广播电视和网络视听深度伪造防范技术要求

（2022 版）

广播电视人工智能应用国家广播电视总局重点实验室

2023 年 4 月

目 录

前　言

人工智能合成技术在媒体领域具有广泛的应用前景，但是深度伪造技术的滥用会危及国家安全、政治安全、经济安全和社会安全。深度伪造是指以误导他人为目的，采用深度学习等手段篡改、伪造视听内容的一类技术，通常涉及面部替换、表情操纵、人脸合成、语音剪辑、语音合成、场景编辑、场景合成等。

本技术要求按照 2019 年 11 月国家互联网信息办公室、文化和旅游部、国家广播电视总局联合发布的《网络音视频信息服务管理规定》、2021 年 9 月 17 日国家互联网信息办公室、中央宣传部、教育部、科学技术部、工业和信息化部、公安部、文化和旅游部、国家市场监督管理总局、国家广播电视总局联合发布的《关于加强互联网信息服务算法综合治理的指导意见》，以及 2022 年 12 月 11 日国家互联网信息办公室、工业和信息化部、公安部联合发布的《互联网信息服务深度合成管理规定》，结合深度伪造技术和深度伪造治理技术发展现状编制。

广播电视人工智能应用国家广播电视总局重点实验室将根据国家和行业有关深度伪造技术治理要求和发展现状，及时对本技术要求以及涉及的评价数据集进行完善和更新。

1 范围

本文件提出了深度伪造视听内容防范总体技术要求、面部识别技术要求、深度伪造鉴别技术要求、黑名单技术要求以及深度伪造防范能力评估要求，给出了深度伪造防范能力系统参考设计。

