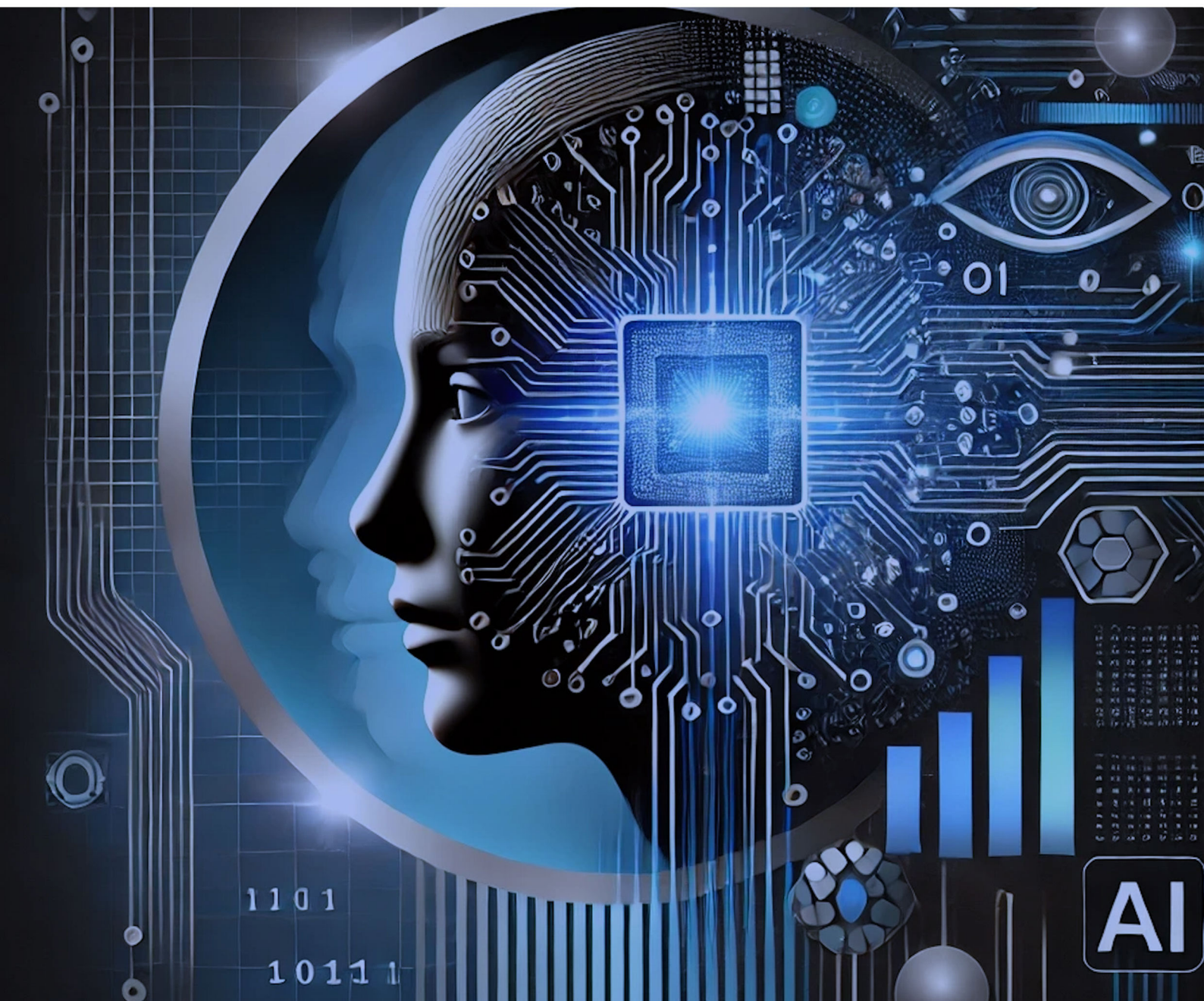


AI时代深度伪造和 合成媒体的安全威胁与对策



版权声明

本报告为北京谷安天下科技有限公司（以下简称“本公司”）旗下媒体平台安全牛研究撰写，报告中所有文字、图片、表格均受有关商标和著作权的法律保护，部分文字和数据采集于公开信息，所有权为原著者所有。未经本公司书面许可，任何组织和个人不得以任何形式复制或传递本报告的全部或部分内容，不得将本报告内容作为诉讼、仲裁、传媒所引用之证明或依据，不得用于营利或用于未经允许的其他用途。任何未经授权使用本报告的相关商业行为都将违反《中华人民共和国著作权法》和其他法律法规以及有关国际公约的规定。未经授权或违法使用本报告内容者应承担其行为引起的一切后果及法律责任，本公司将保留追究其法律责任的权利。

免责声明

本报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为本公司的当然客户。任何非本公司发布的有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以本公司发布的本报告完整版本为准。

本报告中的行业数据主要为分析师市场调研、行业访谈及其他研究方法估算得来，仅供参考。因调研方法及样本、调查资料收集范围等的限制，本报告中的数据仅服务于当前报告。本公司以勤勉的态度、专业的研究方法，使用合法合规的信息，独立、客观地出具本报告，但不保证数据的准确性和完整性，本公司不对本报告的数据和观点承担任何法律责任。同时，本公司不保证本报告中的观点或陈述不会发生任何变更。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的行为建议，也没有考虑到个别客户特殊的目的或需求。客户应考虑本报告中的任何意见是否符合其特定状况，若有必要应寻求专家意见。任何出现在本报告中的包括但不限于评论、预测、图表、指标、指标、理论、陈述均为市场和客户提供基本参考，您须对您自主决定的行为负责。本公司不对因本报告资料全部或部分内容产生的，或因依赖本报告而引致的任何损失承担任何责任，不对任何因本报告提供的资料不充分、不完整或未能提供特定资料产生的任何损失承担任何责任。

目 录

第一章 AI 时代深度伪造和合成媒体安全概念界定及背景分析 11

1.1 深度伪造和合成媒体概念界定	11
1.1.1 深度伪造概念界定与特征分析	11
1.1.2 合成媒体概念界定与特征分析	11
1.2 深度伪造和合成媒体安全发展阶段	12
1.3 深度伪造和合成媒体技术应用场景	13
1.4 深度伪造和合成媒体安全影响和痛点分析	14
1.5 深度伪造和合成媒体安全的政策、规范与标准	17
1.5.1 国外深度伪造和合成媒体安全法律分析	17
1.5.2 国内深度伪造和合成媒体安全政策、规范和标准分析	19

第二章 AI 时代深度伪造和合成媒体安全技术、威胁及对策研究 22

2.1 深度伪造和合成媒体安全技术原理分析	22
2.1.1 计算机图形学 (CG) 的原理与应用	22
2.1.2 自编码器 (AE) 的原理与应用	22
2.1.3 人工神经网络 (ANN) 的原理与应用	23
2.1.4 生成对抗网络 (GAN) 的原理与应用	24
2.1.5 扩散模型 (DDPM) 的原理与应用	25
2.2 深度伪造和合成媒体技术应用分析	25
2.2.1 视频领域：换脸、动作转移、表情操纵等	25
2.2.2 图像领域：人脸生成、属性编辑、风格转换等	26
2.2.3 语音领域：克隆、转换、合成等	27

2.3 深度伪造和合成媒体安全威胁分析.....	28
2.3.1 对个人层面的安全威胁	28
2.3.2 对社会层面的安全威胁	29
2.3.3 对国家层面的安全威胁	29
2.4 深度伪造和合成媒体安全检测类技术分析.....	30
2.4.1 数字水印溯源技术研究	31
2.4.2 基于区块链的内容认证机制	31
2.4.3 可信硬件支撑环境的构建.....	32
2.5 深度伪造和合成媒体安全检测类技术分析.....	33
2.5.1 聚焦深度伪造内容本身的分析检测方法	33
2.5.2 聚焦深度学习的检测方法.....	34
2.5.3 基于多模态融合的检测方法	35
第三章 AI 深度伪造和合成媒体安全的市场现状分析	37
3.1 深度伪造和合成媒体安全的市场表现分析.....	37
3.1.1 深度伪造和合成媒体市场规模分析	37
3.1.2 深度伪造和合成媒体公开中标信息分析	37
3.1.3 深度伪造和合成媒体厂商投融资情况分析.....	38
3.2 深度伪造和合成媒体安全生态图谱分析	39
3.3 深度伪造和合成媒体安全市场成熟度分析.....	41
3.4 深度伪造和合成媒体安全消费者敏感度分析	42
3.4.1 深度伪造和合成媒体安全消费者认知度分析	42
3.4.2 深度伪造和合成媒体安全消费者使用度分析	43
3.4.3 深度伪造和合成媒体安全消费者需求度分析	45
3.4.4 深度伪造和合成媒体安全消费者敏感度分析	46
3.4.5 深度伪造和合成媒体安全消费者保护度分析	47

第四章 AI 深度伪造和合成媒体安全市场技术落地分析 49

4.1 深度伪造和合成媒体安全能力建设分析	49
4.1.1 国内深度伪造和合成媒体安全研究侧能力分析	50
4.1.2 国内深度伪造和合成媒体安全服务侧能力分析	52
4.2 深度伪造和合成媒体安全防控体系分析	54
4.2.1 深度伪造和合成媒体安全防控体系—治理层面	54
4.2.2 深度伪造和合成媒体安全防控体系—管理层面	55
4.2.3 深度伪造和合成媒体安全防控体系—技术层面	56
4.3 深度伪造和合成媒体安全技术落地难点分析	59

第五章 国外深度伪造和合成媒体安全工具 / 平台介绍

(按首字母排序) 62

5.1 Deepware.....	62
5.2 Illuminarty	63
5.3 Sentinel.....	64
5.4 WeVerify	65

第六章 国内深度伪造和合成媒体安全厂商典型案例

(按简称字母排序) 66

6.1 顶象典型案例分析	66
6.2 东方通典型案例分析	69
6.3 国投智能典型案例分析.....	72
6.4 瑞莱智慧典型案例分析.....	74
6.5 网易易盾典型案例分析.....	76
6.6 中科睿鉴典型案例分析.....	79

第七章 AI 深度伪造和合成媒体安全市场未来五大趋势分析 82

7.1 深度伪造和合成媒体安全潜在市场规模巨大 82

7.2 未来三年，企业和大众将会深刻感受到深度伪造和合成媒体安全带来的挑战.....83

7.3 技术创新和应用创新双轮驱动，推动深度伪造和合成媒体安全行业快速发展.....83

7.4 国内深度伪造相关技术短期领先，长期发展需全生态协同推进 84

7.5 “融合”成深度伪造和合成媒体安全关键词 85

第八章 附录 86

8.1 本报告主要参考文献 86

8.2 国内深度伪造相关专利信息汇总表.....88

引言

本次《AI 时代深度伪造和合成媒体的安全威胁与对策》研究报告旨在传递两个核心理念：① AI 的发展已是大势所趋。无论个人意愿如何，我们都应以开放的心态去接受并拥抱这一趋势；② 对人类力量的坚定信心。无论技术如何发展，人类始终是创新和应对挑战的主导者。这一双重视角希望帮助读者在面对 AI 安全挑战时，既能正视其带来的影响，又能坚定应对的信心。

「无平不陂，无往不复。 坚贞无咎，勿恤其孚，于食有福」

— 《周易·泰》

根据唯物辩证法，事物是不断发展变化的，可是人们往往更多关注到突变的时刻，容易忽略事物变化的整个过程。本次《AI 时代深度伪造和合成媒体的安全威胁与对策》研究报告传达的理念也可以用“无平不陂，无往不复”来表达。这句话出自《周易·泰》泰卦，原文是“无平不陂，无往不复。坚贞无咎，勿恤其孚，于食有福”。其意是世上不存在只有平地而没有坡地的，没有只去不来的。只要坚定不移走下去，就没有什么好后悔的，可能有挫折，也可能会得到好的回报。这个卦不是否极泰来的意思，但可以理解为事物前途是光明的，发展的道路是曲折的，需要看准方向，坚定不移地往前走。

「技术可以在很大程度上左右人类这一物种的生物属性，但反过来人类也可以赋予技术以灵魂」

— 《技术与文明》

人类与技术的互动是一个动态的过程，技术的发展可以推动人类社会的进步，而人类的智慧和道德判断则可以引导技术向有益于人类的方向进化。这种相互作用确保了技术不仅仅是推动力，也是人类文化和精神的延伸。我们需要相信人类自身的智慧与力量，相信科技向善的潜力，相信我们能够为技术注入善意的灵魂，赋予它服务于人类福祉的使命！

报告关键发现

■ 深度伪造和合成媒体安全可分为四个发展阶段：大众认知阶段、监管介入阶段、技术对抗阶段、立法明确阶段。每个阶段具有持续性特点，并且会有重叠，这体现了深度伪造和合成媒体安全治理的复杂性和长期性。

■ 深度伪造和合成媒体技术带来的影响是深远和复杂的，安全牛从正反两面进行了总结。从正面来看，其在多个领域的巨大潜力和价值不容忽视，如个人娱乐、历史教育、亲人缅怀、医疗旅游等；从反面来看，随之而来的消极影响也不容小觑，如虚假宣传、名誉损坏、财产损失、道德伦理和安全威胁等多个角度。

■ 整体来看，美国是最早提出对深度伪造技术进行立法的国家，拥有最多的针对性法律法规；我国在相关政策制定和理论研究等方面也在积极推进中。例如，《网络音视频信息服务管理规定》、《互联网信息服务深度合成管理规定》、《生成式人工智能服务管理暂行办法》、《生成式人工智能服务安全基本要求》等的出台，推进了我国在深度伪造和合成媒体监管侧法规依据的不断完善。

■ 目前国内对现有的深度伪造安全防御技术根据策略不同进行划分有两种类型，即主动式防御（信任类）和被动式防御（检测类）。检测类是当前应对深度伪造负面对抗的主要形式。具体的检测技术有多种，目前基于多模态的检测方法是业内公认的有效手段。

■ 安全牛公开调研收集的深度伪造和合成媒体安全市场相关信息显示，2023 年我国深度伪造和合成媒体安全市场规模达到 1.19 亿元，预计 2024 年同比增长 243%，达到 4.08 亿元。其中，各部委、公检法、传媒集团、银行和运营商是深度伪造和合成媒体安全市场的主要采购客户。从投融资信息角度来看，2020 年是资方看好深度伪造和合成媒体创新厂商的主要时机，在这一年瑞莱智慧、深信科创获得了天使轮投资，中关村科金获得了股权投资。2024 年，获得投融资的两家企业为中科睿鉴和瑞莱智慧，两家企业均为国家队成员。

■ 深度伪造技术领域的重度参与者可以按照研究侧和服务侧两部分进行划分。研究侧包括拥有深度伪造专利认证的金融机构、科研院所和高校；服务侧则涵盖了提供深度伪造和合成媒体安全服务的企业，可分为以网络和数据安全为主营业务的企业，以及以人工智能为核心的企业和满足自我检测需求的互联网企业。目前，这三类服务企业均处于初级阶段。在整体能力上来看，存在各自的优势及问题，融合发展有望是未来的解决之道。

■ 安全牛在本研究项目中发起了《大众对深度伪造和合成媒体技术安全认知、态度及影响调研问卷》，从认知度、使用度、需求度、敏感度和保护度五方面进行了消费者分析。主要结论是：大众普遍对相关概念及其典型事件（如三只羊录音事件）有所了解，但在识别利用该技术制造的虚假内容方面存在明显困难；大众普遍对此技术存有好奇心，但同时也对此技术可能引发的诚信问题抱有普遍的担忧；大众对于安全需求保障的意愿较为强烈，特别是对监管侧和技术提供商寄予厚望；受访者普遍对合成媒体有较高的安全意识，大众也有意愿了解相关防范知识，并期待深度伪造检测工具的进一步落地。

■ 安全牛根据公开信息收集了 2021 年至今（2024 年 10 月）的深度伪造和合成媒体专利信息得出以下主要结论：专利申请人主要是高校，占比达到 57%，其次是企业和科研院所；专利申请主要集中在 2022 年和 2023 年；专利的法律状态主要集中在等待（实审提案、年登印费），占比达到 71%，真正获得专利权有效的申请者占比只有 22%。检测类型是深度伪造发明专利的主要类型，占比达到 77%，未来深度伪造技术在防护技术和系统实现方面还有较大发展空间。

■ 深度伪造和合成媒体安全的监管者需要变革。为应对深度伪造和合成媒体技术带来的挑战和问题，监管机构面临的任务愈发复杂且迫在眉睫。因此，监管机构需要采取全面、多维的策略，以确保能够有效识别、监管并防范深度伪造内容的不当使用。安全牛建议：监管机构需要探索不同的开放制度、平台和模型，以精准管控深度伪造和合成媒体的具体方向和落地细节。

