TC260-PG-202512A

网络安全标准实践指南

——人工智能生成合成内容检测 第 1 部分：框架

(V1.0-202508)

全国网络安全标准化技术委员会
National Technical Committee 260 on Cybersecurity of SAC

全国网络安全标准化技术委员会秘书处

2025 年 08 月

本文档可从以下网址获得：
www.tc260.org.cn/

全国网络安全标准化技术委员会
National Technical Committee 260 on Cybersecurity of SAC

前　言

《网络安全标准实践指南》（以下简称《实践指南》）是全国网络安全标准化技术委员会（以下简称“网安标委”）秘书处组织制定和发布的标准相关技术文件，旨在围绕网络安全法律法规政策、标准、网络安全热点和事件等主题，宣传网络安全相关标准及知识，提供标准化实践指引。

本文件是《网络安全标准实践指南——人工智能生成合成内容检测》的第 1 部分。计划发布以下部分：

——第 2 部分：评价方法；

——第 3 部分：图片检测指南；

——第 4 部分：视频检测指南；

——第 5 部分：音频检测指南；

——第 6 部分：文本检测指南。

本文件起草单位：阿里云计算有限公司、中国电子技术标准化研究院、国家计算机网络应急技术处理协调中心、浙江大学、阿里巴巴（中国）有限公司、北京快手科技有限公司、香港中文大学（深圳）、中央网信办数据与技术保障中心、国家计算机网络应急技术处理协调中心江苏分中心、杭州中科睿鉴科技有限公司、北京中科凡语科技有限公司、厦门美图网科技有限公司、OPPO 广东移动通信有限公司、北京抖音信息服务有限公司。

本文件主要起草人：孙勇、曾吉申、张妍婷、任奎、许

晓耕、张震、王志伟、杨锐、刘佳睿、郝春亮、贺敏、陈保营、张立尧、黄天宁、落红卫、谷晨、武执政、王力、周鹏、何覃、邓彪、王泽晋、涂先胜、嵇程、刘梓含、郭建领、熊安、巴钟杰、孙培尧、吕飞霄、黄晴、宋昊、费凡芮。

全国网络安全标准化技术委员会
National Technical Committee 260 on Cybersecurity of SAC

声　明

本《实践指南》版权属于网安标委秘书处，未经秘书处书面授权，不得以任何方式抄袭、翻译《实践指南》的任何部分。凡转载或引用本《实践指南》的观点、数据，请注明“来源：全国网络安全标准化技术委员会秘书处”。

摘　要

贯彻落实《人工智能生成合成内容标识办法》第六条关于网络信息内容传播服务的服务提供者通过检测生成合成痕迹来识别疑似生成合成内容的要求，规范人工智能生成合成内容检测的开发与应用，本文件给出了人工智能生成合成内容检测框架，明确了人工智能生成合成内容检测的目标设定、检测流程、检测算法、检测服务封装等方面内容，并给出了常见检测算法及检测服务封装，为人工智能生成合成内容检测的设计、开发和应用提供标准化指导。

全国网络安全标准化技术委员会
National Technical Committee 260 on Cybersecurity of SAC

目　录

1 范围 2
2 规范性引用文件 2
3 术语和定义 2
4 缩略语 3
5 概述 3
6 目标设定 4
7 检测流程 5
7.1 概述 5
7.2 检测方案设计 6
7.3 检测对象预处理 7
7.4 检测实施 9
7.5 检测结果输出 11
8 检测算法 13
8.1 文本 13
8.2 图片 14
8.3 音频 15
8.4 视频 16
附录 A（资料性）常见检测算法 18
附录 B（规范性）检测服务封装 31

全国网络安全标准化技术委员会 National Technical Committee 260 on Cybersecurity of SAC

1 范围

本文件给出了人工智能生成合成内容检测框架，明确了人工智能生成合成内容检测的目标设定、检测流程、检测算法、检测服务封装等方面内容。

本文件适用于网络信息内容传播服务提供者对文本、图片、音频、视频等内容进行生成合成痕迹检测，也可为第三方测评机构提供参考。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 25069—2022 信息安全技术 术语

GB 45438—2025 网络安全技术 人工智能生成合成内容标识方法

3 术语和定义

GB/T 25069—2022、GB 45438—2025 界定的以及下列术语和定义适用于本文件。

3.1 人工智能生成合成内容 content generated by artificial intelligence

利用人工智能技术生成、合成的文本、图片、音频、视频、虚拟场景等信息。

[来源：GB 45438—2025，3.1]

3.2 人工智能生成合成内容检测 detection of artificial intelligence generated content

使用算法对人工智能生成合成内容（3.1）与非人工智能生成合成内容进行识别和区分的过程。

4 缩略语

下列缩略语适用于本文件。

AI：人工智能（Artificial Intelligence）

API：可通过开发应用程序编程接口（Application Programming Interface）

SaaS：软件即服务（Software as a Service）

SDK：软件开发工具包（Software Development Kit）

5 概述

人工智能生成合成内容检测涉及检测对象、目标设定、检测流程、检测算法等内容。本实践指南中检测对象包括文本、图片、音频、视

频。目标设定包括一般目标、特定目标。检测流程包括检测方案设计、检测对象预处理、检测实施、检测结果输出。检测算法包括文本检测算法、图片检测算法、音频检测算法、视频检测算法。人工智能生成合成内容检测框架如图 1 所示。常见检测算法见附录 A，检测服务封装见附录 B。