本文件适用于广播电视和网络视听机构在内容审核发布环节部署深度伪造防范能力系统时应具备的技术规范和要求。

2 总体技术要求

2.1 基本要求

广播电视和网络视听机构应部署运行深度伪造防范能力系统，全面防范深度伪造内容，包括但不限于以下技术措施：

（1）应具备人技一体的深度伪造防范机制；

（2）应具备重要人物面部识别能力；

（3）应具备深度伪造内容鉴别能力；

（4）应具备深度伪造内容分级管理能力；

（5）应具备黑名单管理能力；

（6）应具备深度伪造内容应急处置能力；

（7）鼓励视听机构开展黑名单可信共享。

2.2 分级治理要求

按照深度伪造内容的危害程度，将其分为三级，其中一级的危害程度最高，三级的危害程度最低。

一级是指涉及重要人物音频、影像的深度伪造内容；

二级是指不涉及重要人物音频、影像的新闻类深度伪造内容；

三级是指不涉及重要人物音频、影像的非新闻类的深度伪造内容。

广播电视和网络视听机构应具备深度伪造内容分级管理能力，要求如下：

（1）应具备将深度伪造内容及时更新到黑名单中的能力；

（2）应具备将一级和二级伪造内容与原始视听内容关联的管理机制，具备支撑应急处置情况下将一级和二级伪造内容及时替换为原始视听内容的能力；

（3）应具备及时下线伪造内容的能力。

2.3 人技一体防控要求

广播电视和网络视听机构深度伪造防范能力系统应具备人技一体的防范机制，确保机构按照章节 2.2 的要求实现深度伪造内容分级治理。具体要求如下：

（1）系统判定在黑名单中的视听内容应交由人工审核处理；

（2）不在黑名单中的内容应同时开展重要人物面部识别和深度伪造鉴别，根据重要人物面部识别和深度伪造鉴别结果做如下处理：

a）识别为包含重要人物，并且深度伪造鉴别为非伪造的，应直接进入人工审核流程，按照本机构的法定许可范围进行相应处理；

b）识别为包含重要人物或与重要人物相似，并且深度伪造鉴别为伪造的，应由人工审核确认为一级深度伪造内容后加入到黑名单中。

（3）识别为不包含重要人物，并且深度伪造鉴别为伪造的，根据内容的属性由人工审核进行处理：

a）如果内容属性为新闻类，应由人工审核确认为二级深度伪造内容后加入到黑名单中；

b）如果内容属性为非新闻类，但人工审核确认属于第三级的深度伪造内容，应禁止通过审核。

（4）识别为不包含重要人物，并且深度伪造鉴别为非伪造的，按照广播电视和网络视听机构日常审核流程处理。

2.4 应急能力要求

广播电视和网络视听机构深度伪造防范能力系统应具备应急处置能力，应急处置能力应至少符合以下要求：

（1）应具备深度伪造内容一键下线能力；

（2）宜具备二级以上深度伪造内容替换能力，及时消除影响。

2.5 安全性要求

广播电视和网络视听机构部署的深度伪造防范能力系统应至少满足以下数据安全、算法安全、应用安全、系统与网络安全要求，涉及到的人工智能算法应满足章节 2.5.1、2.5.2、2.5.3、2.5.4 中对人工智能算法的要求。

2.5.1 数据安全要求

（1）应保障黑名单数据的机密性、完整性和真实性；

（2）应保障人工智能算法训练数据来源等的可靠性；

（3）应保障人工智能算法训练数据集的多样性，训练数据集应包括对抗攻击、数据扰动等训练数据；

（4）应保障人工智能算法训练数据存储和使用等的安全性。

2.5.2 算法安全要求

（1）应保障人工智能算法设计过程的可解释性；

（2）应保障人工智能算法训练步骤的安全性；

（3）应保障人工智能算法训练过程和结果的可复现。

2.5.3 应用安全要求

（1）应保障人工智能算法部署过程的可追溯性；

（2）应保障人工智能算法在应用过程中的完整性和真实性；

（3）应保障人工智能算法所使用的第三方或开源深度学习框架和依赖库等的安全性，包括但不限于开展第三方或开源深度学习框架和依赖库的安全风险评估和安全加固；

（4）应保障人工智能算法对常见攻击的抵抗能力，包括但不限于对抗攻击、数据扰动等；

（5）应保障人工智能算法具备定期更新、自动升级维护的能力。

2.5.4 系统与网络安全要求

广播电视和网络视听机构所建的深度伪造防范能力系统应符合国家和行业网络安全相关要求。

3 面部识别技术要求

3.1 功能要求

广播电视和网络视听机构配备的重要人物面部识别能力应达到以下技术要求：

（1）应支持 MP4、AVI、WEBM、MKV、FLV、MOV 等视频文件格式；

（2）应支持 MPEG2、AVS+、AVS2、H.264、H.265、AV1 等视频编码格式；

（3）应支持定位视频中 64×64 像素以上的所有人物面部；

（4）应支持识别视频中 64×64 像素以上的重要人物面部。

3.2 性能要求

（1）应保障人物面部清晰无遮挡情况下面部定位准确；

（2）应保障人物面部清晰无遮挡情况下重要人物面部定位和识别准确；

（3）应保障人工智能算法的强泛化能力，包括但不限于对常见对抗攻击或数据扰动的防御能力。

3.3 适应性要求

（1）应具备对遮挡眉毛、眼睛、嘴巴、鼻子、下巴、额头、耳朵及脸部轮廓的重要人物面部进行识别的能力；

（2）应具备对面部过亮，面部光线不足或部分细节过暗、不清晰，以及光线不均匀等情形下的重要人物面部识别能力；

（3）应具备对倾斜角不超过 10°、水平转动角不超过 10°的重要人物面部识别能力；

（4）应具备对与重要人物相似面部的识别能力，包括与重要人物相似的卡通形象等。

3.4 安全性要求

面部识别技术安全性要求应满足章节 2.5 的相关规定。

4 深度伪造鉴别技术要求

深度伪造鉴别通常是指采用深度学习等人工智能技术检测视听内容是否被篡改或是否为合成内容，分析所用深度伪造技术类型和（或）原始视听内容等的一类技术。深度伪造鉴别技术应满足以下要求。