■ 深度伪造和合成媒体安全内容方需要变革。随着生成式内容的指数级增长，内容方当前面临深刻变革。一方面，内容方需要顺应时代发展，满足大众日益提高的精神文化需要；另一方面，内容方需要建立专门的内容审查团队，负责监控上传的内容，快速响应并处理深度伪造内容的举报。安全牛建议：内容方需要满足消费者的多重需求，同时提前布局、审查加固和建立举报制度。

■ 深度伪造和合成媒体技术快速发展的背景下，消费者需要变革。消费者需要“擦亮双眼”，在深度伪造和合成媒体内容高速发展的当下，眼见不一定为实，在肉眼分不清现实和虚拟的时候，借助科技的力量已是必需的选择。安全牛建议：消费者在面对精彩纷呈的 AI 世界时，需参考“零信任”原则，并对看到的内容先持怀疑态度，然后再去证实。

■ 安全牛认为，虽然深度伪造和合成媒体安全应用的行业与场景各不相同，但均可按照制定管理计划—管理风险评估—管理框架搭建—构建安全模型—深度内容检测—事件响应计划—事件改进计划—持续更新教育的方式，建立健全有效防御体系。

■ 安全牛总结深度伪造和合成媒体有五大发展趋势：1) 深度伪造和合成媒体安全潜在市场规模巨大，预计从 B 端（主要是监管方、企业和机构）和 C 端（消费者）来看，市场规模可达到 30 亿元；2) 未来三年，企业和大众将会深刻感受到深度伪造和合成媒体安全带来的挑战；3) 技术创新和应用创新双轮驱动，将推动深度伪造和合成媒体安全行业快速发展；4) 国内深度伪造相关安全技术应用会短期领先，但长期发展需全生态协同推进；5) “融合”是深度伪造和合成媒体安全关键词。

第一章 AI 时代深度伪造和合成媒体安全概念界定及背景分析

1.1 深度伪造和合成媒体概念界定

深度伪造 (Deepfake) 和合成媒体 (Synthetic Media) 是人工智能技术发展演进的科技进化产物，如同一把双刃剑，在给我们带来便捷、高效以及愉悦同时，不可避免地会带来滥用、欺诈、伦理乃至更多潜在的风险。本章将重点梳理深度伪造和合成媒体的概念及背景，为下文探讨深度伪造和合成媒体技术原理、安全威胁及其应用奠定理论基础。

1.1.1 深度伪造概念界定与特征分析

深度伪造 (Deepfake) 取自英文 “Deep learning” (深度学习) 和 “Fake” (伪造)，是两者融合的产物，主要表现为利用人工智能技术，尤其是深度学习算法和生成式对抗网络来合成、篡改或伪造音频、视频等多媒体内容的技术手段。通过这项技术，可以生成高逼真的虚假媒体内容，包括但不限于虚假视频、音频、图像等，制造出以假乱真的内容。

学术界对深度伪造有着多种看法，清华大学有学者以 “深度伪造” 技术的典型化事实为例进行分析，界定其为 “颠覆性技术异化”；也有国内研究人员认为，深度伪造可以算作 “宏进化” 和 “微进化” 的结合，并提出其进化的渐进性。安全牛认为，深度伪造技术具有几个典型特性，简单总结为三 “高”：1) 高仿真程度 -- 利用该技术产出的视频、音频、图像细节呈现丰富，普通用户难以识别；2) 高度自动化 -- 只需少量样本或数据，可高效快捷批量生成；3) 高隐蔽性 -- 难以查明篡改痕迹和责任主体。

1.1.2 合成媒体概念界定与特征分析

合成媒体 (Synthetic Media) 是利用人工智能算法生成或操控的数字内容，代表了虚拟内容创作中人工智能技术的前沿成果。合成媒体具有高度的真实感和沉浸感，使得内容与现实世界的媒体难以区分，其典型应用如新闻播报员的数字分身、虚拟偶像或者游戏化身等。从广义来看，深度伪造也是合成媒体的一种，但合成媒体的概念外延更宽，它不局限于视听内容的篡改，更强调利用算法从无到有地 “创造” 数字内容。

安全牛认为，合成媒体具有上述深度伪造的特点，但其在应用场景方面与深度伪造略有不同。深度伪造主要用于生成诈骗影像、伪造色情视频、散布政治谣言等负面性的应用场景；而合成媒体背后动机更加复杂，除上述消极场景外，通常也大量广泛应用于虚拟主播、数字人等正向应用领域。随着技术门槛的不断降低，生成成本大幅下降，合成媒体带来的内容真伪冲击也随之加大。

1.2 深度伪造和合成媒体安全发展阶段

其实，深度伪造和合成媒体背后的针对图像、视频和音频的加工操作早就存在，甚至非正常修改或非法替换的技术也屡见不鲜。例如，利用 Photoshop、美图秀秀、天天 P 图、Final Cut、剪映等工具早已实现了对图像的实质性修改。但深度伪造和合成媒体技术在人工神经网络的快速发展后迎来了大跳跃，尤其是生成对抗网络 (generative adversarial network, GAN) 的出现，使深度伪造和合成媒体技术训练所需的样本数据量大幅下降，而“深度伪造物”的质量在快速提升，甚至出现了“以假乱真”的现实特点。

深度伪造概念首次出现在美国，2017 年 12 月，一位名为“DeepFakes”的用户在美国社交媒体平台发布换脸“假视频”。2018 年 1 月，“DeepFakeApp”用户创建“FakeApp”并开源，“深度伪造 (DeepFake)”一词由此首次正式被提出。此后，该技术逐渐陷入恶意滥用的泥沼。像伪造色情视频、政治人物演讲、企业家虚假新闻等层出不穷。2019 年，美国为了应对深度合成技术可能带来的风险，举办了多项检测深度伪造技术的赛事。如美国国家标准与技术研究院 (NIST) 发起的“媒体取证挑战赛”，旨在推进图像和视频取证技术的发展，以识别被篡改的图像和视频。Facebook 也在 2019 年发起了“Deepfake 检测挑战赛”，中科大俞能海教授—张卫明教授团队在此赛事中荣获了第二名。此次比赛共收到超过 35000 个检测模型，最终得出的最高准确率为 82.56%。

我国深度伪造和合成媒体的应用虽比国外晚了一段时间，但是不论技术产品还是监管力度都在快速发展。安全牛认为，我国深度伪造和合成媒体安全发展可分为四个阶段：



图 1.2 1 中国深度伪造和合成媒体安全发展阶段图

■ **大众认知阶段：**2019 年 2 月，一位名为“换脸哥”的博主在哔哩哔哩上传了一段将杨幂的面孔替换成朱茵版黄蓉的视频，引发社会广泛讨论。直至五年后，围绕名人的深度伪造事件依然频发。如今年国庆期间，通过深度伪造和 AI 配音技术合成的“雷军”视频，以其尖锐言词批评社会热点，迅速在网络中传播，引起公众的广泛关注。

■ **监管介入阶段：**2019 年 9 月，国内首个深度伪造技术应用“ZAO”换脸 App 因涉嫌侵犯隐私被约谈并最终下架，成为监管介入的标志性事件。2021 年，另一款换脸软件“蚂蚁呀嘿”也因监管问题在中国区下架，显示我国监管机构的持续关注和迅速响应。

■ **技术对抗阶段：**“用技术打败技术”从未停止。学术界、产业界一直在研究技术对抗的办法，如 2020 年，北京大学和微软亚洲研究院联合发布 Face X-Ray，此为针对伪造人脸图像的通用检测工具。又如 2021 年 1 月就有企业开始申请深度伪造相关专利。

■ **立法明确阶段：**2024 年 4 月，北京互联网法院一审开庭宣判了中国首例 AI 生成声音侵权案，明确认定在具备可识别性前提下，自然人声音权益的保护范围可及于 AI 生成声音。2024 年 6 月，北京宣判首例“AI 换脸”侵权案，认定使用他人视频“换脸”后制作模板再提供“换脸”服务的网络服务提供者侵害了他人个人信息权益。

值得一提的是，我国深度伪造和合成媒体安全发展的四个阶段具有持续性特点，每个阶段都可能持续较长时间，并且可能会有重叠。这也体现了深度伪造和合成媒体安全治理的复杂性和长期性。

1.3 深度伪造和合成媒体技术应用场景

深度伪造与合成媒体技术虽然在某些应用场景中带来了负面影响，但其在多个领域的巨大潜力和价值不容忽视。如在个人娱乐领域，技术如 FaceSwap 和 DeepFaceLab 使得用户能够轻松地将自己的脸替换到视频中，为社交媒体和个人分享带来新的乐趣；在教育领域也受益于这些技术，通过创造更加生动的教学材料，如历史事件的图像融合，增强了学习体验；在影视制作中，合成媒体技术如 Stable Diffusion 和 Midjourney 被用于生成逼真的背景和特效，节省了传统拍摄的成本和时间；医疗行业通过合成媒体技术模拟手术过程，提高了手术训练的真实性；艺术领域中，合成技术激发了艺术家的创造力，如通过图像融合技术创作数字艺术作品；新闻媒体和娱乐行业则利用这些技术快速生成报道和娱乐内容，或者运用 Heygen 生成数字人，提高了内容生产的效率；在隐私保护方面，合成媒体技术通过替换真实人物的面部信息，保护了个人隐私，尤其是在需要匿名处理的场合；语言翻译方面，讯飞智作等平台提供了语音合成服务，使得多语种的内容创作成为可能。



安全牛认为，当前，众多应用平台正在降低深度伪造和合成媒体技术的门槛，使得普通用户能够轻松地利用这些软件实现具体的应用。无论是国际知名的平台如 OpenAI、Heygen、Midjourney、Runway，还是国内的讯飞智作、剪映等，都在积极推动合成媒体技术的发展和應用。这些平台提供了一系列从文本到视频的全面服务，极大地简化了内容创作的流程，提高了效率。此外，开源应用如 Stable Diffusion、FaceSwap 等，为个人用户和开发者提供了强大的工具，进一步促进了合成媒体技术的普及和创新。随着技术的持续进步，相信合成媒体技术将在更多领域扮演关键角色，将对社会产生深远影响。

1.4 深度伪造和合成媒体安全影响和痛点分析

深度伪造和合成媒体技术带来的影响是深远和复杂的，安全牛总结这些影响涉及虚假宣传、名誉损坏、财产损失、道德伦理和安全威胁等多个角度。以下结合具体事件展开详细分析：

表 1.4.1 2017-2024 国内外深度伪造和合成媒体典型事件汇总表

事件时间	国家 / 地区 名称	事件名称	事件内容
2017 年	美国	盖尔·加朵遭色情换脸	Reddit 上的“deepfakes”用户利用深度伪造技术，将《神奇女侠》主演盖尔·加朵的面容合成到了一部成人影片中，引发桃色风波
2018 年	印度	印度记者遭色情报复视频困扰	印度著名调查记者 Rana Ayyub 成为色情复仇视频的受害者

2018 年	美国	奥巴马骂特朗普假视频	一段伪造的奥巴马视频在推特上播放超 200 万次，点赞超 5 万，视频中奥巴马批评特朗普是“彻头彻尾的笨蛋”。示上，这是美国喜剧演员乔丹·皮尔声音，旨在展示深度伪造技术
2018 年	美国	特朗普多次被深度伪造“碰瓷”事件一	特朗普宣布中止全球气候变化协议，随后被比利时某政党利用“深度伪造”技术篡改，做出“特朗普宣告比利时政府也应退出”的假视频，引起比利时民众的公愤
2019 年	美国	特朗普多次被深度伪造“碰瓷”事件二	福克斯旗下西雅图电视台播出了一段被深度伪造的特朗普演讲视频，视频中特朗普脸色异常红润，频繁做出夸张动作
2019 年	美国	扎克伯格被伪造假视频发布演讲	一段关于 Facebook CEO 马克·扎克伯格的人工智能深度合成 (deepfake)“换脸”视频被上传到 Instagram 上，视频中他讨论大数据，但他否认说过这些话
2019 年	中国	杨幂遭换脸风波	网友用杨幂脸替换《射雕英雄传》里黄蓉扮演者朱茵脸，发布的视频引发网络热议
2020 年	美国	爱因斯坦“穿越”视频	合成的爱因斯坦在视频中鼓励大家团结一致共同抗疫
2022 年	中国	古风汉服网红魏某“AI 换脸”App	成都铁路运输第一法院审理了魏某诉四家 AI 换脸软件公司侵犯肖像权案。魏某为汉服网红，其肖像被未经授权用于 AI 换脸视频全球首部运用深度伪造技术合成名人面孔的节目《深度伪造邻居之战》于 2023 年 1 月在英国 ITVX 平台首播
2023 年	英国	《深度伪造邻居之战》播出	全球首部运用深度伪造技术合成名人面孔的节目《深度伪造邻居之战》于 2023 年 1 月在英国 ITVX 平台首播
2023 年	美国	乔·罗根被“宣传”	一段由人工智能制作的假乔·罗根 (Joe Rogan) 宣传男性保健品的深度伪造视频，在某个社交媒体平台迅速上传
2024 年	中国	雷军骂人事件	十一期间，用锋利且刺耳的言辞抨击堵车、调休、游戏等热门话题的“雷军（采用深度伪造技术合成的雷军）骂人”视频引发网络热议
2024 年	中国	三只羊录音事件	9 月，一段伪造的录音在网络上传播，内容涉及三只羊公司创始人卢文庆的言论，最终被证实为 AI 合成，导致公司形象受损并引发公众热议
2024 年	韩国	韩国“N 号房 2.0 事件”爆发	韩国“N 号房 2.0 事件”爆发，深度伪造技术被用于制作和传播淫秽影像，涉及人数高达 22 万人
2024 年	中国	香港深度伪造诈骗案损失 2 亿港币	香港一家跨国公司员工被邀请进入用深度伪造制作的“高管视频会议”，被骗转账 2 亿港币，成为香港历史上损失惨重的“变脸”案例

数据来源：安全牛根据公开信息自主整理，按时间顺序排列，自主绘制

AI 时代深度伪造和合成媒体安全概念界定及背景分析

■ **虚假宣传—秩序扰乱与信任风险：**深度伪造技术可以被用于制作虚假视频和音频，这些内容可以被用于政治宣传或误导公众，从而破坏社会稳定甚至国际关系。例如，2018 年，特朗普宣布中止全球气候变化协议，随后被比利时某政党利用“深度伪造”技术篡改，做出“特朗普宣告比利时政府也应退出”的假视频，引起比利时民众的公愤。再如，一段由人工智能制作的假乔·罗根（Joe Rogan，美国知名博客）宣传男性保健品的深度伪造视频，在某个社交媒体平台上迅速传播，涉及虚假宣传。

■ **名誉损坏—政策挑战与权利侵犯：**深度伪造技术制作的色情换脸视频侵犯了公民的名誉权和肖像权，对被伪造者造成严重的心理和社会影响。这些视频的传播可能导致受害者产生焦虑、抑郁等心理障碍，甚至影响其工作就业与社会交往。多国社会名人、企业家受到波及。如印度著名调查记者 Rana Ayyub 成为色情复仇视频的受害者；《神奇女侠》主演盖尔·加朵的面容合成到了一部成人影片中，引发桃色风波。

■ **财产损失—经济风险与经济犯罪：**深度伪造技术被用于诈骗和勒索，导致个人和企业遭受经济损失。例如，语音克隆技术能够在采集目标人物几句话的基础上，就通过算法生成这个人的全套语音，让人难辨真假。目前，利用假视频、假音频进行碰瓷、诈骗的现实案例层出不穷。如 2024 年，香港一家跨国公司员工被邀请进入用深度伪造制作的“高管视频会议”，被骗转账 2 亿港币，成为香港历史上损失最惨重的“变脸诈骗”案例。

■ **道德伦理—伦理风险与法律风险：**深度伪造技术的应用引发了伦理问题，如侵犯已故明星与名人的主体价值完整性，以及对逝者的“赛博重生”可能违反了公民应有的被遗忘权原则等。如疫情期间，合成的爱因斯坦在视频中鼓励大家团结一致共同抗疫引发大众热议。此外，深度伪造技术还可能导致道德行为失序，如伪造视频可能引发对个体身份的操纵和数字暴力。

■ **安全威胁 - 国家安全与国际关系：**深度伪造技术可能被用于制作虚假的战争命令或解锁重要设施的人脸识别系统，从而对军事安全构成威胁。此外，深度伪造技术还可能被用于网络钓鱼攻击、交易欺诈等，对企业安全构成威胁。深度伪造技术可能被用于国家间的政治抹黑、军事欺骗，增加国家间的危机和冲突，尤其是增加拥有核武器国家之间爆发核战争的风险。例如，在敏感关系或战争国家间流传的深度伪造视频可能破坏国际的战略互信，助长冲突双方的仇恨。