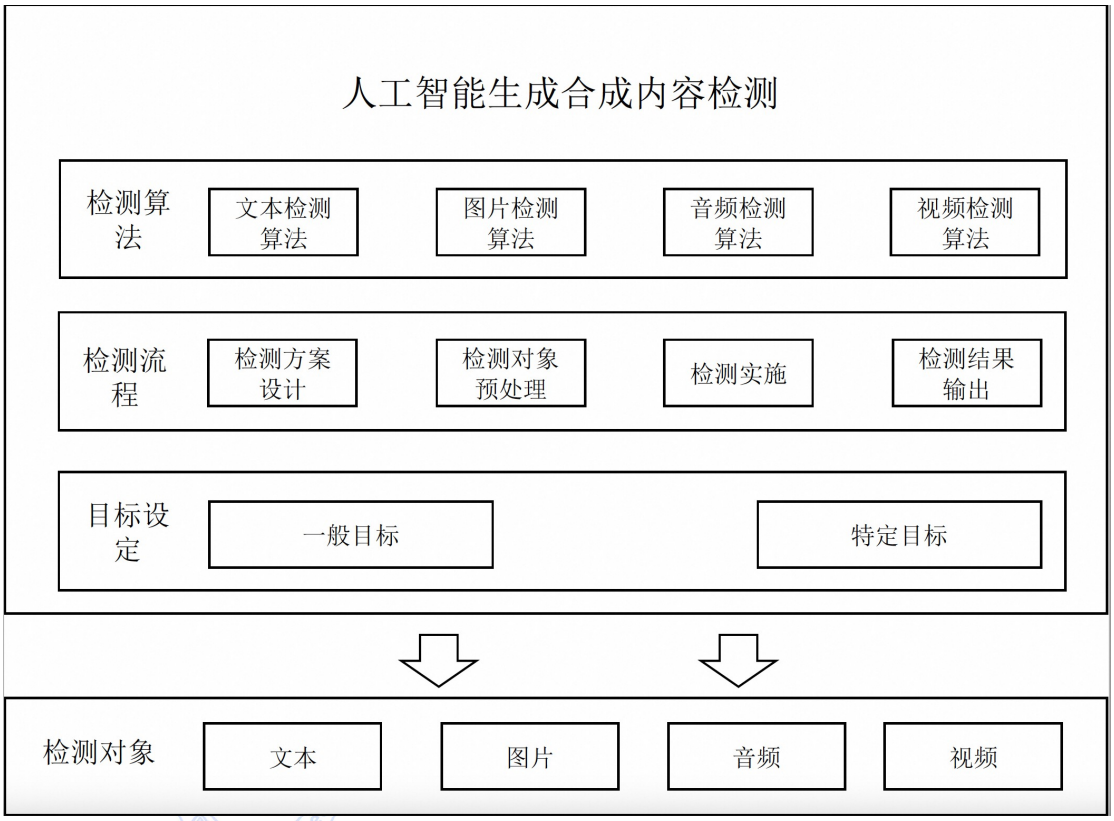


图 1　人工智能生成合成内容检测框架

6　目标设定

检测目标设定包括：一般目标、特定目标。一般目标为对常见内容进行检测，在检测对象遭受裁剪、形变、压缩、编辑、社交媒体传输等失真干扰情形下，检测指标（准确率、召回率、误检率等）需满足业务预期目的；特定目标为针对明确应用场景，检测需支持的个性

化需求。

a) 一般目标:

1) 保证检测准确：检测能够准确地区分人工智能生成合成内容和非人工智能生成合成内容，其准确率、召回率、误检率等符合预期;

2) 检测对象适配：检测能够适用于常见的文本、图片、音频、视频等格式;

3) 抵御失真干扰：检测在检测对象遭受裁剪、形变、压缩、滤波、去噪、编辑、社交媒体传输等情形下，仍然能够检测出符合预期的人工智能生成合成内容。

b) 特定目标:

1) 支持局部检测：在有明确局部内容检测要求的应用场景中，如图片、视频中的人脸检测，检测需满足相应的要求;

2) 支持实时应用：在有明确检测时效要求的应用场景中，检测需满足对应的实时性要求;

3) 生成算法判定：检测能够判定出人工智能生成合成文本、图片、音频、视频所使用的算法类别，如：生成对抗网络、扩散模型等。

7 检测流程

7.1 概述

人工智能生成合成内容检测流程旨在通过一系列步骤来识别和验证检测对象内容是否由人工智能技术生成合成。这一流程包括检测方案设计、检测对象预处理、检测实施以及检测结果输出等关键环节。人工智能生成合成内容检测流程如图 2 所示。

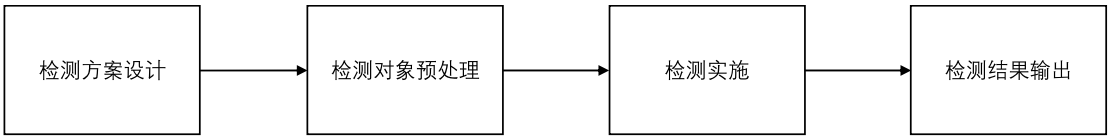


图 2 检测流程

7.2 检测方案设计

检测方案设计环节是制定人工智能生成合成内容检测策略的基础。在此环节，首先需要明确检测使用场景，确定检测对象格式、准确率需求等。其次评估对应场景下潜在的失真干扰及具体需求，以便确定一般目标和特定目标。最后根据具体目标设计出检测方案。