4.1 功能要求

（1）应支持 MP4、AVI、WEBM、MKV、FLV、MOV 等视频文件格式；

（2）应支持 AAC、WAV、MP3 等音频文件格式；

（3）应支持 MPEG2、AVS+、AVS2、H.264、H.265、AV1 等视频编码格式；

（4）应支持 AAC、AC3、MP3 等音频编码格式；

（5）应支持视频深度伪造鉴别，包括但不限于面部替换、表情交换、姿态迁移、人脸合成、场景编辑、场景合成；

（6）应支持视频中分辨率 64×64 像素以上的人物面部深度伪造鉴别；

（7）应支持音频深度伪造鉴别，包括但不限于语音剪辑、语音合成。

4.2 性能要求

（1）应保障在行业公认评价数据集上人物面部深度伪造鉴别准确；

（2）应保障在行业公认评价数据集上重要人物面部深度伪造鉴别准确；

（3）应保障人工智能算法的强泛化能力，包括但不限于对常见对抗攻击或数据扰动的防御能力。

4.3 适应性要求

深度伪造鉴别技术应具备对内容优化、媒体调和、数据处理等有意或无意掩盖深度伪造痕迹且未改变视听内容语义的“后处理”相关技术的适应性，具体要求如下：

（1）应具备对后处理的深度伪造视频内容的鉴别能力。视频后处理方法包括但不限于视频压缩等编码压缩处理，遮罩、美化、风格迁移、风格变化等场景后处理，以及瘦脸、磨皮、祛痘等美颜操作；