综上所述，深度伪造技术的影响深远且复杂，需要通过技术研发、立法管制、加强宣传教育以及推进国际安全合作治理等手段来应对其被滥用而带来的安全挑战。

1.5 深度伪造和合成媒体安全的政策、规范与标准

本节将从国外和国内两个方面进行深度伪造和合成媒体安全相关法律、政策和规范的分析与描述。

1.5.1 国外深度伪造和合成媒体安全法律分析

安全牛认为，从整体来看，美国是最早对深度伪造技术进行立法的国家，拥有最多的针对性立法，并在联邦及州政府层面采取了广泛的行动。2018 年 12 月，美国参议院提出《恶意深度伪造禁令法案》，旨在打击制作和传播深度伪造内容的行为，可对相关个人和社交媒体平台处以罚款和监禁。若伪造内容引发严重后果，如煽动暴力或扰乱选举，监禁可长达 10 年。2019 年 6 月，众议院提出《深度伪造责任法案》，要求制作者在深度伪造视频上添加不可删除的水印和文本说明，否则视为犯罪。同月，众议院在《2020 财政年度达蒙·保罗·尼尔森和马修·杨·波拉德情报授权法》中建议举办技术竞赛以促进鉴伪技术的发展。同时，两党议员在众议院和参议院提出《深度伪造报告法案》，明确“数字内容伪造”的定义，并要求国土安全部定期发布相关报告，由分管科学和技术事务的副部长负责。在州级层面，得克萨斯州通过法案，将使用深度伪造技术干扰选举的行为定为刑事犯罪。弗吉尼亚州将制作和传播深度伪造视频（例如色情暴力等），违法者可面临监禁和罚款。加州州长签署法案，将发布对政客产生负面影响的虚假音视频定义为犯罪。

表 1.5 1 国外深度伪造和合成媒体安全的法律汇总表

国内 / 国外	时间	法律名称	立法机构	法律内容
国外	2018	《恶意伪造禁令法案》	美国国会	对制作深度伪造引发犯罪和侵权的个人、明知内容的性质仍进行分发的媒体平台处以罚款和监禁
国外	2018	《反虚假信息行为准则》	欧盟委员会	认识到虚假信息的传播存在众多侧面且受到生态系统中各类不同部门多与者的促进，并对这些参与者造成影响，所有利益相关者在消除虚假信息的传播方面均应当发挥作用
国外	2019	《非同意色情法》	联邦法律 - 美国弗吉尼亚州	将利用“深度伪造”技术合成他人色情视频并予以传播的行为犯罪化
国外	2019	《关于伪造欺诈视频影响选举结果的刑事犯罪法案》	联邦法律 - 美国得克萨斯州	将利用“深度伪造”技术合成候选人虚假视频并予以传播的行为犯罪化
国外	2019	《深度伪造责任法案》	美国国会	任何创建深度伪造视频媒体文件的人，必须用“不可删除的数字水印以及文本描述”来说明该媒体文件是篡改或生成的，否则将属于犯罪行为

● AI 时代深度伪造和合成媒体安全概念界定及背景分析

国外	2020	《2020 财政年度达蒙·保罗·尼尔森和马修·杨·波拉德情报授权法》	美国众议院	针对深度伪造对服役人员及其家人构成的威胁进行情报评估，包括评估该技术的成熟度以及如何使用该技术进行信息作战
国外	2020	《识别生成对抗性网络的输出法案》	美国国家标准与技术研究院	旨在支持开发必要的测量和标准的研究，以开发工具来检查 GANs 或其他合成或操纵内容的技术的功能和输出
国外	2020	《2020 财年国防授权法案》	美国众议院	要求国家情报局负责人就潜在的影响国家安全的深度伪造技术、以及为外国用以实施传播虚假信息等恶意行为的深度伪造技术的潜在与实际影响向国会情报委员会进行汇报
国外	2021	《人工智能法案》	欧盟	将深度伪造列为“高风险”人工智能技术，欧洲议会认为该归类的合理性来自于深度合成技术对个人基本权利及安全的威胁，《人工智能法案》于 2021 年被首次提出，2024 年 8 月 1 日正式生效。

数据来源：安全牛根据公开信息自主整理，自主绘制。注意：本表仅罗列与深度伪造最为相关的部分国外法律。

其他国家（或组织）在应对深度伪造技术方面的举措。欧盟主要通过个人信息保护和虚假信息治理的法律法规来进行限制。2018 年 9 月，欧盟发布了《反虚假信息行为准则》，认识到虚假信息的传播存在众多侧面，且受到生态系统中各类不同部门参与者的促进，并对这些参与者造成影响，所有利益相关者在消除虚假信息的传播方面均应发挥作用。

其他国家虽然没有专门针对深度伪造技术的直接立法，但是通过现有的公民信息保护、虚假信息治理、社交媒体平台管控和网络审查等法律法规，对涉及造假技术的犯罪案件进行审理和规制。例如，新加坡通过《**广播法**》和《**互联网操作规则**》等法规，要求电信服务供应商承担屏蔽、封堵特定网站和信息言论的义务，以及及时发现并依法查处虚假信息；英国则通过《**数据保护法**》保护个人脸部图像等信息，适用于可能被用于制作深度伪造内容的数据；韩国通过《**电子传播商务法**》和《**个人信息保护法**》等法规，对网络舆情内容进行审查，并保护个人数据不被未经同意的情况下收集与使用；俄罗斯则通过一系列法规，要求网站迅速清理非法信息，并对违法行为进行打击。这些国家的法律 / 法规 / 政策虽然不是专门针对深度伪造技术，但目前主要用于应对该技术可能带来的法律问题和犯罪行为。

1.5.2 国内深度伪造和合成媒体安全政策、规范和标准分析

随着深度伪造技术快速发展，我国在政策制定和理论研究等方面也积极推进。国内针对深度伪造的法律规定主要涉及肖像权、互联网法规以及个人信息等领域。《中华人民共和国民法典》第一千零一十九条规定：“任何组织或个人不得以丑化、污损，或者利用信息技术手段伪造等方式侵害他人的肖像权。未经肖像权人同意，不得制作、使用、公开肖像权人的肖像，但是法律另有规定的除外。”根据该规定，即使没有营利目的和主观恶意，未经本人同意的深度伪造行为仍可能侵害肖像权。

2019 年，国家互联网信息办公室（以下简称网信办）、国家广播电视总局、文化和旅游部联合发布《**网络音视频信息服务管理规定**》，明确提出网络音视频信息服务提供者和网络音视频信息服务使用者利用基于深度学习、虚拟现实等新技术新应用制作、发布、传播非真实音视频信息的，应当以显著方式予以标识。同时，不得利用基于深度学习、虚拟现实等新技术新应用制作、发布、传播虚假新闻信息。同年，网信办还发布了《**网络信息内容生态治理规定**》、《**数据安全管理办法**》和《**具有舆论属性或社会动员能力的互联网信息服务安全评估规定**》，明确提出网络信息内容服务使用者和生产者、平台不得开展网络暴力、人肉搜索、深度伪造、流量造假、操纵账号等违法活动。再次提出网络运营者利用大数据、人工智能等技术自动合成新闻、博文、帖子、评论等信息，应以明显方式标明“合成”字样；具有舆论属性或社会动员能力的互联网信息服务提供者拒不按照本规定开展安全评估的，应当通过全国互联网安全管理服务平台向公众提示该互联网信息服务存在安全风险，并依照各自职责对该互联网信息服务实施监督检查。

2020 年，全国人大三次会议表决通过的《**中华人民共和国民法典**》对于保护公民肖像权所作出的规定关注到了“深度伪造”这一技术前沿问题，但采用的是“利用信息技术手段伪造”的表述。2021 年，网信办、工信部、公安部和市场监督管理总局发布《**互联网信息服务算法推荐管理规定**》，明确将生成合成类算法列入需监管的五大类应用算法（生成合成类、个性化推送类、排序精选类、检索过滤类、调度决策类）推荐技术之一，并要求对 AI 生成图片、视频进行标识。2022 年，网信办、工信部和公安部发布《**互联网信息服务深度合成管理规定**》采用“深度合成”的概念，定义“深度合成技术”包括人脸生成、人脸替换、人物属性编辑、人脸操控、姿态操控等生成或者编辑图像、视频内容中生物特征的技术。明确任何组织和个人不得利用深度合成服务制作、复制、发布、传播法律、行政法规禁止的信息，不得利用深度合成服务从事危害国家安全和利益、损害国家形象、侵害社会公共利益、扰乱经济和社会秩序、侵犯他人合法权益等法律、行政法规禁止的活动。《互联网信息服务深度合成管理规定》还对深度伪造提出较有针对性的规制。如因深度伪造涉及人脸等生物识别信息的显著编辑功能，《管理规定》第十四条第 2 款规定，深度伪造服务提供者和技术支持者应当提示服务使用者依法告知人脸被编辑的个人（即肖像权人），并取得其单独同意。此外，为避免深度伪造可能导致公众混淆或者误认，《管理规定》第十七条第 1 款第 3 项规定，服务提供者应当在生成或者编辑的信息内容的合理位置、区域进行显著标识，向公众提示深度合成情况。

2023 年，网信办发布《生成式人工智能服务管理暂行办法》，明确生成式人工智能服务提供者应当按照《互联网信息服务深度合成管理规定》对图片、视频等生成内容进行标识。2024 年，全国网络安全标准化技术委员会发布《生成式人工智能服务安全基本要求》明确指出要对生成式人工智能的生成内容进行标识，并对利用生成式人工智能服务进行的违法行为要依法依规采取有关处置措施。

表 1.5 2 国内深度伪造和合成媒体安全的政策、规范与标准汇总表

国内 / 国外	时间	政策 / 规范名称	发布机构	政策内容
国内	2019	《网络音视频信息服务管理规定》	网信办、国家广播电视文化和旅游部	网络音视频信息服务提供者和网络音视频信息服务使用者利用基于深度学习、虚拟现实等的新技术新应用制作、发布、传播非真实音视频信息的，应当以显著方式予以标识，不得利用基于深度学习、虚拟现实等的新技术新应用制作、发布、传播虚假新闻信息
国内	2019	《网络信息内容生态治理规定》	网信办	明确提出网络信息内容服务使用者和生产者、平台不得开展网络暴力、人肉搜索、深度伪造、流量造假、操纵账号等违法活动
国内	2019	《数据安全管理办法》	网信办	网络运营者利用大数据、人工智能等技术自动合成新闻、博文、帖子、评论等信息，应以明显方式标明“合成”字样；不得以谋取利益或损害他人利益为目的自动合成信息
国内	2019	《具有舆论属性或社会动员能力的互联网信息服务安全评估规定》	网信办	网信部门和公安机关发现具有舆论属性或社会动员能力的互联网信息服务提供者拒不按照本规定开展安全评估的，应当通过全国互联网安全管理服务平台向公众提示该互联网信息服务存在安全风险，并依照各自责对该互联网信息服务实施监督检查
国内	2020	《中华人民共和国民法典》	全国人大	对于保护公民肖像权所作出的规定关注到“深度伪造”这一技术前沿问题，但采用的是利用信息技术手段伪造”的表述
国内	2021	《互联网信息服务算法推荐管理规定》	网信办、工信部、公安部、市场监督管理总局	将生成合成类算法列入监管的五大类应用算法（生成合成类、个性化推送类、排序精选类、检索过滤类、调度决策类）推荐技术之一，并要求对 AI 生成图片、视频进行标识

国内	2022	《互联网信息服务深度合成管理规定》	网信办、工信部、公安部、市场监督管理总局	采用的是“深度合成”的概念，明确任何组织和个人不得利用深度合成服务制作、复制、发布、传播法律、行政法规禁止的信息，不得利用深度合成服务从事危害国家安全和利益、损害国家形象、侵害社会公共和益、扰乱经济和社会秩序、侵犯他人合法权益等法律、行政法规禁止的活动
国内	2023	《生成式人工智能服务管理暂行办法》	网信办	明确生成式人工智能服务提供者应当按照《互联网信息服务深度合成管理规定》对图片、视频等生成内容进行标识
国内	2024	《生成式人工智能服务安全基本要求》	全国网络安全标准化技术委员会	明确指出要对生成式人工智能的生成内容进行标识，并对利用生成式人工智能服务进行的违法行为要依法依规采取有关处置措施

数据来源：安全牛根据公开信息自主整理，按时间顺序排列，自主绘制

整体来看，这些政策的制定与执行彰显了我国在深度伪造及合成媒体安全管理方面的坚定立场，其目标是维护国家安全、社会公共利益，以及保障公民、法人和其他组织的合法权利。此外，这些措施也为深度合成技术的有序发展和合规使用提供了支持。

第二章 AI 时代深度伪造和合成媒体安全技术、威胁及对策研究

2.1 深度伪造和合成媒体安全技术原理分析

安全牛认为，深度伪造和合成媒体的关键技术主要有以下四类：计算机图形学（Computer Graphics，CG）、自编码器（AutoEncoder，AE）、人工神经网络（Artificial Neural Network，ANN）、生成对抗技术（Generative Adversarial Network，GAN）和扩散模型（Denoising Diffusion Probabilistic Models，DDPM）。本章节将分别就这些关键技术的基本原理进行重点剖析。

2.1.1 计算机图形学（CG）的原理与应用

计算机图形学（Computer Graphics，简称 CG） 技术在过去二十年中不断进步，已成为实现高真实度人脸伪造的关键技术。在人脸领域的应用，CG 技术主要依赖面部检测、追踪、建模和纹理映射等核心技术。得益于其高度成熟和人工调整手段的结合，CG 技术能够创造出极其逼真的人脸效果，曾一度被认为是最先进的人脸伪造技术之一。然而，为了实现这种高度逼真的效果，需要投入大量的硬件、软件资源以及专业团队，导致成本极高。因此，CG 技术在人脸合成和替换方面的应用主要局限于成本较高的商业领域，如电影和游戏产业。

在视频行业中，CG 渲染与合成技术也被用于深度伪造技术。例如，FaceSwap 技术通过获取人脸关键点，利用 3D 模型进行渲染，并不断调整以缩小与目标关键点的差异，最终通过色彩校正使渲染结果更加自然。Thies 等研究人员通过 RGB-D 相机重建源人脸和目标人脸的 3D 模型，并进行融合渲染，以实现实时的表情控制。尽管基于图形学的方法通常需要精确的 3D 建模和较高的计算成本，但近年来，一些研究开始将传统图形学方法与深度学习相结合，使用神经渲染技术替代手工 3D 建模，即使在不完美的 3D 内容上也能实现良好的操控效果。不难发现，未来深度伪造技术将更多地和其他技术融合，提升产业效率，降低生产成本。

2.1.2 自编码器（AE）的原理与应用

自编码器（AutoEncoder，AE） 技术在深度伪造领域扮演了重要的角色。作为一种无监督学习模型，自编码器通过编码器将输入数据压缩成低维表示，再通过解码器恢复成原始数据。在深度伪造技术中，自编码器被用来学习人脸特征，并生成逼真的人脸图像。例如，深度伪造技术利用自编码器学习原始和目标人脸特征，通过交换解码器实现人脸替换，降低了技术门槛，仅需少量照片即可完成换脸。

随着深度学习技术的发展,自编码器的变体如变分自编码器 (VAE) 和生成对抗自编码器 (AAE) 被开发出来,通过引入额外机制改善生成模型性能。这些技术的发展使得深度伪造技术只需少量肖像即可完成换脸,推动了自动编码器、生成对抗网络等技术在深度伪造中的应用。自编码器还可识别数据中的异常模式,用于欺诈检测等领域。

自编码器通过其独特的结构和训练机制,在无监督学习领域提供了一种强大的工具,能够自动学习数据的内在结构和特征,进而在多种应用中发挥作用。随着深度学习技术的不断进步,自编码器及其变体在解决复杂问题上展现出巨大的潜力。

2.1.3 人工神经网络 (ANN) 的原理与应用

人工神经网络 (Artificial Neural Network, ANN) 模拟人脑神经元连接,通过调整神经元之间的权重,学习样本数据中蕴含的特征。深度伪造所用的 ANN 一般采用卷积神经网络 (CNN) 和循环神经网络 (RNN) 等典型架构。