对检测方案进行设计时，需要考虑的因素包括但不限于：

a) 使用场景：确定检测应用于哪些具体场景，如：新闻验证、司法取证等，同时分析场景中的可能遇到的挑战，如：大规模数据处理、多样性的媒体内容等；

b) 检测对象的基本信息：明确检测对象的类型（文本、图片、音频、视频等）及其特定的格式和技术参数。如：图片的格式、分辨率、位深等；

c) 准确性需求：设定检测需要达到的准确性要求，包括准确率和召回率，并考虑在不同误报率下的检测性能；

d) 生成算法判定需求：确定是否需要识别用于生成合成内容的具体人工智能技术或算法，开发或采用能够分析和识别生成技术特征的检测手段；

e) 抵御失真干扰需求：评估检测对象可能经历的各种失真干扰，如压缩、噪声、格式转换等，设计检测算法以提高对这些干扰的鲁棒性；

f) 实时性需求：如果场景要求实时检测，确保检测能够快速响应并及时更新检测结果；

g) 局部检测需求：根据需求，决定检测是否需要识别局部修改或生成的检测对象，选择或调整检测算法以识别和定位局部异常；

h) 检测对象的标识信息：考虑是否需要提取或识别检测对象中可能存在的标识信息，这些信息可能包括内容显式标识、元数据标识、隐式内容标识等。

7.3 检测对象预处理

在此环节，根据上述的检测方案对待检测内容进行必要的初步处理，以便提取特征或调用模型。针对文本、图片、音频、视频等不同类型的内容，预处理考虑因素会有所不同，但都要确保数据的质量和一致性。

a) 对文本进行预处理，需要考虑的因素包括但不限于：

1) 格式转换：将文本文件转换为适合处理的格式，例如将其转换为纯文本格式或将特殊字符转义；

2) 分词：将文本拆分为词语或短语，以便进行进一步的处理和分析；

3) 清洗：去除文本中的无用信息，如 HTML 标签、无效 URL 链接、停用词等；

4) 矫正拼写错误：使用自然语言处理技术纠正文本中的拼写错误，以提高后续处理的准确性；

5) 特征提取：从文本中提取关键特征，如标点符号、锻炼数量、主谓宾个数、主题、命名实体等，以为后续任务提供必要的输入数据。

b) 对图片进行预处理，需要考虑的因素包括但不限于：

1) 格式转换：将图片文件转换为检测算法所需的格式；

2) 几何变换：通过缩放、旋转、平移、剪切等操作校正图片的几何畸变；

3) 分辨率调整：根据检测算法的要求，调整图片的大小和分辨率，以优化计算效率和检测效果；

4) 去噪：使用滤波器等技术减少图片中的噪声，提高后续特征提取的准确性；

5) 特征提取：提取图片的关键特征，如边缘、纹理、颜色分布等，为检测算法提供必要的输入数据。

c) 对视频进行预处理，需要考虑的因素包括但不限于：

1) 视频分离：将视频分解为单独的帧或关键片段，以便对每

一帧或片段进行独立分析；

2) 帧率调整：根据检测需求，可能需要调整视频的帧率，以减少处理的数据量或提高处理速度；

3) 分辨率和编码格式调整：调整视频的分辨率和编码格式，以适应检测算法的性能要求；

4) 音频分离：如果视频中包含音频，可能需要将其从视频中分离出来，单独进行音频预处理；

5) 特征提取：提取视频的关键特征，如纹理、颜色分布、频谱特征、音高、节奏等信息，也可使用计算机视觉技术来识别视频中的物体、人物和活动，提取相关的特征向量。

d) 对音频进行预处理，需要考虑的因素包括但不限于：

1) 格式转换：将音频文件转换为检测算法可处理的格式；

2) 采样率调整：根据检测算法的要求，调整音频的采样率，以优化处理效率和检测性能；

3) 去噪：使用噪声抑制技术清除背景噪声，提高音频内容的清晰度；

4) 特征提取：提取音频的关键特征，如频谱、梅尔频率倒谱系数、节奏和音调等，为检测算法提供输入。

7.4 检测实施

检测实施环节是将设计好的检测方案应用于预处理后的内容上，执行具体的检测任务。此环节的关键在于选用合适的策略与算法来分

析检测对象内容，以识别是否由人工智能技术生成。

注：某些检测方式（如利用多模态大语言模型技术）可能已经将检测对象预处理环节包含在模型内部。这些方式对人工智能生成合成内容进行检测设计和训练优化，直接处理上传的内容文件，并根据输入的检测要求给出检测结果。

a) 文本、图片、音频、视频等内容检测时，需要考虑的策略包括但不限于：

1) 全局检测：选择完整内容进行检测实施，适用于内容整体上是否有 AI 生成判定；

2) 局部检测：选择局部内容（如图片或视频中的人脸）进行检测实施，适用于部分内容的判定；

3) 单模态检测：针对视频检测时，可分解为视频帧、音频，参考图片检测算法、音频检测算法，分别对视频帧、音频进行检测；

4) 多模态融合检测：对于包括图片和音频的视频，采用多模态融合检测可以提高检测的准确性。如对深度伪造视频的检测时，单纯分析图片或音频可能难以发现所有的操纵痕迹，当结合两者进行分析时，可能会发现人物的嘴型与语音不同步，或者视频中的动作与声音发出的时间点不匹配等问题；