（2）应具备对后处理的深度伪造音频内容的鉴别能力。音频后处理方法包括但不限于添加噪音、杂音、歌曲、他人语音等背景声后处理，以及加速、减速、低沉、尖锐等编辑方法。

4.4 可解释性要求

（1）应具备对面部深度伪造定位的能力；

（2）应具备对深度伪造技术和原始视听内容溯源的能力；

（3）应可清晰描述深度伪造鉴别算法技术原理。

4.5 安全性要求

深度伪造鉴别技术的安全性要求应满足章节 2.5 的相关规定。

5 黑名单技术要求

5.1 功能要求

黑名单是指禁止播出的视听内容基本信息及其视音频特征信息的集合。广播电视和网络视听机构部署的黑名单技术系统应满足以下功能要求：

（1）黑名单中的内容应包括：鉴别为伪造且包含重要人物音频和影像的内容、鉴别为伪造的新闻内容等；

（2）黑名单元数据应包括：内容类型、基本信息、视音频特征信息、存储路径、音视频文本内容等；

（3）黑名单技术系统应具备视频特征分析技术，支持视音频内容的匹配与过滤；

（4）黑名单技术系统应具备对黑名单数据增加、删除、修改等管理功能。

5.2 视音频特征分析技术要求

用于黑名单技术系统的视音频特征分析技术应满足以下健壮性要求：

（1）视频叠加高斯、椒盐或斑点等噪声不应影响视频特征匹配；

（2）视频进行高斯滤波、中值滤波、均值滤波或锐化等处理不应影响视频特征匹配；

（3）对视频进行转码处理（如 H.264 转码为 H.265）不应影响视频特征匹配；

（4）对视频逐帧进行平移处理（如水平方向或垂直方向，平移距离相对于原视频图像比例不超过 5%）不应影响视频特征匹配；

（5）对视频图像行与列方向进行不等比例缩放处理不应影响视频特征匹配；

（6）对视频图像行与列方向进行等比例缩放处理不应影响视频特征匹配；

（7）对视频绕画面中心旋转不应影响视频特征匹配；

（8）对视频进行水平镜像处理不应影响视频特征匹配；

（9）对视频进行亮度调整（如亮度调整范围为 80%～120%）不应影响视频特征匹配；

（10）对视频进行对比度调整（如对比度调整范围为 90%～110%）不应影响视频特征匹配；

（11）对视频进行色度调整（如色调：-5º～5º，饱和度：95%～105%）不应影响视频特征匹配；

（12）对视频四个边侧区域进行部分遮挡处理（如图像四周用 5%

黑边遮挡、16:9 图像两侧用黑边遮挡变为 4:3 图像、弹幕遮挡、字幕遮挡、台标遮挡等）不应影响视频特征匹配；

（13）对视频进行帧率转换处理（如 50fps 与 25fps 帧率互换）不应影响视频特征匹配；

（14）对视频进行软件录屏处理，或使用摄影设备进行现场拍摄不应影响视频特征匹配；

（15）对视频进行信号采集处理不应影响视频特征匹配；

（16）视频中人物面部、表情、动作、对白、身体等的微小篡改不应影响视频特征匹配；

（17）视频片段重排等时序方面的编辑操作不应影响视频特征匹配；

（18）不应通过上述视音频特征逆向分析出视音频敏感内容。

6 深度伪造防范能力要求

深度伪造防范技术包括但不限于深度伪造鉴别、重要人物面部识别、深度伪造内容分级管理、深度伪造内容应急处置、黑名单管理等。深度伪造防范能力是广播电视和网络视听机构部署的用于防范深度伪造内容的一系列深度伪造防范技术的综合体现。

广播电视和网络视听机构应定期对部署的深度伪造防范能力系统进行评估，根据评估结果及时对深度伪造防范能力系统核心功能进行升级维护。

广播电视和网络视听机构应建立相应组织机构和机制，定期开展深度伪造防范能力培训、人技一体的深度伪造防范应急演练，确保深度伪造防范能力系统持续、有效地发挥作用。

参考文件

[1] 网络音视频信息服务管理规定，2019 年 11 月 18 日国家互联网信息办公室、文化和旅游部、国家广播电视总局（国信办通字〔2019〕3 号）

[2] 关于加强互联网信息服务算法综合治理的指导意见，2021 年 9 月 17 日国家互联网信息办公室、中央宣传部、教育部、科学技术部、工业和信息化部、公安部、文化和旅游部、国家市场监督管理总局、国家广播电视总局（国信办发文〔2021〕7 号）

[3] 2022 互联网信息服务深度合成管理规定，2022 年 11 月 3 日国家互联网信息办公室、中华人民共和国工业和信息化部、中华人民共和国公安部令（第 12 号）

