Vol.7 Iss.3 2026 Communication & Education Review
https://doi.org/10.37420/j.cer.2026.035

A Review of AI-Generated Content Detection Technologies: From DeepFake Recognition to Multimodal Provenance

Lei Weibin*
Department of Broadcasting, Henan Radio and Television Station
No. 2 Weiyi Road, Jinshui District, Zhengzhou City, Henan Province, China
Corresponding author, E-mail: lucyboy2022@163.com

Abstract:
Generative adversarial networks (GANs), diffusion models, and multimodal synthesis pipelines have reached a level where machine-produced images, audio-visual clips, and cross-platform narratives can closely mimic authentic content. Detection strategies built around isolated visual artifacts are, as a consequence, losing much of their former effectiveness. We survey representative work from IEEE Xplore, Web of Science, and CNKI covering the period 2018–2025, drawing on 38 references—28 core studies plus 10 supplementary sources on benchmarks, methodological norms, and thematic surveys—to build an analytical framework organized by detection target, forensic feature, model architecture, and deployment context. A clear trajectory emerges from the literature: early detectors that relied on local texture descriptors and frequency-domain cues have gradually given way to more integrated forensic approaches, including real-face representation learning, cross-dataset robustness enhancement, audio-visual consistency verification, and generation-source attribution. While multimodal fusion and open-set provenance analysis broaden the reach of detection to complex real-world settings, practical obstacles—platform compression, adversarial perturbation, real-time throughput limits, and the disconnect between algorithmic outputs and governance workflows—are far from resolved. Building on these findings, we map current technical routes, trace their evolutionary logic, and outline considerations for detection system design in social media governance contexts.

Keywords:
AI-generated content detection; DeepFake; diffusion models; multimodal forensics; source attribution; cross-dataset generalization; social media governance

1 引言

1.1 研究背景与问题提出

过去五年间，以 GAN 和扩散模型为代表的生成式 AI 经历了数轮迭代，图像合成、语音克隆乃至视频编辑的操作门槛被大幅拉低。多篇针对人脸替换和表情交换的综述研究已注意到，机器合成的视觉内容在保真度上与真实采集素材之间的差距正在快速缩小 [11,15]。在此背景下，DeepFake 不再局限于早期的人脸替换，而是逐步扩展到表情驱动、语音模仿和跨模态联动生成；与此同时，围绕生成图像来源识别的研究也表明，检测任务正在由单纯识别真伪扩展到生成源归因与责任追踪 [23]。

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

相关风险也已从实验环境进入实际传播链条。针对深度伪造媒体取证现状与未来挑战的综述指出，伪造内容可被嵌入政治传播、商业营销、身份冒用和网络诈骗等场景，影响公众判断并削弱平台信任机制 [3]。与此同时，AIGC 在新闻辅助生产、短视频创作和数字版权流通中的应用不断增多，关于 AI 生成内容挑战与解决路径的综述进一步表明，研究重点已从“是否伪造”延伸到来源追踪、责任认定与版权归属等治理议题 [1,23]。因此，检测研究正在由单一识别任务转向兼顾发现、解释与溯源的复合型技术体系。

1.2 现有综述局限与本文切入点

已有综述为理解 DeepFake 检测奠定了基础，但其研究视角仍以单模态视觉任务为主，对音视频联合伪造、文本—图像协同误导以及跨平台传播后的溯源问题讨论不足。尤其是 2022 年以来扩散模型快速进入主流应用后，生成痕迹分布与既有 GAN 伪影存在明显差异，部分早期综述已难以准确概括当前检测对象与方法谱系。

出于这一考虑，本文选择将检测与溯源放进同一个分析视角，而不是分开讨论——具体地说，是从生成机制、伪造类型、检测线索到部署场景，搭建一条贯通的考察路径。我们在 IEEE Xplore、Web of Science 和 CNKI 中检索了 2018 至 2025 年间的相关成果，最终选定 38 篇文献作为分析基础：其中 28 篇直接服务于主体论证，另外 10 篇则提供基准数据集、方法学规范或专题综述方面的补充参照 [27-31]。

全文结构安排如下：第 1 章说明研究背景、问题界定与综述切入点；第 2 章梳理 DeepFake 生成机制及相应检测路径的演进；第 3 章讨论多模态检测、生成源归因与社交媒体场景中的部署问题；第 4 章在比较现有路线的基础上总结瓶颈，并提出后续研究关注点。