5) 多算法交叉检测：采用多种不同的算法对同一内容进行独立分析，随后通过交叉验证的方式来整合这些算法的检测结果。

b) 文本、图片、音频、视频等内容的检测算法选择参考第 8 章。

7.5 检测结果输出

检测结果输出环节是将检测实施环节得到的分析结果整理并呈现，以便使用者可以根据这些结果做出决策。对于一般目标，常见的输出形式是给出判断的结论，例如是人工智能生成，或不是人工智能生成，也可以进一步给出置信度。此外，根据检测方案，可进一步给出更多的判断信息，如由具体生成算法生成，或脸部由人工智能生成等。

对于检测结果的输出，需要考虑的因素包括但不限于:

a) 可视化展示: 可视化展示是帮助使用者快速理解检测结论的有效手段。在检测到人工智能生成内容时，可利用图片标注、热力图、高亮框等技术，直观呈现异常区域。例如，在图片检测场景中，若发现图片部分区域由人工智能生成，可使用不同颜色的边框对该区域进行标记，并在旁边标注生成类型；在文本检测中，对于疑似人工智能生成的段落，可通过不同底色高亮显示；在音频检测中，可采用热力图高亮伪造区域，帮助用户直观识别音频中的异常片段;

b) 置信度评分: 增强检测结果的可靠性，建议为每个判断提供置信度评分。置信度评分应基于检测的算法原理和数据特征，采用科学的计算方法得出。例如，可通过机器学习模型输出的概率值作为基础，结合历史检测数据的验证结果进行校准。同时，为便于用户理解，需对不同置信度区间进行明确说明，如高、

中、低等不同程度的确信，并在检测结果展示中，将置信度评分与判断结论一同呈现；

c) 结果验证：为增强检测结果的准确性和可靠性，需建立系统化验证机制。通过引入专家复审环节，基于领域专家的经验和专业知识对检测结论进行人工评估，判断其合理性；同时采用交叉检验的方法，运用多种检测算法对同一对象进行独立检测，对比多个结果之间的一致性。若出现不一致的情况，需进一步分析原因，重新进行检测或优化检测方案；

d) 报告编制：详细的检测报告是检测结果输出的重要组成部分。报告内容包括检测背景、检测方法、检测结果的详细说明、异常或可疑的标注及分析、采用的人工智能生成合成等。报告格式宜规范统一，采用模块化设计，方便用户快速定位所需信息。对于复杂的检测项目，可在报告中增加图表、案例分析等内容，增强报告的可读性；

e) 反馈机制：建立有效的反馈机制是持续改进检测方案和提升检测性能的关键。可通过在线表单、邮件反馈、用户评价系统等多种渠道，收集使用者对检测结果的评价和建议。对收集到的反馈信息进行分类整理和分析，及时发现检测过程中存在的问题，如误判、漏判等情况，并针对性地优化检测算法和参数；

f) 结果的可操作性：为满足其他系统或工具对检测结果进一步处理或分析的需求，需提供结果的机器可读版本。可通过开发

API，实现检测结果的实时传输和调用；也可将检测结果以标准化的数据格式（如 JSON、XML）输出，方便与其他系统进行集成。

8 检测算法

8.1 文本

人工智能生成合成文本检测算法旨在通过 7.3.a)提取文本关键特征或通过端到端文本分类器来区分人工智能生成合成文本与非人工智能生成合成文本。根据不同的目标需求，可采用多种算法进行检测。

a) 针对文本检测一般目标：

1) 确保检测准确：检测算法宜选择基于字符级别的统计模型分析算法、特定符号统计分析算法、端到端文章文本分类器算法、微调大模型检测器算法、基于对比的文本检测算法等；

2) 文本兼容适配：检测算法宜选择基于字符级别的统计模型分析算法、端到端文章文本分类器、微调大模型检测器算法、基于对比的文本检测算法等；

3) 抵御失真干扰，检测算法宜选择文本数据增强分类算法、基于对抗训练的文本分类方法等。

b) 在上述 8.1.a）基础之上，针对特定目标：

1) 支持局部检测，检测算法宜选择基于局部差异分析的方法、

基于字词级别的统计模型分析方法等;

2) 支持实时应用，检测算法宜选择轻量级检测算法，如基于轻量深度网络的文本分类方法、集成学习的文本分类方法等;

3) 生成算法判定，目前暂无有效检测算法。

8.2 图片

人工智能生成合成图片检测算法旨在通过7.3.b)中提取出的关键特征分析区分人工智能生成合成图片与非人工智能生成合成图片。根据不同的目标需求，可采用多种算法进行检测。

a) 针对图片检测一般目标:

1) 保证检测准确，检测算法宜选择基于物理特征的检测算法、空域特征检测算法、频域特征检测算法、基于生成过程的伪影特征检测算法、基于图片与文本描述多模态检测算法、生成图片重构残差检测算法等;

2) 图片兼容适配，检测算法宜选择基于数据增强检测算法、基于特定学习范式检测算法等;

3) 抵御失真干扰，检测算法宜选择鲁棒特征检测算法、数据增强检测算法。

b) 在上述8.2.a）基础之上，针对特定目标:

1) 支持局部检测，检测算法宜选择基于局部差异分析检测算法、图片局部区域定位算法;