CNN 善于提取图像中的多层次特征。典型的 CNN 由多个卷积层和池化层交替堆叠而成,逐层将图像抽象为更高级的特征。这些特征再输入到全连接层进行分类、回归等任务。CNN 已成为图像篡改检测的主流方法。例如 Rossler 等人利用 Xception 架构对视频帧提取特征进行真假分类,取得了优于传统方法的性能。RNN 则擅长处理序列数据。相比前馈网络, RNN 引入了记忆机制,可以捕捉时序特征。如 Guera 等人用 CNN 和 LSTM 级联,让 LSTM 学习视频帧的时序信息以检测深度伪造。

Transformer 是一类相对较新的神经网络,彻底改变了自然语言处理。它们依靠注意力机制来捕获输入序列不同部分之间的关系,使它们能够比 RNN 更有效地处理远程依赖关系。Transformer 在深度伪造和合成媒体中的应用: 1) 基于文本的深度伪造: Transformer 可用于生成逼真且连贯的文本,可用于创建合成对话或处理深度伪造中的现有文本; 2) 音频合成: Transformer 越来越多地用于语音合成,生成高质量且自然的语音; 3) 视频生成: Transformer 在视频生成方面显示出良好的效果,可以捕获远程时间依赖性并创建更连贯、更逼真的视频; 4) 多模态深度伪造: Transformer 可用于组合不同的模态,例如文本、音频和视频,以创建更复杂、更具说服力的深度伪造。

无论采取何种架构, ANN 都需要大量标注数据来训练。目前深度伪造检测领域面临高质量数据集缺乏的问题。一些研究者通过数据增强等方法缓解这一问题。如 Cozzolino 等人在训练样本中加入 JPEG 压缩、模糊等扰动,使得检测模型的鲁棒性和泛化性大幅提升。此外,一些研究还探索将对抗样本技术用于提升检测模型鲁棒性。

2.1.4 生成对抗网络 (GAN) 的原理与应用

生成对抗网络 (GAN) 是深度伪造最为核心的技术，是人工神经网络 (Artificial Neural Network, ANN) 的一个特殊类型。与传统深度学习技术的单链条模式不同，GAN 引入了独特的“对抗”机制，其通过采用两条并行的学习神经网络，在互动式对抗中逐步优化学习成果。其中，一组神经网络被定位为“生成器”，其职责是依据“源数据”构建目标图像模型，进而生成伪造图像；另一组则定位为“鉴别器”，以真实目标图像为标准，对“生成器”的合成作品进行严格“检验”。相较于传统深度学习技术，“生成对抗网络”可以理解为一个设有产品检验标准的自动生产线。“生成器”合成的非真实作品会被送至“鉴别器”，以判断其伪造的逼真程度。若逼真度未达特定标准，作品将被退回“生成器”继续修改。经过无数次的退回与再修正，最终能够合成出“高逼真”的非真实音视频作品。正是生成对抗网络 (GAN) 凭借“生成器网络”与“鉴别器网络”的对抗机制，使合成的非真实音视频作品无限贴近真正版本。

生成对抗网络 (GAN) 在深度伪造领域应用广泛，特别是基于人脸的生成、编辑、替换和重演。例如，深度伪造人脸生成可以分为四类：人脸完全生成技术、人脸属性编辑技术、人脸身份替换技术和人脸面部重演技术，这些不同分类运用了不同的“变体 GAN”，下表罗列了不同分类的“变体 GAN”及其衍生技术类别。

表 2.1 1 深度伪造人脸技术生成分类表

深度伪造人脸技术生成分类	人脸完全生成技术					
	GAN	DCGAN	WGAN	StyleGAN	TediGAN	Collaborative Diffusion
	人脸属性编辑技术					
	IcGAN	StarGAN	AttGAN	STGAN	HifaFace	HFGI
	人脸身份替换技术					
	FSGAN	FaceShifter	InfoSwap	MegaFS	3dSwap	DiffSwap
	人脸面部重演技术					
	Face2Face	A2V	ICface	3dSwap	DAE-Talker	etc...

数据来源：杨宏宇，李星航，胡泽. 深度伪造人脸生成与检测技术综述，华中科技大学学报（自然科学版），安全牛自主绘制

总体来看，生成对抗网络 (GAN) 技术推动了图像和视频生成、处理技术的革新，其提供了一种强大的技术手段，使得伪造内容的生成变得更加真实和高效，但同时也带来了检测和防范的进一步挑战。

2.1.5 扩散模型（DDPM）的原理与应用

DDPM (Denoising Diffusion Probabilistic Models) 是一种生成模型, 在 2015 年由 Sohl-Dickstein 等提出。DDPM 的核心原理基于扩散过程, 这一过程包括两个主要阶段: 前向扩散过程和逆向去噪过程。①前向扩散过程: 这一过程逐渐将高斯噪声添加到图像中, 直到图像最终变成纯噪声。这个过程可以视为一个马尔可夫链, 其中每一步都是在前一步的基础上添加噪声, 噪声的强度由预定义的方差序列控制, 满足逐渐增大的条件。②逆向去噪过程: 这是一个学习的过程, 目标是从纯噪声开始, 逐步去噪, 最终生成一个实际的图像。这个过程同样是一个马尔可夫链, 但是它是参数化的, 需要通过训练神经网络来学习如何从噪声中恢复出图像。

DDPM 的主要作用是生成高保真的图像。通过模拟物理世界中的扩散过程, DDPM 能够从简单的噪声分布生成复杂的数据样本, 其训练过程稳定, 但保真度不及 GAN。此外, DDPM 具有可解释性, 能够更好地理解图像生成过程。DDPM 的这些特性使其在图像生成、多模态图像处理等领域展现出广泛的应用潜力, 例如在多模态图像融合中, DDPM 可以用于组合不同模态下的图像并保留互补信息。尽管 DDPM 在计算成本和实时推理方面存在挑战, 但通过不断的研究和改进, DDPM 在训练效率和推理速度方面仍有提升空间。

2.2 深度伪造和合成媒体技术应用分析

深度伪造和合成媒体技术广泛应用于视频、图像与语音领域中。安全牛通过公开调研和企业访谈发现, 按照视频领域 (换脸、动作转移、表情操纵)、图像领域 (人脸生成、风格转换)、语音领域 (克隆、转换、合成) 等来看, 不同落地领域的占比分别是 37%、56% 和 7%。

2.2.1 视频领域: 换脸、动作转移、表情操纵等

视频篡改和合成是深度伪造的主要应用场景。利用换脸、动作转移、表情操纵等技术, 可以生成以假乱真的视频内容。**视频换脸**是将目标人物的面部嫁接到源视频人物上, 从而创造目标人物做特定动作、说特定言论的假象。如 2018 年初, 一段换上奥巴马面部的乔丹·皮尔视频在网上疯传, 视频中“奥巴马”抨击特朗普, 引发轰动。典型的视频换脸方法有基于 2D 信息和 3D 建模两大类。基于 2D 的方法直接在图像空间进行替换, 如深度伪造利用人脸对齐, 用自动编码器逐帧替换关键面部特征。基于 3D 的方法在三维空间重建人脸, 如 FaceSwap 提取人脸关键点并拟合 3D 模型, 调整姿态和表情, 再融合渲染到目标视频中。**动作转移**指将源人物的动作姿态迁移至目标人物, 使其做出从未有过的动作。如 Aberman 等提出了一种根据骨架驱动的动作迁移方法, 可以将专业舞者的动作映射到普通人身上。动作转移可用于虚拟教学、辅助创作等。**表情操纵**则聚焦于人脸细微的表情变化, 在唇形对齐、眼神迁移等方面对视频进行篡改, 有学者利用 RNN 技术学习人的口型, 再合成目标人物说特定言论的视频。

这些视频篡改技术给取证带来很大挑战。一方面，经过换脸处理的视频具有很强的欺骗性，难以靠肉眼识别。另一方面，视频数据量大、内容复杂，全面检测的难度和成本高。一些涉案视频可能经过多重篡改，既有换脸，又有表情操控，甚至植入虚假背景，给视频取证带来很大挑战。尽管已有诸多尝试，但视频篡改检测依然面临样本少、场景复杂、迭代快等技术挑战。未来深度伪造视频可能更加逼真，更难识别。如何在海量、高分辨率视频中快速、准确检测深度伪造视频仍需要学界和业界的持续攻关。

2.2.2 图像领域：人脸生成、属性编辑、风格转换等

伴随 AI 在计算机视觉领域的进步，近年来出现了一系列逼真的人脸图像合成技术，为深度伪造注入了新动能。一方面，高分辨率的人脸生成模型层出不穷，如 PGGAN、BigBiGAN 等，可生成以假乱真的大尺寸人脸。2022 年，GFPGAN-v1.3 通过修复和超分辨率，生成了接近真人的 1024x1024 人脸图像，在 FFHQ 数据集上以 0.5612 的人类评分优于 StyleGAN3(0.5087)。神经辐射场、隐空间探索等技术进一步放大了人脸生成的潜力空间。

另一方面，人脸属性编辑技术不断突破，可任意操控人脸的性别、年龄、表情等。如 Grigory 等用条件 GAN 改变人脸年龄，Rui 等用 GAN 生成多视角人脸并保持细节，大大拓展了图像伪造的灵活性。基于 3DMM 的参数化人脸建模，则可独立调整人脸形状、反射率、姿势等要素，支持更精细的编辑。一些网红换脸应用，如 ZAO、FaceApp 等，将人脸编辑技术进一步推向大众，引发广泛争议。

在人脸生成和编辑之外，风格转换也是图像合成的重要方向。借助 AdaIN、DualStyleGAN 等模型，可将一幅人脸的风格（如卡通塑形、复古滤镜）迁移到另一幅人脸上，在保持内容纹理的同时实现风格融合。CLIP 引导的扩散模型，更是以文本—图像特征对齐为基础，用语言描述引导人脸的风格变换，极大降低了图像处理的技术门槛。

AI 驱动下，虚假人脸图像变得触手可及，个人照片随时可能被非法篡改，用于换脸色情、诈骗等。这不仅严重侵犯公民肖像权和隐私权，也给社会舆论和司法取证蒙上阴影。对此，学界开展了卓有成效的探索，在生成痕迹分析、内容一致性分析、物理异常分析等维度取得重要突破。尽管取得了可喜进展，但面对深度伪造图像的快速迭代，现有的检测方法还存在诸多不足：一是缺少大规模、高质量的对比数据集，难以全面评估模型性能；二是对未知生成方法的泛化能力不足，容易过拟合已知模型；三是对压缩、调整大小等常见操作的鲁棒性不够，导致实际应用受限。此外，随着生成模型不断进化，一些传统的伪造痕迹可能会逐渐消失，如人脸遮挡、姿态不自然等，需要及时发掘新的识别线索。未来，深度伪造图像检测除了要进一步扩大样本量、优化模型架构，还要重视跨模态分析、主动对抗生成等新思路，多管齐下应对日益复杂的图像篡改威胁。

2.2.3 语音领域：克隆、转换、合成等

语音合成和转换是深度伪造的又一重要应用，其原理是用机器学习模仿特定人的语音，生成其本人并未说过的内容。具体而言，语音伪造主要有以下几种类型：①**语音克隆**指在给定少量目标人物语音的情况下，合成其声音特征。如仅用 5 秒原始语音，就可合成酷似原声的语音。语音克隆可用于虚拟助手、有声读物等，但也可能被滥用。②**语音转换**指在不改变内容的情况下，将语音的个人属性（如音色、语调）转换为目标人物。如 AN 等人用 CycleGAN 实现说话人的声音转换。语音转换可以增强语音的表现力，但也可能用于伪造他人言论。③**语音合成**则是用计算机生成一段原本不存在的语音，其内容完全由输入的文本决定。如 DeepVoice、Tacotron 等系统可根据文本自动合成语音。可见，语音合成最具威胁，完全摆脱了对原始语音的依赖，诈骗人员利用它就能伪造各种假消息。

针对语音伪造，传统的说话人识别技术可能会失效。因为神经网络可以分析人的语音特征，提取音素、基频等关键参数，再重组为与原声高度相似的语音。一些研究者尝试从音质、韵律、情感等方面入手，探索语音真伪鉴别的新思路。如：①**音质分析**。合成语音在频谱、相位等音质指标上往往有异常，可据此设计检测算法。②**韵律分析**。机器难以完美模仿人类语音的节奏、重音等韵律特征，形成其独特的“声纹”。③**情感分析**。人类语音饱含情感，而情感合成是语音伪造的技术瓶颈。有研究者发现，真实语音的情感变化更加自然连贯，由此提出了一种基于情感相似度的检测框架。

总的来看，语音领域的深度伪造发展非常迅速，给司法取证、个人隐私等带来严峻挑战。恶意者可能伪造公众人物讲话，误导舆论；伪造他人声音，实施诈骗；甚至制造虚假证据，干扰司法。这不仅危害个人名誉、财产安全，也可能引发社会动荡。未来语音取证除了要提高检测精度，还要与说话人识别、情感分析等技术深度融合，从语音事件的背景、场景等多方面排查可疑点，以应对不断演进的语音伪造威胁。同时，还要加强对语音的溯源、保护和管理，完善相关法律，筑牢防伪造的制度防线。

2.3 深度伪造和合成媒体安全威胁分析

本章将从个人、社会和国家三个层面全面剖析深度伪造的安全威胁，呼吁各界提高警惕，采取有效举措应对相关威胁。



图 2.3.1 深度伪造和合成媒体安全威胁分析图

2.3.1 对个人层面的安全威胁

深度伪造对个人层面的侵犯，最直观的就是侵犯大众合法权益，使其财产受到损失，影响个人生活，安全牛总结其为三个细化方面：

■ **隐私泄露：个人数据被恶意采集与滥用。**根据第 53 次《中国互联网络发展状况统计报告》显示，截止至 2023 年 12 月，中国网民规模已达 10.9 亿，其中使用过人脸识别功能的用户占比超七成，主要集中在安防、金融、电商、出行等场景。随着人脸识别在身份认证、移动支付等领域的普及，海量人脸数据被存储在互联网各个角落。个人信息的过度采集、非法买卖和滥用，已成为网络时代的顽疾。攻击者利用深度学习技术，对这些易于获取的人脸数据进行训练，生成以假乱真的人脸图片或视频，进而对个人隐私和声誉造成损害。

■ **身份盗用：冒用他人身份实施违法犯罪活动。**随着深度伪造技术的发展，利用该技术伪造他人生物特征、冒用身份实施诈骗等犯罪活动的现象日益增多。安全牛分析发现，身份盗用多集中在色情及诈骗两大场景。一方面，攻击者可通过视频换脸，将受害人的面部“移植”到任意视频中，使其“现身”于色情、政治等场景，常见于公众人物。另一方面，语音合成和变声技术也为网络诈骗提供了新工具。犯罪分子只需搜集公众人物一定数量的语音素材，就能训练出与其极为相似的语音模型，从而伪造身份进行诈骗。仅 2023 年一年，内地就发生多起“AI 换脸诈骗”事件，涉案金额高达数亿元。

■ **声誉损害：伪造淫秽、暴力等内容污蔑个人。**不法分子利用深度伪造技术诋毁、污蔑他人，是近年来网络谣言和网络暴力案件高发的一个重要原因。恶意人员获取一定的图像、视频、音频等数据后，通过 AI 算法，快速创制以假乱真的违法、有害内容，冒用他人身份在网上传播，给被害人个人名誉和形象带来严重损害。不

法分子还利用深度伪造制造名人“讲话”、“采访”等虚假内容，误导公众。

2.3.2 对社会层面的安全威胁

深度伪造对安全的威胁不局限于个人层面，目前，随着技术的快速发展，已逐步影响到社会运转的方方面面。安全牛在本节将重点从虚假信息扩散、舆论操纵、诚信危机三个方面进行社会层面的安全威胁分析。

■ **虚假信息扩散：**伪造内容混淆视听，误导公众。深度伪造技术利用 AI 换脸和语音合成制造真假难辨的视音频内容，通过社交媒体迅速传播，误导公众。例如，今年十一期间，用锋利且刺耳的言辞抨击堵车、调休、游戏等热门话题的“雷军（采用深度伪造技术合成的雷军）骂人”视频即典型案例。又如，伪造钟南山院士声音的音频内容在微信群内大量传播，在疫情期间引发社会恐慌。有研究表明，即使“事实后真相”出现，即信息被证实为假，但负面影响却难以消除。

■ **舆论操控：**恶意引导、分化社会舆论走向。AI 技术的发展使得信息可以被定制化，形成“信息茧房”，这为深度伪造提供了操纵舆论的好机会。一些组织和个人利用这种技术制造符合受众偏见的内容，以操纵和分化舆论。有研究表明，人们更易接受与自己立场相符的虚假信息，伪造内容越迎合受众偏见，就越容易被接受和传播。而“信息茧房”使得大众难以发现自己已是局中棋子，失去自由意志。