为提高综述过程的可重复性，本文参照 PRISMA 2020 系统综述框架组织检索与筛选流程 [27]。英文检索词主要包括“deepfake detection”“synthetic media forensics”“multimodal provenance”“diffusion fake”“source attribution”等，并采用 AND/OR 逻辑进行组合；中文检索词主要包括“深度伪造检测”“多模态取证”“生成内容溯源”“扩散模型伪造”等。初检共获得约 240 篇文献，去重后保留约 160 篇，经标题与摘要相关性筛选后缩减至 80 篇左右。纳入标准为：（1）聚焦 AI 生成内容检测或溯源的原创研究或系统综述；（2）发表于同行评审期刊、顶级会议或高引预印本；（3）发表时间为 2018—2025 年。排除标准为：（1）仅涉及生成模型而不涉及检测或溯源的研究；（2）摘要或全文无法获取的文献；（3）研究内容与本文主题重叠度不足的边缘性工作。最终结合全文质量评估纳入 38 篇文献（28 篇核心研究与 10 篇背景支撑文献），形成本文的分析基础。

2 DeepFake 生成机制与检测技术演进

DeepFake 技术在过去几年里分化出了多条并行的发展路线。目前最主要的三类生成机制分别是：以对抗博弈驱动的 GAN 路线，通过反复去噪实现合成的扩散模型路线，以及借助三维重建和图形学渲染完成拼接的传统混合路线。由于这三类路线在频谱、纹理和时序特征上各自留下不同类型的痕迹，检测端也需要有针对性地选择分析策略。

从伪造操作类型看，研究中常见任务包括人脸替换、表情重演、属性篡改和全脸合成。前两类更容易在身份一致性、口型同步和局部边缘区域暴露异常，后两类则更多表现为纹理分布、光照关系或高频细节的不自然偏移。Amerini 等 [3] 在关于深度伪造媒体取证现状与未来挑战的综述中指出，生成模型能力的持续增强正在压缩传统视觉取证的有效空间；Guarnera 等 [16] 围绕公开检测挑战赛的总结也强调，多媒体操控工具的易得性显著提升，使检测任务转向面向大规模流通内容的系统比

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

较与治理问题。

Malik 等 [15] 在针对人脸图像与视频深度伪造检测的综述中进一步表明，随着伪造工具面向普通用户普及，检测任务已由“识别某类固定伪影”转向“辨认跨模型、跨场景、跨后处理条件下的合成规律”。这意味着仅依赖单一特征的分类器越来越难以覆盖复杂传播环境。

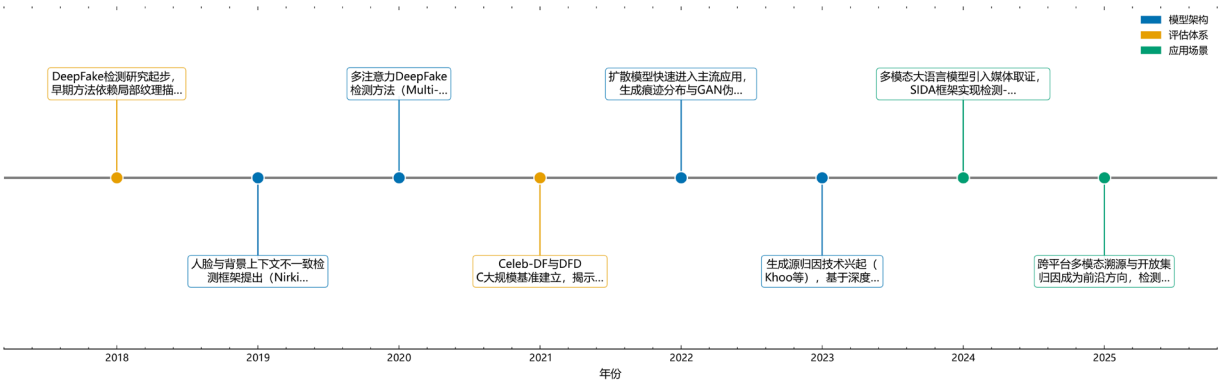


图 1 AI 生成内容检测技术发展时轴

2.1 基于视觉特征的检测方法

在所有检测路径中，视觉分析的成熟度相对最高，不过内部方法之间的差异也最为显著。如果按照检测线索的抽象程度来看，大致可以区分出三个层面：底层特征方法侧重像素统计和频域异常，对压缩伪影比较敏感；表征学习方法则更关注眼部、口部等语义区域的细节差别；关系建模方法把重心放在人脸与周围环境的协调性上，试图抓住时空关系中的不自然之处。