2) 支持实时应用，检测算法宜选择轻量级检测算法；

3) 生成算法判定，检测算法宜选择生成载体指纹检测算法、生成载体特征空间聚类检测算法。

8.3 音频

人工智能生成合成音频检测算法旨在通过7.3.c)中提取出的关键特征分析区分人工智能生成合成音频与非人工智能生成合成音频 。根据不同的目标需求，可采用多种算法进行检测。

a) 针对音频检测一般目标:

1) 保证检测准确，检测算法宜选择时域统计特征检测算法、频域特征检测算法，以及结合时域和频域特征提取技术的深度学习检测算法；

2) 音频兼容适配，检测算法宜选择自适应预处理方法以处理不同格式的音频；

3) 抵御失真干扰，检测算法宜选择鲁棒特征检测算法、数据增强检测算法。

b) 在上述 8.3.a）基础之上，针对特定目标:

1) 支持局部检测，检测算法宜选择基于时域特征的生成音频定位算法或语音编辑区间定位算法；

2) 支持实时应用，检测算法宜选择轻量级检测算法；

3) 生成算法判定，检测算法宜选择生成载体指纹检测算法、针对特定生成伪影特征的语音深度伪造检测算法。

8.4 视频

人工智能生成合成视频检测算法旨在通过7.3.d）中提取出的关键特征分析区分人工智能生成合成视频与非人工智能生成合成视频。根据不同的目标需求，可采用多种算法进行检测。

a) 针对视频检测一般目标:

1) 保证检测准确，检测算法宜选择基于时序不一致检测算法、关键帧生成痕迹检测算法，基于物理特征的检测算法;

2) 视频兼容适配，检测算法宜选择自适应预处理算法、多尺度特征提取算法、基于特定学习范式检测算法等;

3) 抵御失真干扰，检测算法宜选择鲁棒视频生成检测算法、鲁棒特征检测算法、数据增强检测算法。

b) 在上述8.4.a）基础上，针对特定目标:

1) 支持局部检测，检测算法宜选择基于局部差异分析检测算法、基于帧级空间局部差异分析检测算法、面部关键部位伪造特征检测算法、基于时空特征联合的局部差异分析检测算法、基于多模态特征联合的检测算法、基于时序局部差异分析检测算法、片段定位检测算法以及基于帧级检测的检测算法;

2) 支持实时应用，检测算法宜选择轻量级检测算法;

3) 生成算法判定，检测算法宜选择生成载体指纹检测算法、时间序列生成指纹算法、视频帧生成指纹算法、多模态特

征联合视频指纹算法、生成载体特征空间聚类检测算法。

全国网络安全标准化技术委员会
National Technical Committee 260 on Cybersecurity of SAC

附录 A

（资料性）

常见检测算法

A.1　检测算法

检测算法是识别人工智能生成合成内容与非人工智能生成合成内容的关键技术。本节给出了各种检测对象的常见检测算法，可结合具体业务场景、实现目标选择合适的检测算法。

a) 常见的文本检测算法包括但不限于：

1) 基于字符级别的统计模型分析算法：该算法通过计算字符出现的频率及其组合方式，比较真实文本与生成文本之间的差异性，从而判断是否为人工智能生成合成文本。适用于一般目标下的生成文本检测场景；

2) 特定符号统计分析算法：该算法通过分析特定符号（如标点符号、特殊字符等）的使用情况及出现位置，判断是否为人工智能生成合成文本。适用于一般目标下的生成文本检测场景；

3) 端到端文章文本分类器算法：该算法通过对大量真实与虚假文章进行训练，构建一个端到端的文章文本分类模型。当新文本被输入时，系统会自动将其分类为真实或虚假。适用于一般目标下的生成文章文本检测场景；

4) 微调大模型检测器算法：该算法利用已经训练好的大型语言模型，对其进行微调以适应特定任务。当新的文本被输入时，系统会利用微调后的模型判断其真实性。适用于一般目标下的生成文本检测场景；

5) 轻量级检测算法：该算法通过设计轻量级的神经网络模型，如轻量化的注意力机制模型或模型量化推理等，来进行文本分类和检测。适用于特定目标下的生成文章文本检测场景。

6) 基于对比的文本检测算法：该算法通过利用大模型将输入文本进行润色或改写，对比改写前后两个文本的差异性大小，来判断文本内容是否由人工智能算法生成。这种方法适用于一般目标下的生成文本检测场景。

7) 文本数据增强分类算法：该算法通过对训练数据实施多种增强策略（如回译、同义替换、随机插入/删除、风格迁移、字符扰动等）生成更多语义或表述变体，以扩充训练集并在增强后数据上训练分类器或用于对比学习，从而提升检测器对润色、二次编辑和噪声变体的鲁棒性，适用于一般目标下的生成文本检测场景；

8) 基于对抗训练的文本分类方法：该算法在训练环节引入攻击器或生成器以构造针对检测器的强对抗样本（如词级替换、句法重排、回译攻击或训练型伪造器），并通过交替/

联合优化使检测器学习抵御这些伪装手段，从而增强在有能力主动规避的攻击者面前的稳健性，适用于一般目标下的高对抗与安全敏感检测场景。

b) 常见的图片检测算法包括但不限于：

1) 基于物理特征的检测算法：该算法通过分析图片中人物或物体的物理特征（如头部姿势、人眼特征、牙齿特征等），利用这些特征与真实世界物理规律的差异性，实现对生成图片的检测。主要针对图片的物理特征进行检测并分类，适用于一般目标下的生成图片检测场景；

