USA: IEEE: 686-696 [DOI: 10.1109/WACVW65960.2025.00083]

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]

Yan Z Y, Luo Y H, Lyu S W, Liu Q S and Wu B Y. 2024. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8984-8994 [DOI: 10.1109/CVPR52733.2024.00858]

Yan Z Y, Zhang Y, Fan Y B and Wu B Y. 2023. UCF: uncovering common features for generalizable deepfake detection//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22355-22366 [DOI: 10.1109/ICCV51070.2023.02048]

Yang G. 2025. FreqCross: a multi-modal frequency-spatial fusion network for robust detection of stable diffusion 3.5 generated images//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision. Honolulu, USA: IEEE: 6577-6584 [DOI: 10.1109/ICCVW69036.2025.00685]

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]

Yao W D, Li P C, Zhao Y and Wu H C. 2025. Review of research on face deepfake detection methods. Journal of Image and Graphics, 30(7): 2343-2363 (姚文达, 李盼池, 赵娅, 吴洪超. 2025. 人脸深度伪造检测方法研究综述. 中国图象图形学报, 30(7): 2343-2363) [DOI:10.11834/jig.240586]

Yin Z X, Wang J K, Xiao Y S, Zhao H Q, Li T L, Zhou W B, et al. 2024. Improving deepfake detection generalization by invariant risk minimization. IEEE Transactions on Multimedia, 26: 6785-6798 [DOI: 10.1109/TMM.2024.3355651]

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]

作者简介

李佳业, 男, 硕士研究生, 主要研究方向为图像取证。
E-mail: lijiaye0605@163.com
张敏情, 通信作者, 女, 教授, 博士生导师, 主要研究方向为图像取证。E-mail: api_zmq@126.com
黄思远, 男, 博士研究生, 主要研究方向为图像取证。
E-mail: huangsiyuan828@163.com
冯佩, 女, 副教授, 主要研究方向为图像取证。
E-mail: 83101958@qq.com
刘浪, 男, 硕士研究生, 主要研究方向为智能体信息隐藏。
E-mail: 768014672@qq.com