在语义级建模路线中，多注意力机制具有代表性。Zhao 等提出的 Multi-attentional Deepfake Detection 方法 [18] 注意到伪造痕迹并不会稳定集中于单一区域，因此通过多注意力网络同时关注眼周、口周、轮廓边界等多个局部区域，以降低模型对单一显著区域的依赖。这类方法的优势在于能够提升细粒度异常的覆盖度，但对训练数据标注质量和场景分布较为敏感。在 FaceForensics++（HQ）基准上，该方法的帧级 AUC 达到 99.29%，显著优于单注意力基线；然而在跨数据集评测（如 Celeb-DF）中，性能下降较为明显，反映出域内高精度与域外泛化之间的张力 [24]。

关系不一致性检测则强调“伪造区域与环境之间是否协调”。Nirkin 等 [19] 提出基于人脸与上下文差异的检测框架，将人脸区域与背景区域联合建模，检验二者在纹理、光照、边界和空间关系上的偏差。这条路线之所以有效，是因为换脸拼接通常很难把面部区域与原始背景在光照、纹理上做到完美衔接——只要检测器同时观察面内与面外区域，就有机会发现这种跨区域的不连续。

从实验结果的横向对比来看，没有哪一类方法能在所有维度上占据绝对优势。多注意力网络的长处在于覆盖面——它能够同时关注多个面部区域的语义异常，但换到域外数据时效果波动比较大，计算成本也不算低。人脸与背景不一致检测走的是另一条路，重点看拼接边界附近的光照和纹理是否协调，好处是输出结果比较容易解释，复杂度也相对可控。至于二元分类器，部署门槛最低、推理最快，但抗数据分布偏移的能力明显不足，数据集一换指标就容易缩水。

由于不同研究使用的数据集、压缩设置和评测协议并不一致，现阶段尚难对上述方法给出完全统一的量化排序，因此更适合从检测线索、适用场景与部署代价三个方面进行横向比较。从公开基准的使用情况看，FaceForensics++ 与 DFDC 已成为视觉检测研究中的两类核心评测锚点：前者更适合

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

受控条件下的人脸操控识别，后者更强调大规模、竞赛式和真实扰动条件下的模型比较 [28-29]。已有研究表明，基于 Xception 或注意力增强的视觉检测器在数据集内评测中通常能够取得较高指标，但学习域内一致性与域间差异的泛化框架 [32]、结合频率掩码的 Transformer 适配方法 [33]、基于可解释原型的 ExplaNET 框架 [34] 以及针对视觉深度伪造检测技术与局限的系统综述 [35] 都共同指出：在跨压缩、跨生成器和跨数据集设置下，模型性能仍会出现明显退化。例如，基于 Xception 架构的检测器在 FaceForensics++（HQ）上可达到 99% 以上的视频级准确率，但在 Celeb-DF 跨库测试中 AUC 降至 65.7% 左右 [24]，在 WildDeepfake 等野外场景中衰减更为显著 [25]。这意味着，受控基准上的优异结果并不能直接等同于真实平台环境中的稳定取证能力。

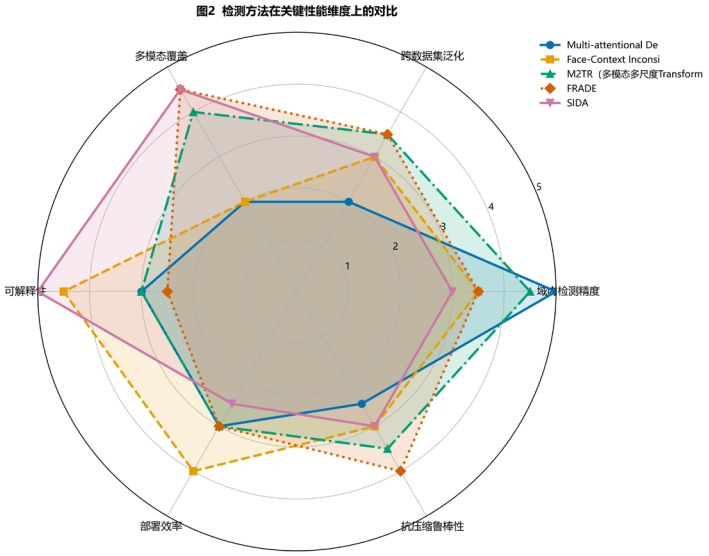


图 2 检测方法在关键性能维度上的对比

2.2 跨数据集泛化与对抗鲁棒性

跨数据集泛化不足，是当前检测系统进入真实业务场景时最先暴露的问题。Das 等 [22] 围绕利用动态人脸增强提升检测能力的分析指出，许多模型在封闭数据集上可以取得较高指标，但当伪造算法、压缩方式或拍摄条件发生变化时，性能往往明显衰减。这说明相当一部分检测器学习到的仍是数据集特有伪影，而非具有迁移能力的生成规律。