■ **诚信危机：**动摇社会诚信基础，加剧不信任感。在人工智能时代，诚信这一社会基石正受到深度伪造技术的严重威胁。曾经作为事实真相铁证的图像和视频，其可信度已大幅下降，公众开始质疑“眼见为实”的原则。个人隐私泄露和声誉受损只是表面现象，它们反映了社会诚信体系的悄然崩溃。在人工智能时代，如何重塑信任成为一个严肃的命题。虽然深度伪造尚未对社会运行造成颠覆性影响，但其潜在风险已经显现。我们必须采取预防措施，及早应对这一挑战。

2.3.3 对国家层面的安全威胁

深度伪造对安全的威胁除个人和社会层面外，对国家层面的潜在安全威胁更令人担忧。安全牛在本节将重点从政治传播、经济挑战和国际合作三个方面进行国家层面的安全威胁分析。

■ **政治传播：**制造虚假政治事件，引发舆论博弈。深度伪造技术可以被用来制造虚假的政治事件，影响政治稳定。例如，通过伪造政治人物的演讲或行为，破坏领导人形象，煽动不满情绪，甚至影响选举结果。在美国，深度伪造技术已经被用于制造虚假的政治宣传，影响公众对政治人物的看法，引发舆论博弈。

■ **经济挑战：**捏造不实信息，影响经济运行。通过发布深度伪造视频，捏造政府行为、编造官员讨论政策调整等手段，对国家经济稳定造成潜在损害。在经济和金融危机时期，利用深度伪造生成的内容放大先前存在的经济担忧和市场的波动，对股票、期货市场带来重大冲击，对国家的经济运行造成负面影响。

■ **国际冲突**：引发国际信任赤字，对和平、安全构成复杂挑战。深度伪造技术可能会被用于国际虚假信息战，破坏国际关系和战略互信，甚至可能引发国际危机和冲突。它降低了国家间的信任，增加冲突风险，如 2019 年印巴紧张时的伪造视频激化了民族主义情绪。深度伪造还可扰乱军事决策，如虚构的核警戒提升信息可能引发其他拥有核武器的国家提高其核力量的警戒等级。此外，恐怖组织也利用深度伪造招募人员，激发针对特定目标的暴力。值得一提的是，深度伪造已成为国际社会日益关注的挑战。

2.4 深度伪造和合成媒体安全检测类技术分析

安全牛研究发现，根据保护策略不同，可以把目前国内深度伪造安全防御技术划分为两大类别：主动式防御（信任类）和被动式防御（检测类）。

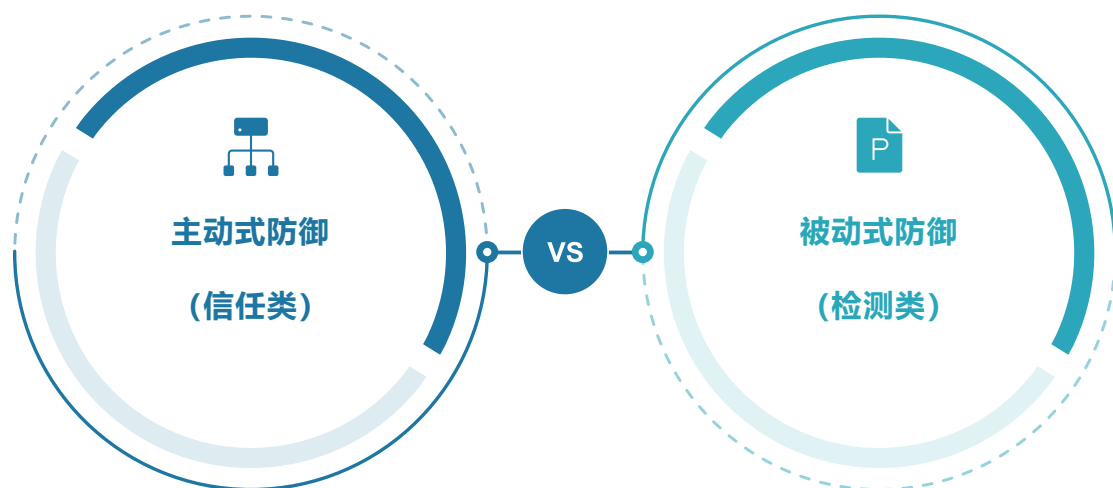


图 2.4 1 深度伪造和合成媒体安全防御技术分类图

主动式防御技术着眼于事前的预防措施。这类技术在视音频内容发布之前，通过嵌入隐蔽的信息，如数字水印或信任认证等，来实现主动的溯源或防止恶意使用。这样的方法旨在主动保护数据内容不被滥用于伪造视频，从而在源头上实现防御；相对地，被动式防御技术主要关注于事后的取证工作，即对已经制作并传播的视频进行分析，以判断其是否包含伪造的视音频内容。这种技术的核心在于识别和验证视频的真实性，通常在视音频发布后进行。本章节重点就主动式防御（信任类）技术进行分析，下一章节就被动式防御（检测类）进行深度分析。

从技术角度来看，要从源头遏制深度伪造，需在内容真实性验证、虚假内容追踪溯源等环节形成有力抓手。目前，业界和学术界已经开展一系列卓有成效的研究，主要集中在数字水印、区块链认证、可信硬件等方面。

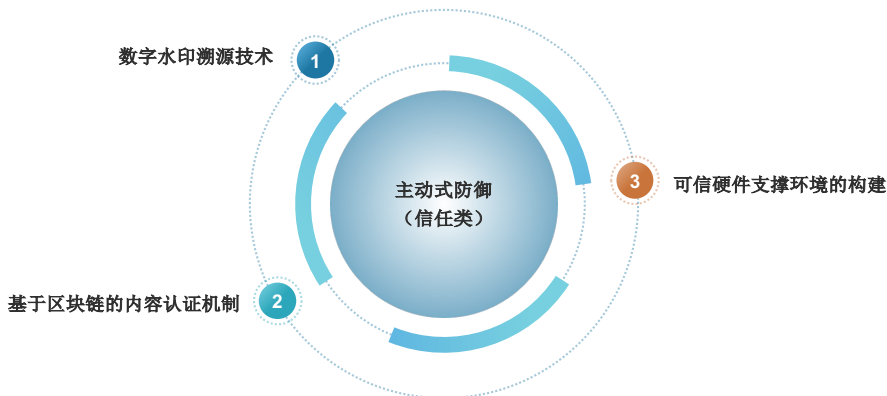


图 2.4 2 深度伪造和合成媒体安全主动式防御（信任类）细分类别图

2.4.1 数字水印溯源技术研究

数字水印是一种隐藏在数字媒体载体中的标识信息，具有很强的鲁棒性和隐蔽性，可有效防止内容被篡改和非法使用。常见的水印方案有基于空域、变换域、压缩域的水印嵌入算法等。**在深度伪造视频和合成媒体中，我们可以在视频制作源头嵌入数字水印，记录视频的原始属性、制作过程等可信信息。一旦发现疑似伪造的视频，可以提取其中的水印，溯源到视频制作主体，进而确定内容真实性和责任归属。**这种源头防伪溯源的思路，不仅可以震慑潜在的造假行为，提高数字资产保护效率，也有利于事后取证和行为人追责。

然而，数字水印技术要真正发挥实效，尚需攻克一些技术难题。首先是如何设计能够抵抗深度伪造的水印方案。由于深度伪造模型的特殊性，传统的鲁棒水印方案往往难以直接适用。一些学者尝试使用对抗训练等方法，生成能抵御深度神经网络的对抗性水印，但在保真度和鲁棒性之间如何平衡仍是一大挑战。其次是如何在嵌入水印的同时，不影响视频内容本身的质量，这对于一些对画质要求较高的场景尤为重要。此外，面对海量视频内容，我们还需研究更高效、更智能的水印检测与提取算法，以实现大规模溯源和快速定位。因此，未来深度伪造数字水印技术的研究，应着眼于水印的对抗鲁棒性增强、视频质量保持、检测效率提升等方向，为实际应用奠定更加坚实的基础。

2.4.2 基于区块链的内容认证机制

区块链以其去中心化、防篡改等特性，为解决互联网内容可信问题提供了新思路。区块链底层由一个个记录数据、互相链接的区块构成，每个区块包含前一个区块的哈希值，从而保证了链上数据的一致性和不可篡改性。将视频、图片等内容的摘要信息上链存证，就可以建立数字内容的第三方公证平台，实现内容来源可追溯和可校验。在此基础上，结合数字签名、哈希指纹等密码学技术，还能构建数字内容版权管理、确权维权的应用生态。这对于遏制深度伪造内容的滥用传播具有重要意义。

目前，国内外已有一些将区块链技术应用于内容认证和版权保护的成功案例。如 Chainalysis 推出了一套基于比特币区块链的数字资产追踪系统，通过在数字内容中嵌入比特币地址，一旦发现资产外流，即可通过该

地址追踪到泄露源。百度超级链则为图片、视频等数字内容提供存证服务，发证方将内容指纹与元数据一同上链，持证方可随时验证内容真实性。阿里巴巴的蚂蚁区块链，现升级为“蚂蚁链”（ANTCHAIN），为 3000 万作品提供数字存证。这些探索表明，区块链有望成为深度伪造防治体系中的重要一环。

不过，在视频防伪领域应用区块链仍面临一些挑战。海量视频数据的存储和同步是一大难题。若全量上链，将给区块链系统带来巨大的存储和计算压力；而仅存储视频指纹，则可能损失一些重要的内容特征，影响事后取证效果。因此，设计兼顾安全、高效的数据上链机制是区块链对抗深度伪造的关键。同时，如何平衡用户隐私保护和内容可追溯性，也考验着相关机制的制度设计。

总的来说，现阶段区块链在视频溯源领域的应用还处于探索阶段，距离大规模落地还需一段时间。未来，随着区块链性能的不断优化以及跨链、侧链等扩展方案的成熟，区块链有望与数字水印等技术深度融合，形成多方协同的深度伪造治理方案。

2.4.3 可信硬件支撑环境的构建

可信硬件环境的构建对深度伪造技术的防范和治理起着至关重要的作用。可信硬件是指通过硬件技术机制确保系统安全及可信，并以此为基础，构建安全可信的操作环境，也称为可信计算环境。它可以为系统提供更高级别的安全保护，有效抵御各种针对软件的攻击，包括深度伪造等新兴的 AI 安全威胁。可信硬件支撑环境的构建有赖于 TPM(Trusted Platform Module)、可信执行环境 TEE(Trusted Execution Environment) 等核心技术。TPM 是一种符合国际可信计算组织规范的安全芯片，能够为系统提供硬件级的密码运算、密钥管理、平台完整性检测等安全功能。TEE 是一种提供隔离执行环境的技术，可以为敏感代码或数据提供安全隔离，保护其不受应用程序或操作系统层面的攻击。

构建可信硬件支撑环境对深度伪造防范的价值主要体现在以下几个方面：首先，可信硬件环境可以对系统关键操作和应用流程提供安全的运行保障。例如，可以将深度伪造检测算法放在可信硬件如 TPM 或 TEE 环境中运行，确保检测过程不被恶意程序所干扰。这大大提高了检测的可靠性。其次，可信硬件环境可以帮助系统建立信任根，提供可信的执行环境认证机制。利用 TPM 等硬件安全模块嵌入到深度伪造检测设备中，可以在启动时检验执行环境的完整性，防止检测软件被篡改。基于此还可以实现远程认证，保障执行环境的可信度。此外，可信硬件支撑的密钥管理能力，可用于构建深度伪造内容溯源机制。除了上述直接作用，构建可信硬件支撑环境还有助于提升相关行业的安全能力，提高抵御深度伪造风险的整体韧性。政府和行业组织应加大支持力度，推动可信硬件支撑环境的建设。

总之，可信硬件支撑环境的构建是应对深度伪造安全威胁的重要举措。通过在硬件层面嵌入安全机制，可以极大提升深度伪造检测的可靠性，强化内容溯源能力，提高整体防护水平。这需要产业链各方通力合作与协同推进。伴随新一代信息技术的发展和网络安全形势的变化，构建可信硬件支撑环境将成为共识，在深度伪造治理和网络空间安全中发挥越来越大的作用。

2.5 深度伪造和合成媒体安全检测类技术分析

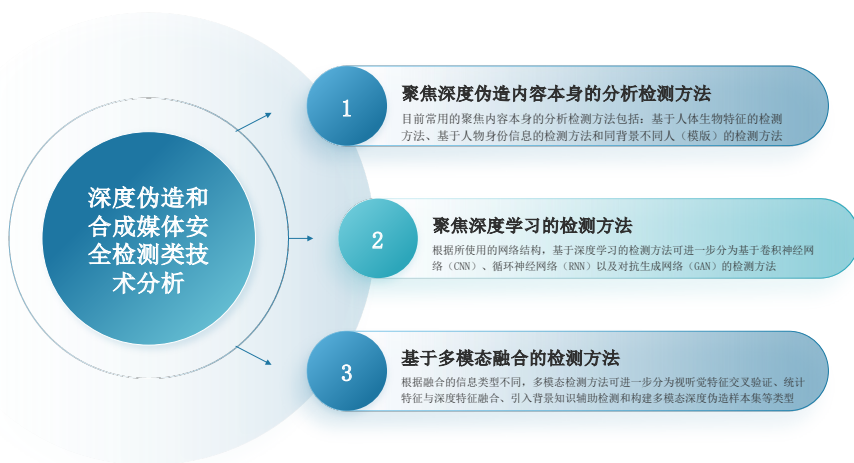


图 2.5 1 深度伪造和合成媒体安全被动式防御（检测类）细分类别图

不同于主动式实施条件的苛刻及主动化要求，被动式检测因其技术实现的可行性和对现有视频内容的广泛适用性，成为当前深度伪造防御技术的主流。为应对深度伪造带来的安全威胁，学术界和产业界开展了一系列深度伪造检测类技术研究，本章将围绕目前主流的深度伪造检测技术的具体方法，即聚焦深度伪造内容本身的分析检测方法、聚焦深度学习的检测方法以及多模态融合的方法，进行展望与发展研究。

2.5.1 聚焦深度伪造内容本身的分析检测方法

基于内容本身的检测方法主要利用深度伪造视频与真实视频在某些物理特征上的差异进行检测。相比于深度学习方法，此类方法具有更好的可解释性，能够指出造假痕迹所在。目前常用的聚焦内容本身的分析检测方法包括：基于人体生物特征的检测方法、基于人物身份信息的检测方法和同背景不同人（模板）的检测方法。

■ **基于生物特征的检测方法。**基于生物特征的检测方法利用伪造人脸与真实人脸在生物特征上的差异进行检测。安全牛根据公开资料总结这些差异包括但不限于眨眼频率、角膜镜面高光、眼神凝视、嘴部运动、人脸与头部姿势的一致性以及心跳节律等。例如，Li 等人通过分析眨眼频率的异常模式来区分真假人脸，而 Hu 等人发现生成对抗网络（GAN）生成的人脸在角膜镜面高光上存在不一致性。Peng 等人则关注于眼神凝视的细微特征，Haliassos 等人利用嘴部的不规则运动来识别伪造视频，Yang 等人通过检测人脸姿势与头部姿势的一致性来辨别真伪，而 Qi 等人通过监控心跳节律的变化来进行检测。这些方法在准确率上表现出色，但它们的泛化能力受到限制，尤其是在面对不断进步的 GAN 技术时，低级图像异常减少，使得检测方法需要关注更深层次的生物特征。

■ **基于人物身份信息的检测方法。**针对有影响力的人物，如国家领导人和明星等，基于身份信息的检测技术在保护重要人物免受深度伪造安全影响方面发挥了重要作用。通过分析视频中人物的身份特征，利用先

进的计算机视觉算法和机器学习模型，提高了伪造人脸视频的检测准确性，如 Agarwal 等利用 OpenFace2 提取面部数据，通过 SVM 区分真假视频。Dong 等专注于人脸中心区域，提出 OuterFace 算法，以提高检测的泛化性能。然而，这些方法需要真实视频对比，不适用于社交媒体的大规模内容检测。为了解决这一问题，Cozzolino 等提出了 ID-Reveal 方法，利用 3D 可变形模型提取人脸特征，生成身份不一致样本，与真实样本一起训练分类网络，仅用真实视频训练即可获得高泛化性的判别器，显著提高了高压视频压缩中面部重现伪造的检测准确度。尽管基于身份信息的检测器在性能和实用性上表现良好，但部分方法因缺乏训练数据而对普通用户有局限性。