2) 空域特征检测算法：该算法通过提取图片的空域特征或将图片空域信号输入端到端网络进行训练，利用生成图片与真实图片在这些特征上的差异，实现对生成图片的辨别。主要针对图片的色彩空域像素信号进行检测并分类，适用于一般目标下的生成图片检测场景；

3) 频域特征检测算法：算法通过分析图片的频域特征（如局部频域特征、全局频域特征、自适应频域特征等），分析生成图片在频率分布上的异常模式，实现对生成图片的检测。主要针对图片频域特征进行检测并分类，适用于一般目标下的生成图片检测场景。

4) 基于生成过程的伪影特征检测算法：该算法通过对生成图片中存在的特定伪影（这些伪影是由于生成过程中的限制

或算法内在缺陷所导致的，如融合边界、上采样伪影、残差特征等）的检测，实现对生成图片的识别。适用于一般目标下的生成图片检测场景；

5) 基于图片与文本描述多模态检测算法：该算法结合图片与相关文本描述，通过对比图片内容与文本描述的一致性，以及生成图片在跨模态信息融合方面的不自然性，实现对生成图片的检测。主要针对图片与关联文本描述的多模态信息进行检测并分类，适用于一般目标下的生成图片检测场景；

6) 生成图片重构残差检测算法：该算法利用扩散模型对图片进行重构，比较重构后图片与原图片的绝对误差（即重构残差），通过深度模型对这些残差特征进行分类，以此识别生成图片。主要针对重构残差特征进行检测并分类，适用于一般目标下的生成图片检测场景；

7) 基于数据增强检测算法：该算法利用数据增强技术（如基于人脸关键点、非敏感区域的数据增强）对图片进行处理，增强图片中可能被忽视或掩盖的生成痕迹，从而提高检测准确性。主要针对经过数据增强处理后的图片特征进行检测并分类，适用于一般目标下的生成图片检测场景；

8) 基于特定学习范式检测算法：该算法运用知识蒸馏、元学习、测试时自适应等特定学习范式，引导特征提取器提取

更能区分生成与非生成图片的特征，并利用分类器进行分类。主要针对特定学习范式引导下提取的特征进行检测并分类，适用于一般目标下的生成图片检测场景；

9) 生成载体指纹检测算法：该算法通过分析图片中由特定生成模型（如生成对抗网络、扩散模型）留下的独特指纹特征，实现对生成图片的追溯与识别。主要针对生成图片的指纹特征进行检测并分类，适用于一般目标下的生成图片检测场景；

10) 生成载体特征空间聚类检测算法：该算法首先利用特征提取器提取图片的表征，然后在特征空间中进行聚类，基于聚类结果判断图片是否为生成内容。主要针对图片在特征空间中的聚类结果进行检测并分类，适用于一般目标下的生成图片检测场景；

11) 鲁棒特征检测算法：该算法关注图片中不易受干扰、具有稳定性的特征（如几何学特征），或者通过端到端深度网络中数据增强来获得鲁棒的检测网络，通过分析这些特征或通过鲁棒端到端网络来识别生成图片。主要针对图片的鲁棒特征或通过将图片输入鲁棒端到端网络来进行检测并分类，适用于一般目标下的生成图片检测场景

12) 基于局部差异分析检测算法：该算法通过对图片局部区域的特征差异（如局部频域特征差异、局部空域特征差异）

进行细致分析，发现生成图片中可能存在的局部异常并实现生成图片检测。主要针对图片的局部差异特征进行检测并分类，适用于一般/特定目标下的生成图片检测场景；

13) 图片局部区域定位算法：该算法采用图片分割或目标检测技术，端到端学习定位图片中可能包含的生成区域（图片分割的输出为掩码，目标检测的输出为矩形框）。主要针对图片中局部生成区域的定位结果进行检测并分类，适用于一般/特定目标下的生成图片检测场景；

14) 轻量级检测算法：针对图片的特定实时应用需求，设计计算效率高、资源占用少的轻量级检测算法（如模型剪枝、量化等方式对端到端生成图片检测网络进行轻量化），实现对生成图片的快速识别。主要针对图片的特定轻量级特征进行检测并分类，适用于一般/特定目标下的生成图片检测场景；

c) 常见的音频检测算法包括但不限于：

1) 频域特征检测算法：算法通过分析音频的频域特征（如局部频域特征、全局频域特征、自适应频域特征等），揭示生成图片在频率分布上的异常模式，实现对生成图片的检测。主要针对图片频域特征进行检测并分类，适用于一般目标下的生成音频检测场景；

2) 生成载体指纹检测算法：该算法通过分析音频中由特定生

成模型（如生成对抗网络、扩散模型）留下的独特指纹特征，实现对生成音频的追溯与识别。主要针对生成音频的指纹特征进行检测并分类，适用于一般目标下的生成音频检测场景；

3) 鲁棒特征检测算法：该算法关注音频中不易受干扰、具有稳定性的特征（如声学特征、语言学特征），通过分析这些特征来识别生成音频。主要针对音频的鲁棒特征进行检测并分类，适用于一般目标下的生成音频检测场景；