对抗鲁棒性问题进一步放大了上述不足。若模型仅输出“真 / 伪”二元标签，就很难解释伪造证据来自何处，也难以支撑后续生成源归因。关于深度伪造来源识别的研究表明，随着攻击者主动采用后处理、重压缩与局部修复策略，单纯依赖静态分类边界的检测器很难同时保持稳定识别与可靠归因能力 [23]。

不过，伪造威胁早已不只停留在视觉层面。Sadiq 等 [12] 对社交媒体中机器生成推文的识别研究就说明了这一点——在真实的攻击场景里，伪造视频、合成语音和生成文本经常被打包使用，形成多模态协同的操控链条。如果只盯住单一模态，很容易顾此失彼。

由此可见，当前领域至少存在三类持续张力：其一，生成模型迭代速度快于检测器更新速度，导致方法生命周期被不断压缩；其二，跨域泛化与特定场景精度之间仍需权衡，难以同时达到；其三，平台压缩、对抗扰动与传播链后处理共同削弱了检测输出的稳定性。未来研究若要面向开放世界部署，需要同步推进鲁棒表征学习、域外适应与多模态证据融合。

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

3 多模态检测、溯源与社交媒体场景

3.1 多模态融合检测方法

实际传播中遇到的伪造手法往往不止涉及一种模态——视频画面可能是合成的，配音可能是克隆的，字幕文本可能是生成的。单模态检测器在这种情况下顾不过来，于是多模态联合分析逐渐成为研究热点。以音视频关系为例，真实视频中嘴型和声音之间存在高度时序耦合，而经过伪造处理后这种耦合关系很容易被破坏，哪怕破坏程度很轻微，检测器也有机会捕捉到。基于此，研究者提出声音—人脸同质性检测思路，通过刻画声纹特征与面部表征之间的跨模态匹配关系来识别异常样本 [9]；关于音频深度伪造的综述也进一步说明，声纹伪造、说话人模仿与跨模态错配已成为需要联合建模的重要风险来源 [38]。

从模型结构看，多模态检测已从简单特征拼接转向层级化协同建模。M2TR 作为多模态多尺度 Transformer 框架，尝试同时刻画局部纹理异常与全局语义关系，在 FaceForensics++ 上的帧级 AUC 达到 97.6%，且在跨操控类型评测中表现出优于单模态基线的稳定性 [14]。针对音频信息利用不足的问题，FRADE 则把音频线索以知识蒸馏方式注入视觉分支，使模型在社交网络压缩与传播场景下仍保持较好的鲁棒性 [7]。此外，多模态大语言模型开始被用于媒体取证任务，相关研究表明，这类模型在跨模态推理和解释生成方面具有一定潜力，但在精细定位与证据稳定性方面仍存在明显不足 [5]。DFBench 对通用多模态大模型的识别能力进行了系统基准评测，LEDNet 则进一步代表了面向鲁棒检测的多模态基础模型路线，两者共同提示通用模型尚未稳定超过专用检测器 [30,31]。与此同时，基于内容引导的空间—频率关系动态图学习方法 [8] 以及面向音视频时序伪造的嵌入级融合框架 [36] 也表明，多线索协同与时序约束仍是提升复杂样本识别能力的重要方向。

3.2 生成源溯源与归因技术

如果说检测解决的是“真还是假”的问题，那么溯源与归因追问的则是“假的话，是谁造的”。这类工作的前提假设并不复杂：每种生成模型在输出中多少都会留下自己的统计印记。Waseem 等 [11] 在综述中就梳理了这一点——GAN 常在频谱上留下周期性伪影，扩散模型则会残留特征性的去噪噪声模式，不同架构还可能引入各自独特的纹理偏差。这些痕迹都可以作为归因分析的切入点。因此，溯源任务并非仅判断内容是否被篡改，而是尝试进一步恢复其生成来源及可能的技术路径。