■ **同背景不同人（模板）检测方法。**这种攻击方式涉及使用相同的模板对多个人的面部特征进行替换。攻击者可能会保留视频中的人像背景和衣服等素材，只替换面部特征，以制作出看似真实的视频。这种攻击方式的挑战在于，它可能更容易绕过基于算法的图像检测系统，因为除了面部特征外，视频的其他元素都是真实的。当前有企业根据人脸照的纹理、深度信息，以及图像序列的特征来进行聚类 and 分类，将图像分组，并通过比较这些组内和组间的特征差异来判断图像是否为攻击图片（模板），并给予异常评分，对于评分较高的图片系统会自动入模板库，较低评分的模版，会通过人工进行二次确定，决定是否加入模板库。

2.5.2 聚焦深度学习的检测方法

与基于内容本身的检测分析方法相比，基于深度学习的检测方法能够自动从大量数据中学习判别深度伪造的特征，无需依赖先验知识和专家经验。根据所使用的网络结构，基于深度学习的检测方法可进一步分为基于卷积神经网络（CNN）、循环神经网络（RNN）以及对抗生成网络（GAN）的检测方法。**值得一提的是，基于深度学习的检测方法也不是单一使用，通常会和聚焦内容本身的检测方法进行交叉，进而形成复合的检测方法。**

■ **卷积神经网络 (CNN) 检测。**卷积神经网络善于提取图像中的局部特征，已在计算机视觉领域取得了巨大成功。将 CNN 应用于深度伪造检测，就是让网络直接从视频帧图像中学习判别真伪的特征。Rossler 等人发现，基于 Xception 的 CNN 在 FaceForensics++ 数据集上对高质量视频的检测准确率超过 98%，激发了后续研究对 CNN 结构的优化。Nguyen 等人引入胶囊网络提升检测性能，而 Bayar 等人和 Stehouwer 等人分别通过限制卷积滤波器和加入注意力机制来增强模型对篡改痕迹的识别。Cozzolino 等人提出基于自动编码器的零样本学习方法，提高了模型对未知生成方式的泛化能力。Nirkin 等人进一步结合对抗性自动编码器和多任务学习，增强模型的稳健性。尽管单帧 CNN 检测器已经很强大，但在实际应用中，由于深度伪造视频的帧数和分辨率差别很大，如何高效地处理任意长度的视频序列仍是一大挑战。此外，CNN 虽然能够学习到判别性很强的特征，但泛化能力有限，如何提高 CNN 检测器的泛化能力，是未来一个重要的研究方向。

■ **循环神经网络 (RNN) 检测。**相比于 CNN 善于提取空间特征，RNN 则擅长建模时序数据。将 RNN 用于深度伪造视频检测，能够充分利用视频帧之间的关联信息。Güera 等人最早将 CNN 和 RNN 结合用于深度伪造检测。但 RNN 处理长序列的能力有限，在建模全局语义方面不如 Transformer 等新的序列模型。如何在检

测任务中将 RNN 与 Transformer 结合，充分利用局部和全局时序信息，值得进一步探索。

■ **基于 GAN 生成图像指纹检测。**随着生成对抗网络（GAN）技术的进步，生成的图像越来越难以用肉眼区分，但总会留下特定的生成伪影。早期方法通过识别 GAN 模型的伪影来检测，但这种方法在 GAN 架构变化时准确率会降低。Marra 等人发现 GAN 生成的图像会留下类似相机拍摄照片的 PRNU 模式痕迹，这种独特的标记被称为指纹，可以用来进行可靠的伪造人脸检测。Yu 等人进一步分析了 GAN 指纹的内在机制，揭示了其与生成器中 BN 层的权重压缩有关。Wang 等人从 GAN 模型的神经元覆盖角度出发，设计了一种测试 GAN 生成样本的框架，在多个数据集上取得了良好的检测效果。为了提升 GAN 生成方法的检测器的泛化性能，一些研究者提出利用对抗训练和元学习来优化检测器。Wang 等人将对抗样本引入训练，使检测器在应对对抗攻击时鲁棒性更强。

总的来说，基于深度学习的检测方法依托大规模数据集和强大的学习能力，在检测领域取得了瞩目成绩。但这些方法也存在计算开销大、可解释性差、泛化能力不足等局限性。同时，它们大多只关注视频中的视觉信息，忽略了音频、文本等其他模态的信号。

2.5.3 基于多模态融合的检测方法

深度伪造视频通常包含视觉、听觉等多种模态信息，仅依靠单一模态进行检测往往难以全面刻画深度伪造的特征。因此，研究者开始探索多模态融合的检测框架，综合利用不同模态的互补信息，以期获得更加准确和鲁棒的检测效果。根据融合的信息类型不同，多模态检测方法可进一步分为视听觉特征交叉验证、统计特征与深度特征融合、引入背景知识辅助检测和构建多模态深度伪造样本集等类型。

■ **视听觉特征交叉验证。**在人类辨别视频真伪时，视觉和听觉线索都是关键因素。研究者们尝试将唇语与语音的一致性作为深度伪造检测的依据。Korshunov 等人首次探索了视听觉特征在深度伪造检测中的应用，他们提取嘴唇区域的视觉特征和语音的 MFCC 特征，使用独立检测器建模并加权平均输出，发现融合后性能优于单一模态。Mittal 等人提出基于注意力机制的多流网络，使网络在编码唇语和语音特征时能够相互参考，聚焦于关键区域。然而，现有方法主要关注假唇音不同步的视频，对于换脸、换头等无需篡改音频的伪造方式，检测效果有限。Chugh 等人引入预训练的说话人识别和人脸识别模型，通过比较输出差异判断视听觉身份一致性，实验表明该方法在多个数据集上表现良好。尽管如此，利用视听觉信息进行深度伪造检测仍面临挑战。高质量的视听觉深度伪造数据集如 DFDC 在音视频质量和篡改方式多样性方面有待提高，限制了检测器的泛化能力。此外，随着语音克隆等技术的发展，视听觉身份不一致的案例将减少。未来的研究需要挖掘面部细微表情与情绪语气的协同性，并探索更深层次的模态融合方法，以充分发挥多模态信息的协同作用。

■ **统计特征与深度特征融合。**前文提到，基于统计特征的方法和基于深度学习的方法在检测深度伪造方面各有优势。一些研究者尝试将两类方法的优势结合起来，以期获得更好的检测效果。统计特征与深度特征的融合思路为深度伪造检测提供了新的研究视角。相比于端到端的深度学习方法，融合方法具有更好的可解释性，

能够分析统计特征和深度特征在检测中的作用。同时，适当引入统计特征有助于提高深度学习模型的鲁棒性，尤其在应对对抗攻击时。传统的对抗攻击大多针对神经网络的梯度信息，而统计特征则不受其影响。Wang 等人发现，在融合统计特征后，深度学习检测器面对对抗样本时的性能显著提升。未来可以探索将统计建模与对抗训练相结合，以提高检测器的安全性。需要指出的是，现有融合方法所用的统计特征仍较为简单，大多为像素共生矩阵、Markov 转移概率等二阶统计量。一些更高阶的统计特征，如局部像素相关性、谱相关性等，在图像篡改检测中已经显示出优势，但尚未被用于深度伪造检测。如何从理论和实践角度探索新的统计特征，并将其与不同深度学习模型相融合，可能成为一个有前景的方向。

■ **引入背景知识辅助检测。**结合外部知识如事实和常识，可以提升深度伪造检测的准确性。Wu 等人提出了一种多模态检测框架，分析视觉内容和语音转录文本，通过知识库比对捕捉视觉与语义的错误关联，有效识别换脸视频。Wang 等人进一步将检测器与事实验证系统结合，使用 Transformer 编码器和交叉注意力机制学习视觉与文本的关联性，通过对比学习提升检测准确率。然而，该领域仍面临挑战，包括知识库规模小、隐式语义信息处理困难、高质量标注数据获取难等问题。未来研究需探索弱监督学习、自监督对比学习等方法，优化多模态融合框架，实现更准确的篡改检测与定位。随着多模态数据的增长和新学习范式的出现，深度伪造检测技术有望进一步发展，为构建智能、安全的信息社会提供支持。

■ **构建多模态深度伪造样本集。**为了支撑单模态及多模态伪造检测技术，有厂商提出了多模态深度伪造样本集构建，即互联网爬取及 DPI 还原、预训练伪造模型自动生成伪造样本和半自动标注技术。互联网爬取及 DPI 还原主要指通过互联网爬虫技术，可以从公开的网络平台上收集大量的图像和视频数据。这些数据可以包括正常的人脸视频以及已知的深度伪造内容。此外，还可以通过 DPI（深度包检测）技术还原网络流量中的数据包，以获取更多的深度伪造样本。预训练伪造模型自动生成伪造样本指利用已有的生成对抗网络（GAN）或其他深度学习模型，可以自动生成大量的伪造样本。这些模型通常在大规模的真实数据集上进行预训练，然后通过调整参数来生成伪造内容，如人脸替换或表情操纵等。半自动标注技术指为了提高标注效率和准确性，首先使用预训练的模型对样本进行初步标注，然后由人工审核和校正这些标注结果。这种方法可以大幅减少人工标注的工作量，并提高标注的一致性和准确性。通过上述方法，可以构建一个包含多种模态（如视频、音频等）的深度伪造样本集，这些样本集不仅能够支持单模态伪造检测技术的研究，还能够促进多模态融合检测技术的发展。这样的样本集对于训练和评估深度伪造检测模型至关重要，有助于提高模型的泛化能力和检测准确性。

第三章 AI 深度伪造和合成媒体安全的市场现状分析

3.1 深度伪造和合成媒体安全的市场表现分析

本章安全牛将从深度伪造和合成媒体安全的市场规模、公开中标信息及厂商投融资三方面分析其具体市场表现。

3.1.1 深度伪造和合成媒体市场规模分析

根据安全牛公开调研和企业访谈收集深度伪造和合成媒体安全市场相关信息显示，2023 年我国深度伪造和合成媒体安全市场规模 1.19 亿元，预计 2024 年同比增长 27%，达到 4.08 亿元。其中，监管方如各部委、公检法、网信办是深度伪造和合成媒体安全市场的最主要采购方，他们的采购占比超过 50%。其次为金融、银行和运营商。从数据来看，深度伪造和合成媒体目前正处于发展早期阶段，呈现出同比增幅较大，受政务、大客户采购的影响较大的现实趋势。

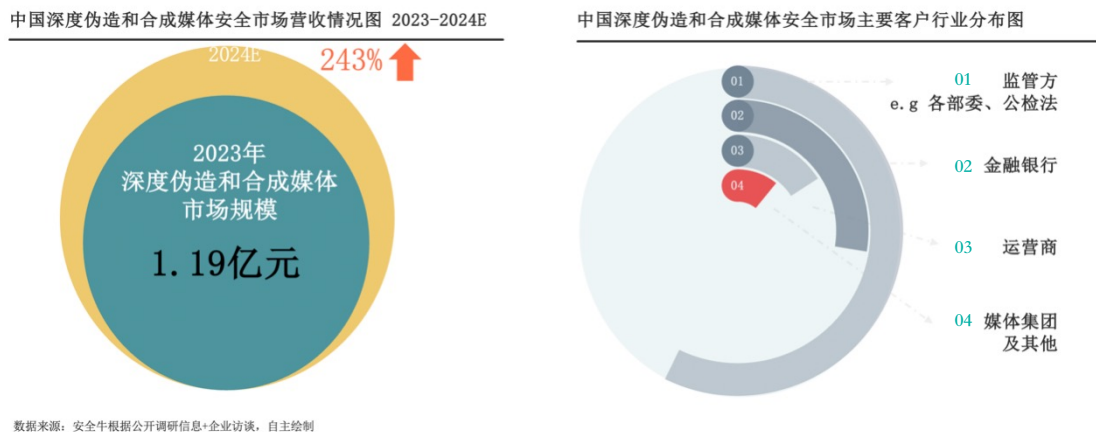


图 3.1.1 中国深度伪造和合成媒体市场营收及主要客户数据分析图

3.1.2 深度伪造和合成媒体公开中标信息分析

安全牛根据公开信息收集 2019 年至今（2024 年 10 月）的深度伪造和合成媒体公开中标标的信息显示，自 2019 年至今，围绕数字内容深度伪造检测与防范关键技术验证进行采购的中标信息总共为 126 件，其中，2019 年 2 件，2020 年和 2021 年各 16 件，2022 年 29 件，2023 年 32 件，2024 年（1-10 月）30 件。从中标金额来看，2019 年中标金额仅有 183 万元，2020 年 865 万元，2021 年 1171 万元，2022 年 2873 万元，2023 年 4713 万元，2024 年仅截至 10 月份，已同比增长达到 345%，为 2.1 亿元。

虽然 2024 年数据“突飞猛进”，但从深度伪造和合成媒体中标的整体情况来看，在数字内容深度伪造检测与防范领域，商业化应用和盈利模式的探索仍在断深入，且有巨大潜力。目前，该领域的主要投资方集中在监管方，而部委的采购数据通常不披露。金融、银行因为具体业务具有较强的采购意愿；媒体集团出于提升内容生产效率、丰富表现形式和增强用户互动的需求，对深度伪造技术的应用和防范投入较多资源；公安部门则更侧重于防范深度伪造技术可能带来的安全威胁，如虚假信息的传播、网络诈骗等，因此在技术防范投入较大。未来，随着技术的进步和应用场景的拓展，预计会有更多行业和领域开始关注并投资于这一技术，推动其商业化进程。

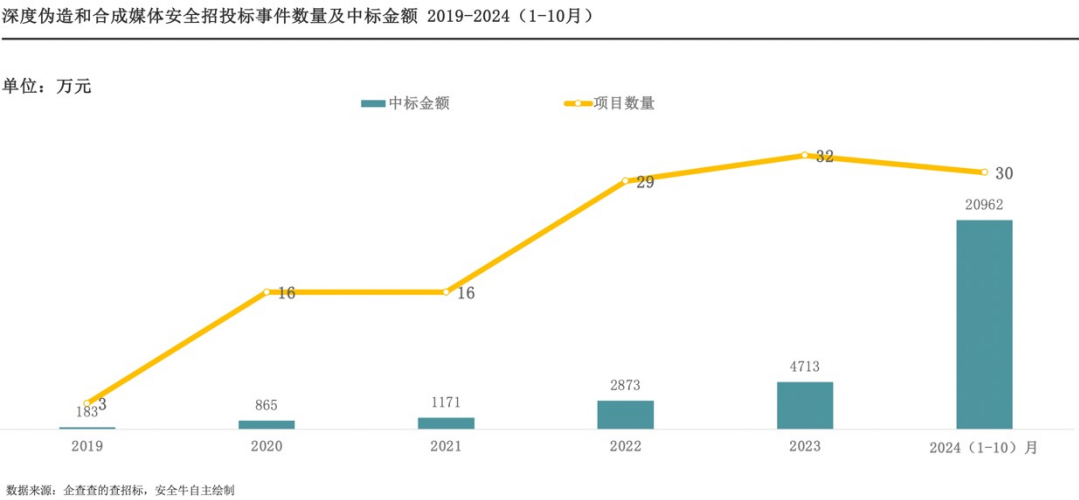


图 3.1.2 中国深度伪造和合成媒体安全招投标事件数量和中标金额分析图

3.1.3 深度伪造和合成媒体厂商投融资情况分析

安全牛根据公开信息收集 2018 年至今（2024 年 11 月）的深度伪造和合成媒体主流厂商投融资信息显示，整体来看，2020 年是资方看好深度伪造和合成媒体创新厂商的主要时机，在这一年瑞莱智慧、深信科创获得天使轮，中关村科金获得股权投资。2021 年，瑞莱智慧获得大额 3 亿元融资，投资方为蚂蚁金服、达泰资本等。2024 年，获得投融资的两家企业为中科睿鉴和瑞莱智慧，两家企业均为国家队成员。

表 3.1.1 2018.01-2024.11 深度伪造和合成媒体主流厂商投融资汇总表

企业名称	轮次	融资时间	融资金额	投资方	地区
中科睿鉴	股权投资	2024/6/19	未披露	达晨财智	北京市
瑞莱智慧	战略投资	2024/4/12	未披露	顺禧基金，智谱 AI 等	北京市