4) 结合时域和频域特征提取技术的深度学习检测算法：该算法同时提取音频的时域特征和频域特征，并通过深度学习网络模型进行联合训练和检测，以识别生成音频。主要针对音频的时域与频域联合特征进行检测并分类，适用于一般/特定目标下的生成音频检测场景；

5) 时域统计特征检测算法：该算法通过计算音频信号中的时域统计特征（如均值、方差、峭度、峰值等），揭示生成音频在时域上的非自然模式，实现对生成音频的检测。主要针对音频时域统计特征进行检测并分类，适用于一般目标下的生成音频检测场景；

6) 针对特定伪影特征的检测算法：该算法专门针对生成音频中可能出现的特定伪影特征（如谐波失真、量化噪声、编码失真等）进行检测，这些特征通常是生成过程中的限制

或算法缺陷所导致的。主要针对音频特定伪影特征进行检测并分类，适用于一般/特定目标下的生成音频检测场景；

7) 时间序列生成指纹算法：该算法针对音频或语音生成内容，通过时间序列分析提取生成过程中的统一指纹，识别特定生成算法的使用。主要针对音频时间序列生成指纹进行检测并分类，适用于一般/特定目标下的生成音频检测场景；

8) 片段定位检测算法：该算法针对局部音频片段进行定位检测，通过分析音频片段内部的特征一致性、过渡自然性等，识别可能的生成区域。主要针对音频片段定位结果进行检测并分类，适用于一般/特定目标下的生成音频检测场景；

9) 轻量级检测算法：针对音频的特定实时应用需求，设计计算效率高、资源占用少的轻量级检测算法（如基于特征压缩、基于模型剪枝的检测算法），实现对生成音频的快速识别。主要针对音频的特定轻量级特征进行检测并分类，适用于一般/特定目标下的生成音频检测场景；

d) 常见的视频检测算法包括但不限于：

1) 基于物理特征的检测算法：该算法通过分析视频中人物或物体的物理特征（如头部姿势、人眼特征、牙齿特征等），利用这些特征与真实世界物理规律的差异性，实现对生成图片的检测。主要针对视频的物理特征进行检测并分类，适用于一般目标下的生成图片检测场景；

2) 频域特征检测算法：该算法通过对视频的频域特征（如局部频域特征、全局频域特征、自适应频域特征等）进行分析，揭示可能由 AI 生成技术导致的频域特异性异常，以此实现对生成视频的识别。适用于一般目标下的生成视频检测场景；

3) 基于特定学习范式检测算法：该算法运用特定学习范式（如知识蒸馏、元学习、测试时自适应学习等）指导特征提取和分类过程，以提升对生成视频的识别能力。适用于一般目标下的生成视频检测场景；

4) 生成载体指纹检测算法：该算法通过分析视频中可能存在的 AI 生成模型指纹特征（如生成对抗网络指纹、扩散模型指纹等），实现对生成视频的溯源和识别。适用于一般目标下的生成视频检测场景；

5) 生成载体特征空间聚类检测算法：该算法首先利用特征提取器提取视频的表征，然后基于特征空间进行聚类分析，根据聚类结果判断视频是否为 AI 生成。适用于一般目标下的生成视频检测场景；

6) 鲁棒特征检测算法：算法侧重于分析视频中的鲁棒特征（如几何学特征），通过这些不易受干扰的特征识别生成视频。适用于一般目标下的生成视频检测场景；

7) 基于局部差异分析检测算法：该算法通过对视频的局部特

征差异（如局部频域特征差异、局部空域特征差异）进行分析，实现对生成视频在局部区域的精准检测。适用于一般/特定目标生成视频检测场景；

8) 关键帧生成痕迹检测算法：该算法专门针对视频的关键帧进行分析，寻找由生成模型（如对抗模型、扩散模型）留下的特定痕迹（如噪声模式、重复纹理），以识别生成视频。适用于一般目标下的生成视频检测场景；

9) 多尺度特征提取算法：该算法通过结合多尺度的特征提取机制（如多层次卷积神经网络），获取视频的细粒度和粗粒度信息，以捕捉更全面地生成视频特征。适用于一般目标下的生成视频检测场景；

10) 基于时序不一致检测算法：该算法针对视频中的时序不一致特征（如空间时序不一致、局部全局时序不一致）进行检测，利用这些时序异常识别生成视频。适用于一般目标下的生成视频检测场景；

11) 时间序列生成指纹算法：该算法通过时间序列分析，从视频生成过程中提取具有辨识性的统一指纹，以识别特定生成算法的使用。适用于一般目标下的生成视频检测场景；

12) 视频帧生成指纹算法：该算法专门针对视频帧进行分析，设计检测算法提取由 AI 生成算法产生的独特指纹，用于识别生成视频。适用于一般目标下的生成视频检测场景；

13) 多模态特征联合视频指纹算法：该算法结合视频中的图片、音频、文本描述等多种模态特征，构建独特的多模态指纹，用于识别和追踪 AI 生成的视频内容。适用于一般目标下的生成视频检测场景；

14) 基于帧级空间局部差异分析检测算法：该算法通过分析单个视频帧内部的局部空间差异（如局部纹理不连贯、光照不一致），检测可能的人工生成内容。适用于一般/特定目标生成视频检测场景；

15) 面部关键部位伪造特征检测算法：该算法专注于识别视频中面部关键区域（如眼部、嘴部、鼻子）的伪造迹象（如不自然的表情动态、纹理细节），以鉴别生成视频。适用于特定目标下的生成视频检测场景；