模型归因框架的关键，在于建立生成源与可观测特征之间的稳定映射。Khoo 等人围绕 AI 生成图像来源识别提出基于深度特征的归因方法，尝试利用不同生成器留下的风格指纹实现细粒度区分 [23]。不过，Sandotra 等 [13] 对特征型 AI 检测技术的系统评估也指出，这类方法一旦进入开放集场景，便会遭遇生成器未见、版本更新和后处理扰动等问题。从已有实验看，同时利用频域和语义两类特征确实有助于提升跨模型归因的准确度；只是一旦遇到重压缩、重采样或者平台端的二次编辑，指纹信号就会受到不同程度的破坏，溯源精度随之打折。

3.3 社交媒体场景检测

到了社交媒体平台上，检测任务面对的输入条件就完全不一样了。图片经过平台压缩，视频被反复转发和跨端转载，有时还会被二次剪辑或拼接——这些操作叠加起来，会把原始取证线索破坏得所剩无几，很多在标准数据集上表现不错的方法一放到真实传播链里就出现明显衰减。针对这一问题，SIDA 框架引入大型多模态模型，同时执行检测、定位与解释任务，并输出自然语言说明，以增强结果的可理解性和可采信度 [4]。对于平台治理而言，这类“检测结论 + 证据解释”的输出形式往往比单一概率分值更具实用价值。

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

另一方面，社交媒体中的失真信息往往并非由单一伪造图像构成，而是通过图像、标题、评论和标签的联动制造误导。因而，文本—图像联合篡改检测逐渐成为重要方向。有害模因识别研究[2]说明，仅理解图像内容或文字含义本身并不足够，模型还需要把握两者组合后的情境指向与传播意图。与此同时，基于关键帧提取的社交媒体视频检测策略[21]为平台侧实时审核提供了更具可行性的工程路径，显示出在效率与准确性之间进行任务分层的必要性。

3.4 基准数据集与评测体系

数据集建设仍是该领域方法迭代的重要前提。FaceForensics++[28]、Celeb-DF[24]与DFDC[29]构成目前使用最广的三类基准，它们覆盖换脸、表情重演与视频伪造等主要任务，但总体上仍偏向受控环境采集。相比之下，WildDeepfake[25]更接近社交媒体中的真实传播条件，因此检测难度更高，也更能检验模型的域外适应能力。DFFMD[10]围绕口罩遮挡场景构建专项数据集，则提示我们：随着社会场景变化，伪造检测也需要持续补充新的评测样本类型。

评测方面一个值得关注的发现来自Groh等[20]：他们对比了普通人群、机器系统以及“机器辅助人群”三种判别模式后发现，人类对深度伪造的直觉判断其实相当不稳定——但如果给受试者提供机器端的提示信息，群体识别效果会有明显改善。换句话说，只拿模型在标准测试集上的数字来评判是不够的，人机交互环节的表现同样应该纳入考量。此外，The Face Deepfake Detection Challenge[16]这类公开挑战赛确实在推动方法比较上起到了积极作用，统一的排行榜和匿名测试集为研究者提供了相对公平的竞技场。但话说回来，目前的评测协议对跨数据集迁移能力、抗对抗攻击能力以及多模态场景的覆盖仍然不够，距离反映真实平台的部署需求还有不小的差距。

4 对比分析、挑战与展望

4.1 主要发现总结

回过头来看，这个领域的发展轨迹并不是一条平滑的上升曲线，倒更像是检测方与生成方之间不断你来我往的攻防拉锯。伴随生成模型持续升级、内容传播路径日趋复杂，加之平台治理标准水涨船高，评价一个检测系统好不好的维度也在变——光看准确率已经远远不够，泛化能力、可解释性和实际可部署性同样被摆上了桌面。

从方法脉络来看，最早一批DeepFake检测工作主要抓的是肉眼或浅层统计能发现的线索——纹理不自然、频域里的异常峰值、帧间时序的跳变等等。LBPNet[17]和双分支循环网络[26]就是这一阶段的典型代表。这类方法在受控数据集上表现尚可，但一旦跨域评测，性能下降十分明显[15]。随后，面向真实人脸基础表示学习的方案[6]开始强调对“真实分布”的建模，以减轻对特定生成器伪影的依赖；再往后，多模态检测逐渐成为重要方向，FRADE[7]利用音频线索增强视频判别，M2TR[14]则通过多尺度Transformer捕捉跨模态与跨层级特征。与此同时，研究目标也从单纯的真伪判别扩展到生成源归因，尝试借助生成器指纹实现更细粒度的取证分析[23]。