瑞莱智慧	股权投资	2022/8/22	未披露	梦景投资, 卓源资本	北京市
深信科创	Pre-A+ 轮	2022/5/24	数千万人民币	北斗资本领投	合肥市
高维数据	Pre-A 轮	2021/12/17	未披露	安恒信息	合肥市
深信科创	Pre-A 轮	2021/10/18	数千万人民币	国汽投资、将门创投领投	合肥市
瑞莱智慧	A 轮	2021/9/27	3 亿人民币	蚂蚁金服, 达泰资本, 中国互联网投资基金等	北京市
中科睿鉴	股权投资	2021/5/12	未披露	深圳柯源	北京市
瑞莱智慧	Pre-A 轮	2021/2/4	数千万美元	松禾资本, 前海母基金, 达泰资本, 基石资本等	北京市
中关村科金	股权投资	2020/8/7	未披露	沅华资本	北京市
深信科创	天使轮	2020/8/4	1000 万人民币	度岩资本领投	合肥市
瑞莱智慧	天使 + 轮	2020/3/4	数千万人民币	松禾资本领投, 同渡资本跟投	北京市
瑞莱智慧	天使轮	2019/1/4	数千万人民币	百度风投、中科创星领投, 水木清华校友种子基金、卓源亚洲等跟投	北京市
红阵网络	股权投资	2018/10/24	未披露	安恒信息	南京市

数据来源：安全牛根据公开信息自主整理，自主绘制

从深度伪造和合成媒体投资方视角来看，创投 / 基金是主要投资方，也有传统网络安全厂商进行战略布局。从获得投融资的厂商所在地区来看，主要集中在北京和合肥，此两地具有较好的政策落地优势。

从整体数据来看，深度伪造投融资市场还有较大增长空间，未探索出较好的盈利模式是其未能“茁壮成长”的主要原因。生成式 AI 的前景广阔，但作为其细分的深度伪造和合成媒体的未来更需要资方关注。

3.2 深度伪造和合成媒体安全生态图谱分析

安全牛根据公开信息和企业访谈，自主绘制下图 (3.2-1 深度伪造和合成媒体安全的生态图谱)。安全牛认为，深度伪造技术领域的重度参与者可以按照研究侧和服务侧两部分内容进行划分。

AI 深度伪造和合成媒体安全的市场现状分析



图 3.2.1 深度伪造和合成媒体安全的生态图谱

研究侧包括拥有深度伪造专利认证的金融机构、科研院所和高校。根据安全牛收集的公开信息显示，**金融机构**获得深度伪造检测认证的企业有阿里巴巴、支付宝和马上消费金融股份有限公司，他们为探索深度伪造技术在金融安全和用户体验方面的应用正做着研究贡献；但从整体研究侧来看，科研院所和高校这些机构在深度伪造的基础研究、技术开发和应用探索方面发挥着关键作用。

科研院所中深度伪造专利获得者包括但不限于：公安部第三研究所、中国科学院计算技术研究所、中国科学院信息工程研究所、中国电子科技集团公司第三十研究所、中国科学院自动化研究所等。

高校则是培养未来技术人才和进行前沿研究的重要基地，南昌大学、暨南大学、齐鲁工业大学（山东省科学院）、浙江大学、南昌大学、中国科学技术大学、大连理工大学、武汉大学、西安交通大学、中山大学、西安电子科技大学等大学均获得深度伪造相关专利证明。

服务侧则涵盖了提供深度伪造和合成媒体安全服务的企业，分为以网络和数据安全为主营业务的企业、以人工智能为核心的企业和满足自我检测需求的互联网企业。以网络和数据安全为主营业务的企业如国投智能、亚信安全、东方通、顶象科技、高维数据、浩瀚深度等，他们不仅可以提供深度伪造检测和防护服务，还在网络安全和数据安全方向具有较大优势。

以人工智能为核心的专精特新企业，如“人工智能四小龙”的商汤科技、旷视科技、云从科技、依图科技等，他们在深度伪造方向更加聚焦人脸识别；又如瑞莱智慧、中科睿鉴和中关村科金等，利用 AI 技术来识别和防范深度伪造内容，推动技术的正向发展。

满足自我检测需求的互联网企业，包括央视频、爱奇艺、哔哩哔哩等，这些企业在内容审核、版权保护和用户安全等方面有强烈需求。它们通过内部研发或与专业服务提供商合作，确保平台内容的真实性和安全性。

随着技术的不断进步，这些参与者在推动技术发展的同时，也在积极构建一个更加安全、可靠的数字环境。

3.3 深度伪造和合成媒体安全市场成熟度分析

根据安全牛调研发现，当前我国深度伪造和合成媒体安全市场正处于发展前期阶段，安全牛从顶层设计、中层构建、基层应用三个层面来具体分析：

■ **顶层设计：**我国正在逐步建立和完善相关法律法规，以应对深度伪造技术带来的挑战。2019 年，国内印发了关于《网络音视频信息服务管理规定》、《网络信息内容生态治理规定》和《具有舆论属性或社会动员能力的互联网信息服务安全评估规定》三项规定，标志着我国开始从立法层面对深度伪造技术进行规范和治理。虽然国内在立法方面在不断进行探索，但与一些发达国家相比，国内在深度伪造的法律规制方面还有较大的完善空间。

■ **中层构建：**在中层构建方面，学术界和产业界正在积极开展相关研究和技术开发。如在学术研究方面，中国的高校和研究机构正在积极开展深度伪造检测技术的研究，专利申请数量与日俱增。如从基于神经网络架构、基于时序特征等多个维度对深度伪造检测技术进行深入分析，并提出未来研究方向及建议。从深度伪造和合成媒体安全服务提供商方面来看，中国的市场还相对较新，许多企业刚开始涉足深度伪造检测服务领域。随着技术的发展和市场需求的增加，预计会有更多的服务提供商进入这一市场。目前，这些服务提供商正在探索有效的商业模式和技术解决方案，以满足市场对深度伪造检测的需求。但整体而言，中层构建还处于起步阶段，需要更多的时间和资源来逐步走向成熟。

■ **基层应用：**安全牛发起的《大众对深度伪造和合成媒体技术安全认知、态度及影响调研问卷》调研数据显示，公众对深度伪造技术的认识和警觉性正在逐步提高，但仍需要开展公民安全意识教育，增强公众参与度，提高全社会对深度伪造技术的识别和应对能力。

● AI 深度伪造和合成媒体安全的市场现状分析



图 3.3 1 深度伪造和合成媒体安全的市场发展阶段图

综上所述，中国的深度伪造和合成媒体安全市场正处于发展前期，随着立法的进一步完善、产学研合作的逐渐深入以及服务提供商的越来越成熟，预计 2027 年可以进入蓬勃发展阶段。

3.4 深度伪造和合成媒体安全消费者敏感度分析

3.4.1 深度伪造和合成媒体安全消费者认知度分析

根据安全牛于 2024 年 9-10 月发起的《大众对深度伪造和合成媒体技术安全认知、态度及影响调研问卷》回收数据显示，31% 的被调研者知道深度伪造和合成媒体技术，60% 的被调研者听说过但不是太了解，只有 9% 的被调研者不知道深度伪造和合成媒体技术。

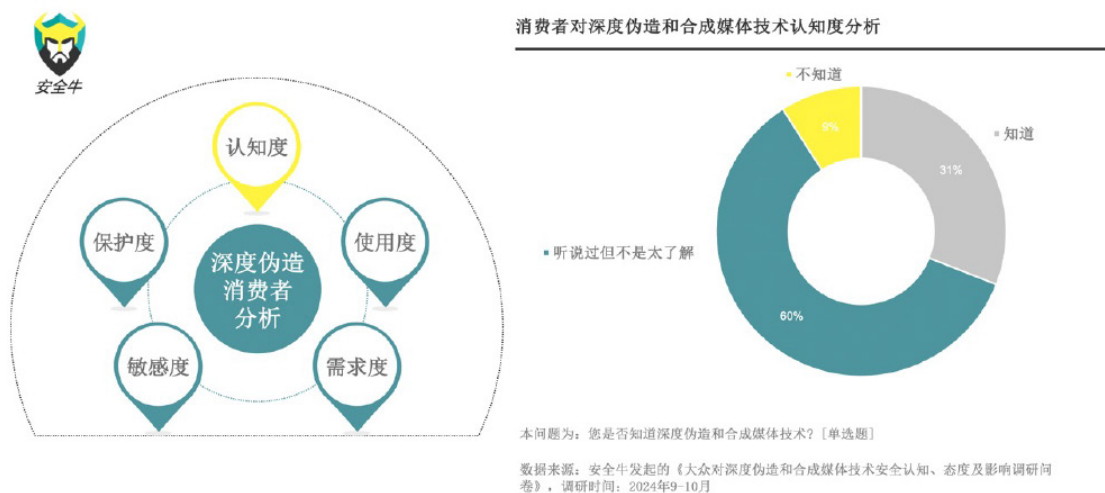


图 3.4 1 安全牛发起的消费者深度伪造和合成媒体技术认知度分析图

根据安全牛于 2024 年 9-10 月发起的《大众对深度伪造和合成媒体技术安全认知、态度及影响调研问卷》

回收数据显示，消费者对深度伪造和合成媒体技术制造的虚假内容整体认知度不高，49% 的消费者认为自己很难辨别虚假内容（但绝大部分对如何识别感兴趣），48% 的消费者认为可以辨别出部分虚假内容，只有 3% 的消费者认为能够完全识别不同类型的深度伪造内容。



安全牛认为，从消费者对深度伪造和合成媒体技术的认知度来看，大众普遍对这一概念及其典型案例有所了解，但在识别利用该技术制造的虚假内容方面存在明显困难。近半数的消费者表示，他们难以识别出深度伪造和合成媒体技术生成的虚假内容，超过四成的消费者认为自己只能识别出一部分。这种情况与该技术的“三高”特点——高仿真度、高自动化、高隐蔽性紧密相关。

3.4.2 深度伪造和合成媒体安全消费者使用度分析

根据安全牛于 2024 年 9-10 月发起的《大众对深度伪造和合成媒体技术安全认知、态度及影响调研问卷》回收数据显示，被问及消费者是否使用或者尝试使用深度伪造和合成媒体技术时，21% 的消费者进行过尝试，但值得一提的是，在尝试过的消费者中有 1/3 的消费者表示以后不想继续尝试。在 79% 的没有尝试过此技术的消费者中，有超半数有意愿在未来进行尝试，有近 1/3 表示不知道尝试的方法 / 途径。

AI 深度伪造和合成媒体安全的市场现状分析



图 3.4.3 安全牛发起的消费者深度伪造和合成媒体技术使用度分析图

根据安全牛于 2024 年 9-10 月发起的《大众对深度伪造和合成媒体技术安全认知、态度及影响调研问卷》回收数据显示，消费者被问及深度伪造和合成媒体技术在哪些消极应用场景中最为常见时，87% 的受访者认为深度伪造技术被用于制造和传播虚假新闻，78% 的人选择此技术被用于网络诈骗和身份欺诈。此外，制造和传播不雅内容、侵犯个人隐私和肖像权以及政治操纵和舆论引导也是公众认为的主要消极场景，分别占到了 72%、63% 和 40%。



图 3.4.4 安全牛发起的消费者认为深度伪造和合成媒体技术消极场景使用度之分析图

安全牛认为，从消费者对深度伪造和合成媒体技术的使用度来看，大众普遍对其存好奇心，但同时也对此技术可能引发的诚信问题抱有普遍的担忧。特别是对信息的真实性和个人隐私的保护，大众显得尤为关注。高比例的制造虚假新闻和网络诈骗、身份欺诈、传播不雅内容和个人隐私权的消极场景进一步凸显了这种担忧。

3.4.3 深度伪造和合成媒体安全消费者需求度分析

根据安全牛于 2024 年 9-10 月发起的《大众对深度伪造和合成媒体技术安全认知、态度及影响调研问卷》回收数据显示，消费者被问及深度伪造和合成媒体应用领域时，影视及视频制作、虚拟现实和游戏是排名前三的领域，占比分别为 88%，77% 和 52%。对于较为创新的方面，如缅怀故人、历史重现和创新旅游，选择的比例分别为 33%，31% 和 28%。医疗救助和法律取证的选取比例最低，分别为 12% 和 7%。

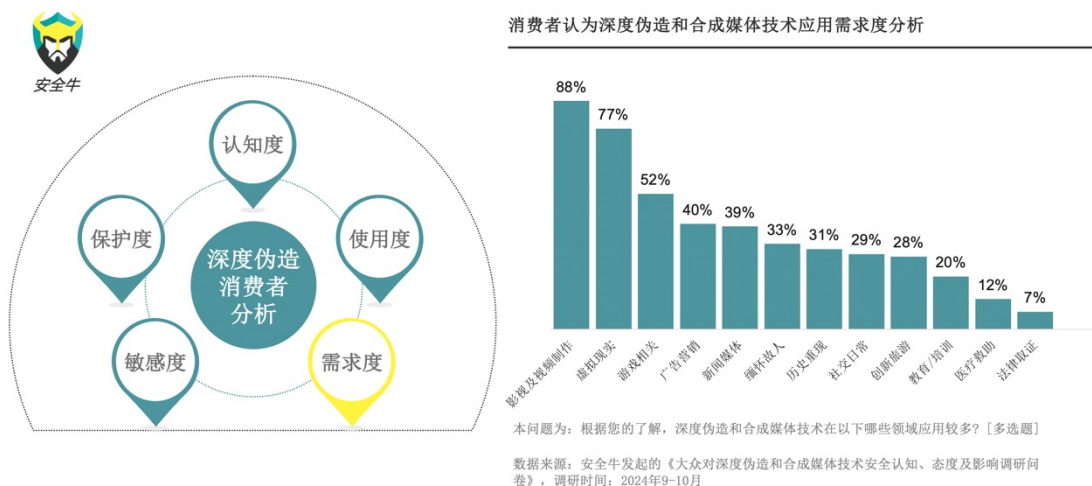


图 3.4.5 安全牛发起的消费者认为深度伪造和合成媒体技术应用需求度分析图

根据安全牛于 2024 年 9-10 月发起的《大众对深度伪造和合成媒体技术安全认知、态度及影响调研问卷》回收数据显示，被问及大众对深度伪造和合成媒体安全保障需求度时，大众普遍认为加强法律制裁、加大平台审核力度和加强大众意识教育是三个最为主要的保障手段，占比分别达到 86%，82% 和 72%。另外，建立举报和预警平台、以技术规制技术和提供免费的检测工具占比也都突破了 50%，分别为 66%，63% 和 60%。

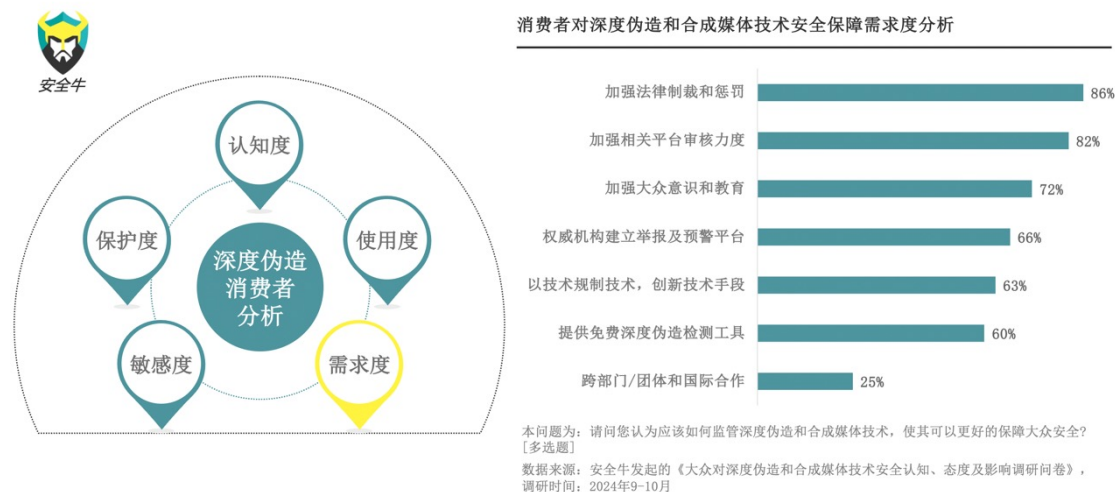


图 3.4.6 安全牛发起的消费者认为深度伪造和合成媒体技术安全保障需求度分析图

● AI 深度伪造和合成媒体安全的市场现状分析

安全牛认为，从消费者对深度伪造和合成媒体技术的需求度来看，大家对于安全需求保障的意愿还是很强烈的，特别是对监管侧和技术提供商寄予厚望。

3.4.4 深度伪造和合成媒体安全消费者敏感度分析

根据安全牛于 2024 年 9-10 月发起的《大众对深度伪造和合成媒体技术安全认知、态度及影响调研问卷》回收数据显示，在被问及是否担心深度伪造和合成媒体技术对个人安全产生不好影响的时候，16% 的受访者表示非常担心，73% 的受访者表示有些担心，只有 11% 的受访者表示不太担心或者完全不担心。