16) 基于多模态特征联合的检测算法：算法通过联合视频的图片、音频、文本描述等多模态特征，提高对 AI 生成视频的检测准确率。适用于一般目标下的生成视频检测场景；

17) 基于时序局部差异分析检测算法：该算法分析视频帧与帧之间的时序局部差异，识别时间连续性中断或不自然过渡区域，从而判断视频是否为生成内容。适用于一般/特定目标生成视频检测场景；

18) 基于时空特征联合的局部差异分析检测算法：该算法利用时空信息联合分析，识别视频中局部区域的生成痕迹（如

运动与静态区域间的不自然过渡）。适用于一般/特定目标生成视频检测场景；

19) 轻量级检测算法：针对特定实时应用需求，设计轻量级检测算法（如基于补丁聚合、基于几何学的检测算法），在保持检测性能的同时，降低计算复杂度，适用于一般/特定目标需求下的生成视频检测场景。

A.2 检测算法与检测对象的适配情况

检测算法与检测对象的适配情况见表 A.1。

表 A.1 实现目标与检测算法的适配情况

算法分类	检测算法名称	一般目标	特定目标
文本检测算法	基于字符级别的统计模型分析算法	✓	×
	特定符号统计分析算法	✓	×
	端到端文章文本分类器算法	✓	×
	微调大模型检测器算法	✓	×
	轻量级检测算法	✓	✓
	基于对比的文本检测算法	×	×
	文本数据增强分类算法	✓	✓
	基于对抗训练的文本分类方法	✓	✓
图片检测算法	基于物理特征的检测算法	✓	×
	空域特征检测算法	✓	×
	频域特征检测算法	✓	×
	基于生成过程的伪影特征检测算法	✓	×
	基于图片与文本描述多模态检测算法	✓	×
	生成图片重构残差检测算法	✓	×
	基于数据增强检测算法	✓	×
	基于特定学习范式检测算法	✓	×
	生成载体指纹检测算法	✓	✓
	生成载体特征空间聚类检测算法	✓	✓
	鲁棒特征检测算法	✓	×
	基于局部差异分析检测算法	✓	✓

	图片局部区域定位算法	✓	✓
	轻量级检测算法	✓	✓
音频检测算法	频域特征检测算法	✓	×
	生成载体指纹检测算法	✓	✓
	鲁棒特征检测算法	✓	×
	结合时域和频域特征提取技术的深度学习检测算法	✓	✓
	时域统计特征检测算法	✓	×
	针对特定伪影特征的检测算法	✓	✓
	时间序列生成指纹算法	✓	✓
	片段定位检测算法	✓	✓
	轻量级检测算法	✓	✓
视频检测算法	基于物理特征的检测算法	✓	×
	频域特征检测算法	✓	×
	基于特定学习范式检测算法	✓	×
	生成载体指纹检测算法	✓	✓
	生成载体特征空间聚类检测算法	✓	✓
	鲁棒特征检测算法	✓	×
	基于局部差异分析检测算法	✓	✓
	关键帧生成痕迹检测算法	✓	×
	多尺度特征提取算法	✓	×
	基于时序不一致检测算法	✓	×
	时间序列生成指纹算法	✓	✓
	视频帧生成指纹算法	✓	✓
	多模态特征联合视频指纹算法	✓	✓
	基于帧级空间局部差异分析检测算法	✓	✓
	面部关键部位伪造特征检测算法	✓	✓
	基于多模态特征联合的检测算法	✓	✓
	基于时序局部差异分析检测算法	✓	✓
	基于时空特征联合的局部差异分析检测算法	✓	✓
	轻量级检测算法	✓	✓

注：
✓ 表示适配，× 表示不适配

附录 B

（规范性）

检测服务封装

B.1 SDK 封装

在本地化部署或应用程序集成的场景中，检测服务宜封装为 SDK 形式。SDK 可以提供灵活的接口，供开发者进行自定义集成和扩展。

注：SDK 封装是一种将检测服务集成到应用程序中的有效方式。通过 SDK 封装，开发者可以将检测服务嵌入到各种软件中。SDK 封装的优势在于高度的定制化和灵活度，开发者可以根据具体业务需求，实现定制化的解决方案。但是，SDK 封装也要求开发者具备一定的技术能力，以便正确地集成和使用 SDK。

B.2 SaaS 封装

在云服务或需远程访问的应用场景下，检测服务宜封装为 SaaS 形式。通过网络接口提供服务，用户可以通过 API 调用进行内容检测，无需关心底层技术实现和算法细节。

注：SaaS 封装是将检测服务作为在线提供的模式。用户通过网络访问服务提供者的服务器，利用 API 等接口形式进行内容检测。SaaS 封装的优势在于无需部署和维护，只需要网络请求即可获得检测服务。然后，SaaS 封装也存在一定的局限性，例如在网络带宽要求较高。

B.3 产品封装

在面向终端用户的场合，检测服务宜封装为产品形式。这些产品可以通过用户友好的图形界面来操作，提供即开即用的功能，适合非技术用户。

注：产品封装是将检测服务封装为独立的产品，通常包含图形化的用户界面，直接面向最终用户。产品封装的优势在于便捷性和易用性，使得非技术用户也能轻松地进行检测。但是，产品封装可能无法满足一些特定的场景，因为其功能和定制化能力有限。