综合比较现有代表性方法可以发现，尽管它们采用的检测线索和模型结构存在差异，但都难以摆脱数据分布依赖这一共同约束：关于深度伪造媒体取证现状与未来挑战的综述已明确指出，生成工具一旦更新，既有检测器便可能部分失效[3]；而围绕多模态大语言模型用于媒体取证的研究则表明，多模态融合虽然通常优于单模态方案，但其收益仍受到模态质量、同步程度与平台后处理方式的显著影响[5]。这也是为什么实验室结果往往难以直接迁移至真实平台场景。与此同时，关于“人工幻象时代”检测偏差的讨论也提醒我们，评价体系本身也可能受到认知偏置影响[37]。

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

方法名称	年份	核心技术	优势	局限性
Multi-attentional Deepfake Det	2020	多注意力网络同时关注眼周、口周、轮廓边界等多个局部区域，降低	能够提升细粒度异常的覆盖度，在FaceForensics++（HQ）基准上帧级AUC达到99.29%	对训练数据标注质量和场景分布较为敏感，在跨数据集评测（如Celeb-DF）中性能下降较为明显，反映出
Face-Context Inconsistency Det	2019	将人脸区域与背景区域联合建模，检验二者在纹理、光照、边界和空	输出结果比较容易解释，复杂度相对可控；换脸拼接难以把面部区域与原始背景在光照纹理上做到完美衔接，联合	文章未明确给出该方法的具体跨数据集量化指标，对生成质量极高的全脸合成类伪造检测效果有限
M2TR（多模态多尺度Transformer）	2022	同时刻画局部纹理异常与全局语义关系的多模态多尺度Transf	在FaceForensics++上帧级AUC达到97.6%，在跨操控类型评测中表现出优于单模态基线的	文章指出通用多模态大模型尚未稳定超过专用检测器，在精细定位与证据稳定性方面仍存在明显不足
FRADE（音频知识蒸馏增强视觉检测）	2023	把音频线索以知识蒸馏方式注入视觉分支，实现音视频联合建模	使模型在社交网络压缩与传播场景下仍保持较好的鲁棒性，针对音频信息利用不足问题提供有效补充	文章未明确给出具体量化性能指标，依赖音视频同步信号，对纯视觉篡改场景覆盖有限
SIDA（大型多模态模型检测-定位-解释框架）	2024	引入大型多模态模型，同时执行检测、定位与解释任务，并输出自然	增强结果的可理解性和可采信度，"检测结论+证据解释"的输出形式比单一概率分值对平台治理更具实用价值	文章指出多模态大语言模型在精细定位与证据稳定性方面仍存在明显不足，对平台生成后的取证线索破坏缺乏专项

图 3 代表性检测方法多维度对比

4.2 核心瓶颈

尽管该领域已形成较完整的方法谱系，但若从真实部署与治理协同角度审视，仍存在几项尚未被根本解决的结构性约束。

首先，泛化性仍是最核心的技术瓶颈。现有模型面对已知伪造方式时往往表现良好，但遇到新型生成架构、未知后处理或跨平台再编码条件时，准确率常出现明显下滑 [3]。Malik 等 [15] 对跨数据集评估的梳理说明，这类衰减并非个别模型问题，而是当前检测范式普遍存在的域依赖现象。基础表示学习 [6] 虽然提供了更稳健的起点，但距离真正的开放世界检测仍有差距；有关检测挑战与“impostor bias”的最新讨论同样表明，模型面对新型人工生成幻象时仍容易产生系统性误判 [37]。

其次，实时性与算力成本直接限制了工程落地。M2TR 等多模态多尺度 Transformer 模型 [14] 与 Multi-attentional Deepfake Detection 这类多注意力网络 [18] 在离线实验中效果突出，但平台审核往往要求在压缩视频、海量并发和延迟受限条件下完成推理，这使高复杂度模型难以直接用于实时流式场景；SIDA 面向社交媒体检测、定位与解释一体化的探索也从侧面反映了这一工程压力 [4]。目前还没有一套公认的工程方案能在模型轻量化、推理吞吐量和检测精度三者之间做好平衡，这仍然是个有待突破的问题。