附录 A 深度伪造防范能力系统参考设计

本附录给出一种符合深度伪造防范技术要求的深度伪造防范能力系统设计架构，为广播电视和网络视听机构设计、部署和运行深度伪造防范能力系统提供参考。

深度伪造防范能力系统设计架构如图 1 所示。

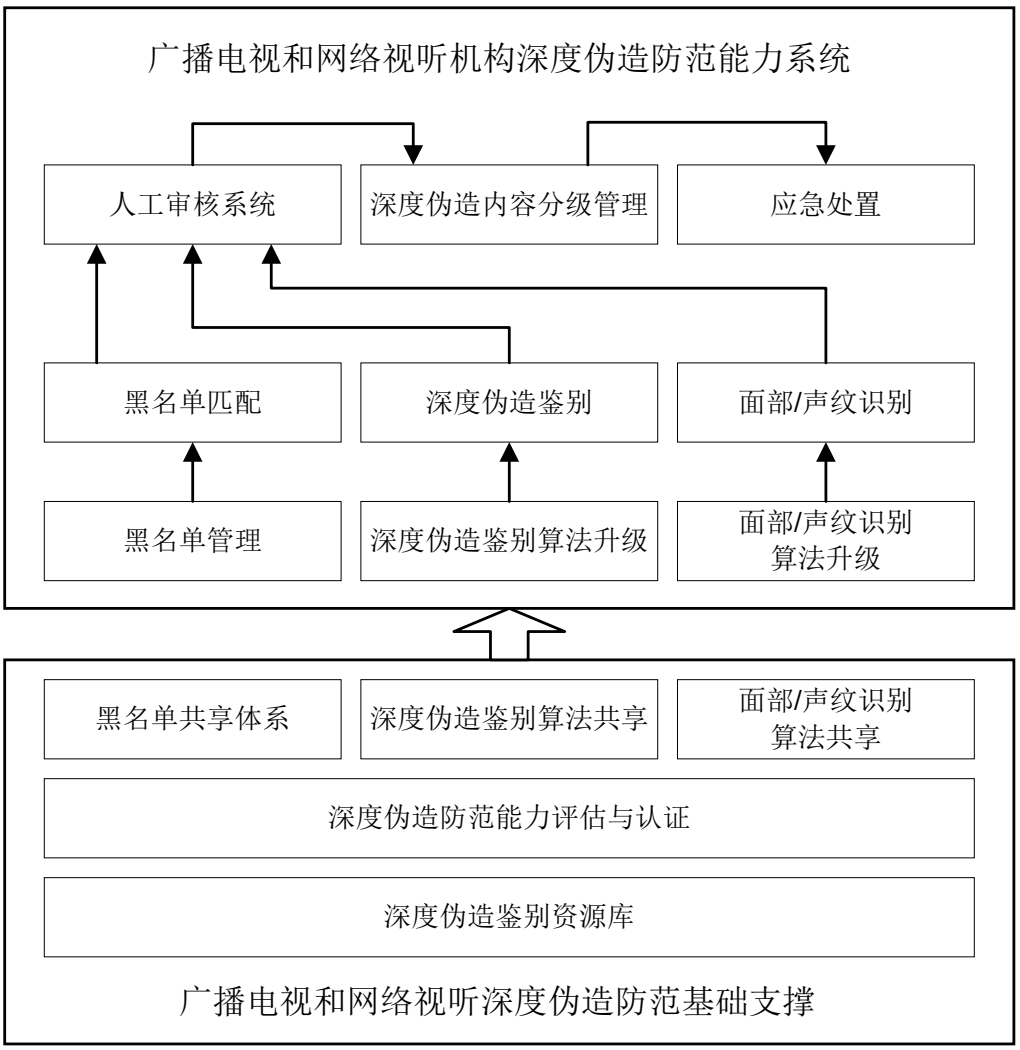

图 1 深度伪造防范能力系统设计架构图

广播电视和网络视听机构应在内容审核、发布、召回等各业务环节具备深度伪造防范技术能力。

（1）内容审核环节

a）部署黑名单管理与匹配等功能，用于黑名单数据与待审核内容的匹配分析；

b）部署深度伪造鉴别和面部/声纹识别技术用于对待审核内容进行分析；

c）根据黑名单匹配结果、深度伪造鉴别结果、面部/声纹识别结果等综合给出人工审核建议；

d）审核出的深度伪造内容应能够及时更新到黑名单和深度伪造内容分级管理环节。

（2）内容管理环节

应部署深度伪造内容分级管理功能，对识别的深度伪造内容进行分级管理，以支持下线、替换等应急处置。

（3）播出环节

应禁止播出深度伪造内容。

（4）应急管理环节

应增加应急处置功能，对漏审的内容进行一键下线，对深度伪造内容分级管理功能提供的关联内容进行下线、替换等应急操作。

（5）系统管理环节

应具备黑名单共享机制和算法升级机制，及时获取新的黑名单数据，及时更新深度伪造鉴别算法和面部/声纹识别算法。

此外，广播电视和网络视听机构应定期开展深度伪造防范能力系统技术评估，确保深度伪造防范能力持续、有效地发挥作用。