再次，数据供给与治理协同问题同样不可忽视。高质量伪造样本的构建、标注和合规使用成本较高，涉及隐私、肖像权与传播伦理等多重约束；无论是 SIDA 这类面向平台解释输出的框架 [4]，还是基于多模态大模型的媒体取证探索 [5]，都提示检测结果必须同时兼顾可理解性与可审计性。与此同时，生成源归因研究 [23] 也说明，如果技术研发与证据标准、申诉机制和责任认定之间缺乏联动，即使检测模型本身取得提升，也难以转化为稳定的社会治理能力。

4.3 未来研究展望

面向后续研究，相关工作更适合沿着“开放世界适应—多模态归因—轻量化部署”三条主线继续推进。

其一，开放世界泛化检测仍应作为优先议题。基础表示学习 [6] 已经显示出通过建模真实内容分布提升域外稳定性的潜力，但关于深度伪造媒体取证未来挑战的综述也表明，如何在无需频繁重训的条件下适应新型生成器，仍需结合持续学习、元学习或测试时自适应等机制进一步探索 [3]。

其二，多模态溯源与可解释性融合值得进一步加强。多模态大模型在媒体取证中的初步应用 [5] 说明，语义层推理有望补足传统伪影检测在证据组织上的不足；若能把生成器指纹识别 [23]、音频蒸馏增强的多模态检测 [7]、音视频时序伪造分析 [36] 与音频深度伪造综述所总结的风险谱系 [38] 结合起来，检测结果将更接近可审计、可申诉和可复核的取证证据。

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

其三，轻量化实时检测架构决定了技术能否真正落地。面向社交媒体流式审核场景[4]，后续研究需要在知识蒸馏、模型压缩、关键帧筛选和边缘推理优化等方面形成系统方案，以缩小实验室性能与平台部署条件之间的差距。

References

[1] WANG Y, PAN Y, YAN M, et al. A survey on ChatGPT: AI-generated contents, challenges, and solutions[J]. IEEE Open Journal of the Computer Society, 2023, 4: 3-25.

[2] CHI Fuyuan, YAN Yu, LIU Zhongyi. Harmful meme recognition with multimodal large models integrating emotion perception and multi-step reasoning[J]. Data Analysis and Knowledge Discovery, 2025, 9(1): 1-19.

[3] AMERINI I, BARNI M, BATTIATO S, et al. Deepfake media forensics: Status and future challenges[J]. Journal of Imaging, 2025, 11(3): 73.

[4] HUANG Z, HU J, LI X, et al. SIDA: Social media image deepfake detection, localization and explanation with large multimodal model[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2025.

[5] JIA S, LYU R, ZHAO K, et al. Can ChatGPT detect deepfakes? A study of using multimodal large language models for media forensics[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE, 2024: 4324-4333.

[6] SHI L, ZHANG J, JI Z, et al. Real face foundation representation learning for generalized deepfake detection[J]. Pattern Recognition, 2025, 161: 111299.

[7] NIE F, NI J, ZHANG J, et al. FRADE: Forgery-aware audio-distilled multimodal learning for deepfake detection[C]//Proceedings of the 32nd ACM International Conference on Multimedia. New York: ACM, 2024: 6297-6306.

[8] WANG Y, WU Z, HUANG Y, et al. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023.

[9] CHENG H H, GUO Y, WANG T, et al. Voice-face homogeneity tells deepfake[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2024, 20(3): 76:1-76:22.

[10] ALNAIM N, ALMUTAIRI Z M, ALSUWAT M S, et al. DFFMD: A deepfake face mask dataset for infectious disease era with deepfake detection algorithms[J]. IEEE Access, 2023, 11: 14829-14843.

[11] WASEEM S, ABU-BAKAR S A R, AHMED B A, et al. DeepFake on face and expression swap: A review[J]. IEEE Access, 2023, 11: 117865-117906.

[12] SADIQ S, ALJREES T, ULLAH S. Deepfake detection on social media: Leveraging deep learning and FastText embeddings for identifying machine-generated tweets[J]. IEEE Access, 2023, 11: 95008-95021.

[13] SANDOTRA N, ARORA B. A comprehensive evaluation of feature-based AI techniques for deepfake detection[J]. Neural Computing and Applications, 2024, 36(8): 3859-3887.

[14] HE Y, LIU K, ZHOU W, et al. M2TR: Multi-modal multi-scale transformers for deepfake detection[C]//Proceedings of the 2022 International Conference on Multimedia Retrieval. New York: ACM, 2022: 615-623.

[15] MALIK A, KURIBAYASHI M, ABDULLAHI S M, et al. DeepFake detection for human face images and videos: A survey[J]. IEEE Access, 2022, 10: 18757-18775.

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

[16] GUARNERA L, GIUDICE O, GUARNERA F, et al. The face deepfake detection challenge[J]. Journal of Imaging, 2022, 8(10): 263.
[17] KINGRA S, AGGARWAL N, KAUR N. LBPNet: Exploiting texture descriptor for deepfake detection[J]. Forensic Science International: Digital Investigation, 2022, 44: 301452.
[18] ZHAO H, ZHOU W, CHEN D, et al. Multi-attentional deepfake detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 2185-2194.
[19] NIRKIN Y, WOLF L, KELLER Y, et al. DeepFake detection based on discrepancies between faces and their context[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 6111-6121.
[20] GROH M, EPSTEIN Z, FIRESTONE C, et al. Deepfake detection by human crowds, machines, and machine-informed crowds[J]. Proceedings of the National Academy of Sciences of the United States of America, 2022, 119(1): e2110013119.
[21] MITRA A, MOHANTY S P, CORCORAN P, et al. A machine learning based approach for deepfake detection in social media through key video frame extraction[J]. SN Computer Science, 2021, 2(2): 98.
[22] DAS S, SEFERBEKOV S, DATTA A, et al. Towards solving the DeepFake problem: An analysis on improving DeepFake detection using dynamic face augmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Montreal: IEEE, 2021.
[23] KHOO B, PHAN R C, LIM C H. Deepfake attribution: On the source identification of artificially generated images[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2021, 12(3): e1458.
[24] LI Y, YANG X, SUN P, et al. Celeb-DF: A large-scale challenging dataset for DeepFake forensics[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 3204-3213.
[25] ZI B, CHANG M, CHEN J, et al. WildDeepfake: A challenging real-world dataset for deepfake detection[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 2382-2390.
[26] MASI I, KILLEKAR A, MASCARENHAS R, et al. Two-branch recurrent network for isolating deepfakes in videos[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer, 2020: 667-684.
[27] PAGE M J, MCKENZIE J E, BOSSUYT P M, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews[J]. BMJ, 2021, 372: n71.
[28] RÖSSLER A, COZZOLINO D, VERDOLIVA L, et al. FaceForensics++: Learning to detect manipulated facial images[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 1-11.
[29] DOLHANSKY B, HOWES R, PFLAUM B, et al. The DeepFake Detection Challenge (DFDC) preview dataset[EB/OL]. (2019-10-23)[2026-03-11]. https://arxiv.org/abs/1910.08854.
[30] WANG J, DUAN H, WANG J, et al. DFBench: Benchmarking deepfake image detection capability of large multimodal models[EB/OL]. (2025-06-03)[2026-03-11]. https://arxiv.org/abs/2506.03007.
[31] TAO R, TANG S, QIN H, et al. LEDNet: A multimodal foundation model for robust deepfake detection[J]. Science China Information Sciences, 2025, 68(6): 160106.
[32] CHEN H, LIN Y, LI B, et al. Learning features of intra-consistency and inter-diversity: Keys toward generalizable deepfake detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023,

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

33(3): 1468-1480.
[33] SHI Z, CHEN H, JIA Y, et al. Customized transformer adapter with frequency masking for deepfake detection[J]. IEEE Transactions on Information Forensics and Security, 2025, 20: 5904-5918.
[34] KHALID F, JAVED A, MALIK K M, et al. ExplaNET: A descriptive framework for detecting deepfakes with interpretable prototypes[J]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2024, 6(4): 486-497.
[35] AHMED N U R, BADSHAH A, ADEEL H, et al. Visual deepfake detection: Review of techniques, tools, limitations, and future prospects[J]. IEEE Access, 2025, 13: 1923-1961.
[36] LIU M, WANG J, QIAN X, et al. Audio-visual temporal forgery detection using embedding-level fusion and multi-dimensional contrastive loss[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(8): 6937-6948.
[37] CASU M, GUARNERA L, CAPONNETTO P, et al. GenAI mirage: The impostor bias and the deepfake detection challenge in the era of artificial illusions[J]. Forensic Science International: Digital Investigation, 2024, 50: 301795.
[38] KHANJANI Z, WATSON G, JANEJA V P. Audio deepfakes: A survey[J]. Frontiers in Big Data, 2023, 5: 1001063.

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)