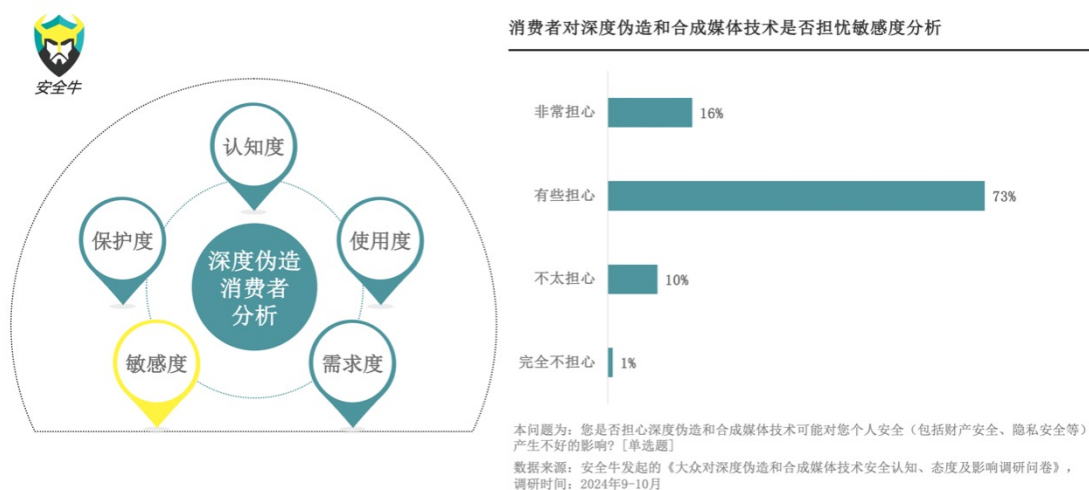


图 3.4.7 安全牛发起的消费者深度伪造和合成媒体技术敏感度分析图

根据安全牛于 2024 年 9-10 月发起的《大众对深度伪造和合成媒体技术安全认知、态度及影响调研问卷》回收数据显示，当被问及如果对深度伪造和合成媒体技术安全现状进行打分，满分 100 分，受访者打出的平均分是 58 分，45% 的受访者打出的分数低于 50 分，21% 的受访者打出 80 分以上的分数。

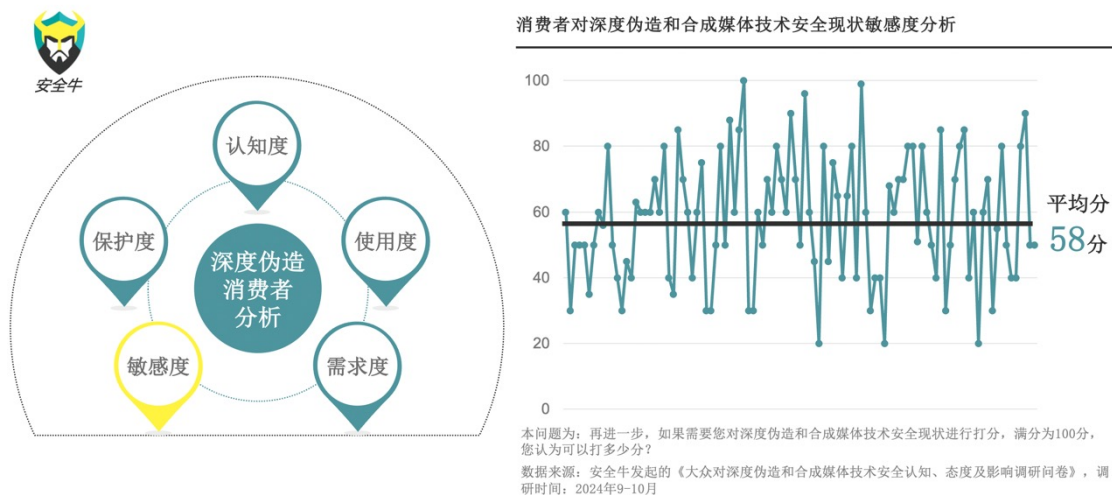


图 3.4.8 安全牛发起的消费者深度伪造和合成媒体技术安全现状敏感度分析图

安全牛认为，从消费者对深度伪造和合成媒体技术的敏感度来看，89% 的受访者对当前深度伪造和合成媒体技术对个人安全产生影响表示担忧，有 45% 的受访者对当前安全现状打出低于 50 分的分数，这表明大众普遍认识到这些技术可能带来的风险，包括但不限于隐私侵犯、名誉损害、信息误导、财产损失等。

3.4.5 深度伪造和合成媒体安全消费者保护度分析

根据安全牛于 2024 年 9-10 月发起的《大众对深度伪造和合成媒体技术安全认知、态度及影响调研问卷》回收数据显示，当被问及对深度伪造和合成媒体的总体看法时，22% 的被调研者表示积极的态度，他们期望看到更多的落地应用；15% 的被调研者表示消极态度，他们期望技术可以受到管控甚至禁止；54% 的受访者持中立态度，他们认为可以应用，但需要受到规范或者管控；9% 的受访者表示不清楚，需要先行观望。



图 3.4.9 安全牛发起的消费者深度伪造和合成媒体技术整体看法保护度分析图

安全牛于 2024 年 9-10 月发起的《大众对深度伪造和合成媒体技术安全认知、态度及影响调研问卷》回收数据显示，当被问及受访者是否采取了相应措施保护自己免受深度伪造技术带来的消极影响时，82% 的受访者表示加强了个人隐私保护，如限制个人信息公开；70% 的受访者表示了解了相关防范知识；56% 的消费者表示向可信平台举报可疑内容；45% 的受访者表示使用深度伪造检测工具，13% 的受访者表示并没有采取任何措施。

AI 深度伪造和合成媒体安全的市场现状分析



图 3.4 10 安全牛发起的消费者深度伪造和合成媒体技术安全保护度分析图

安全牛认为，从消费者对深度伪造和合成媒体技术的保护度来看，超半数受访者对深度伪造和合成媒体持中立态度，认为其需要在监管和规范下进行应用；另外受访者普遍在保护自身意识方面有较高的保护意识，大家也有意愿了解相关防范知识，并期待深度伪造检测工具的进一步落地。

第四章 AI 深度伪造和合成媒体安全市场技术落地分析

4.1 深度伪造和合成媒体安全能力建设分析

《技术与文明》一书阐释道：技术可以在很大程度上左右人类这一物种的生物属性，但反过来人类也可以赋予技术以灵魂。深度伪造和合成媒体安全能力建设是个全生态问题，需要监管侧、研究侧和服务侧的共同配合。通过多方合作，可以更好地促进深度伪造技术的健康发展，同时保护社会大众的利益。

■ **监管侧**—行政与立法机构：深度伪造和合成媒体技术的快速发展引发监管的必要性。多个监管机构如工信部、网信办等，负责制定相关法律法规，确保技术的使用不违反规定，并保护公众的权益。此外，公安、法院和检察院在处理相关案件时也发挥着重要作用。

■ **研究侧**—大学高校与科研院所：这些机构在深度伪造和合成媒体技术的研究中扮演着重要角色。大学不仅负责培养相关领域的人才，还进行基础研究和应用研究。科研院所则专注于技术的开发和标准的制定，确保深度伪造技术的安全和合规使用。

■ **服务侧**—专注 AI 与安全服务的提供商：在深度伪造和合成媒体技术安全应对中，服务提供商的角色不可忽视。专注于 AI 的公司致力于用 AI 的方式推动技术的发展，而专注于安全的公司则关注如何防范深度伪造带来的风险。这些服务提供商通过提供技术解决方案和安全服务，帮助公众应对深度伪造技术带来的挑战。

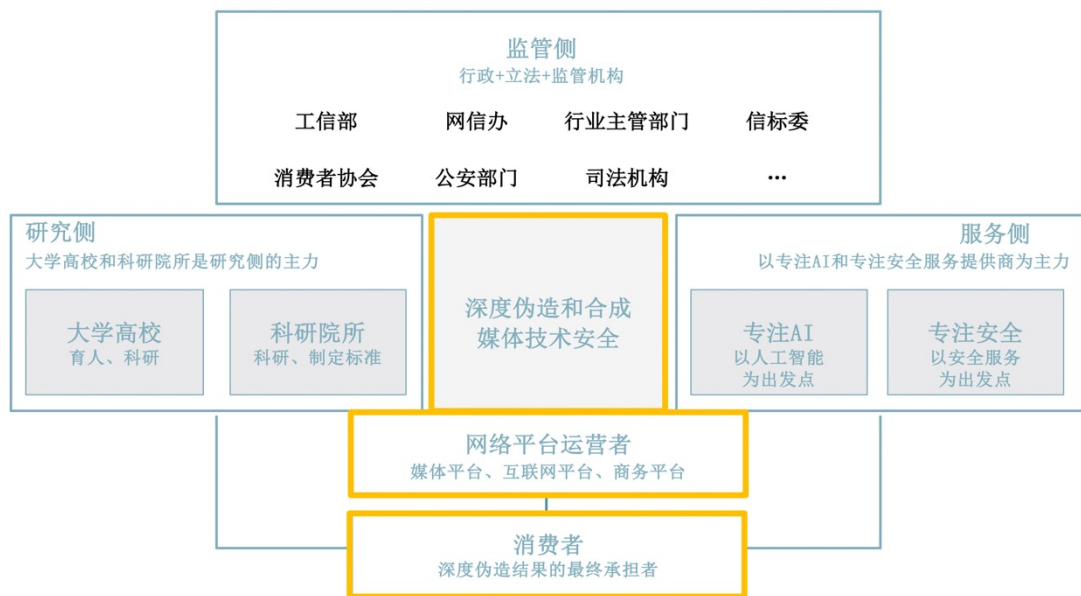


图 4.1.1 中国深度伪造和合成媒体安全能力建设全景图

AI 深度伪造和合成媒体安全市场技术落地分析

随着技术的不断进步,我国的相关法律法规和安全措施也在不断完善(详见第一章第五小节解读具体政策),以应对潜在的风险和挑战。安全牛认为,盘点当前我国深度伪造和合成媒体安全能力建设,重点需要从研发侧和服务侧分别进行能力分析。

4.1.1 国内深度伪造和合成媒体安全研究侧能力分析

安全牛根据公开信息收集 2021 年至 2024 年 1-10 月的深度伪造和合成媒体专利信息得出以下主要结论:

■ 专利申请人主要是高校,占比达到 57%,其次是企业和科研院所,分别为 27% 和 16%。深度伪造专利法律状态为有效的 33 家申请者中,高校占比 58%,企业占比 30%,科研院所占比 12%。

■ 专利申请主要集中在 2022 年和 2023 年,深度伪造和合成媒体的研究热度在这段时间较为明显。

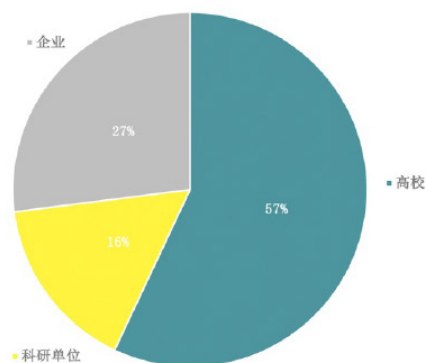
■ 专利的法律状态主要集中在等待(实审提案、年登印费),占比达到 71%,真正获得专利权有效的申请者占比只有 22%。

■ 检测类型是深度伪造发明专利的主要类型,占比达到 77%,鉴别、取证、溯源类占比较低,分别不足 6%。

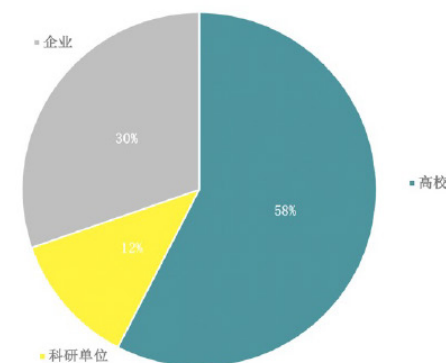
下面就具体数据做详细分析,另外,本报告附录部分列举了翔实数据供参阅。

根据安全牛公开调研收集的深度伪造专利相关数据,从深度伪造专利申请机构的类型分布来看,高校占比最多,达到 57%,其次是企业,占比为 27%,科研单位的申请比例最低,为 16%。从法律状态为专利权有效的申请类型分布来看,也呈现相同的趋势,高校:企业:科研单位占比可以简化为 6:3:1。

深度伪造专利申请机构类型分布图



法律状态为专利权有效的深度伪造申请机构类型分布图



数据来源:安全牛根据公开调研信息自主绘制

图 4.1.2 中国深度伪造专利申请机构类型分布图

根据安全牛公开调研收集的深度伪造专利相关数据，2021 年，深度伪造专利申请件数为 20 件，2022 年为 44 件，2023 年达到 59 件，2024 年仅上半年就达到 27 件，专利申请数量呈现明显的上升趋势。特别是 2022 年和 2023 年，申请数量同比增长较快，说明深度伪造和合成媒体的研究热度在这段时间较为明显。另外，从当前数据来看，每年下半年的申请量普遍高于上半年，预计 2024 年有望突破 80 件。

深度伪造专利申请数量年度分布图 2021-2024上半年

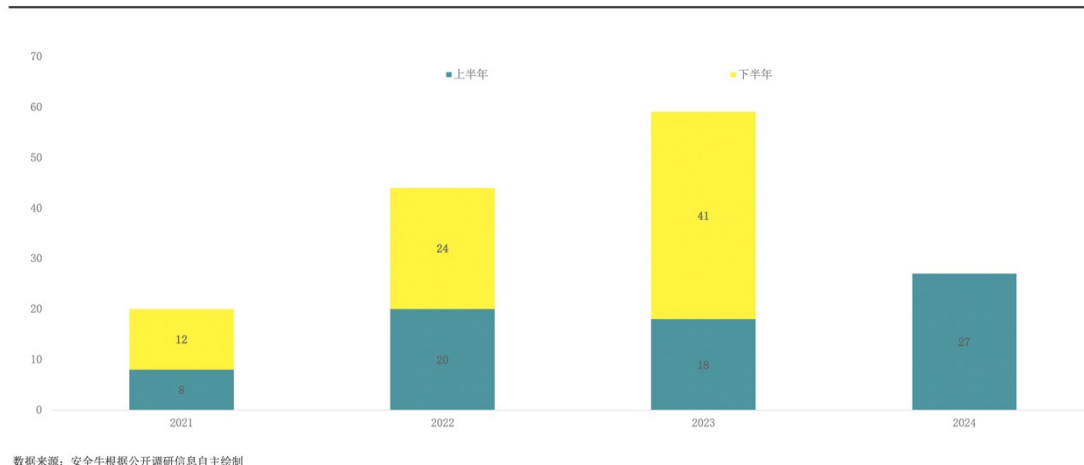


图 4.1.3 中国深度伪造专利申请数量年度分布图

安全牛公开调研收集的深度伪造专利相关数据显示，从深度伪造专利申请案件状态来看，等待实审提案的案件数量最多，占比达到 67%，专利权有效的仅占 22%，另外，还有多种因各种原因未处于专利权有效状态的案件，占比达到 11%。

深度伪造专利申请案件状态分布图

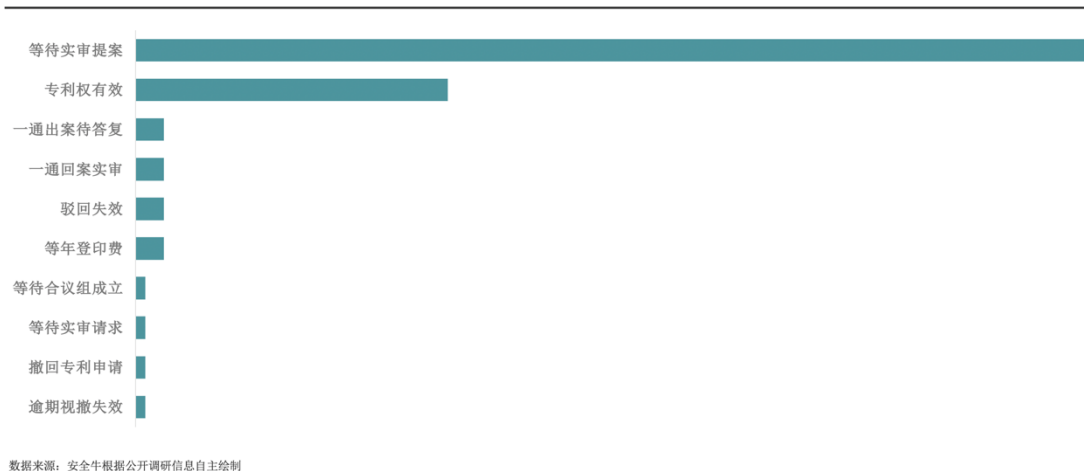
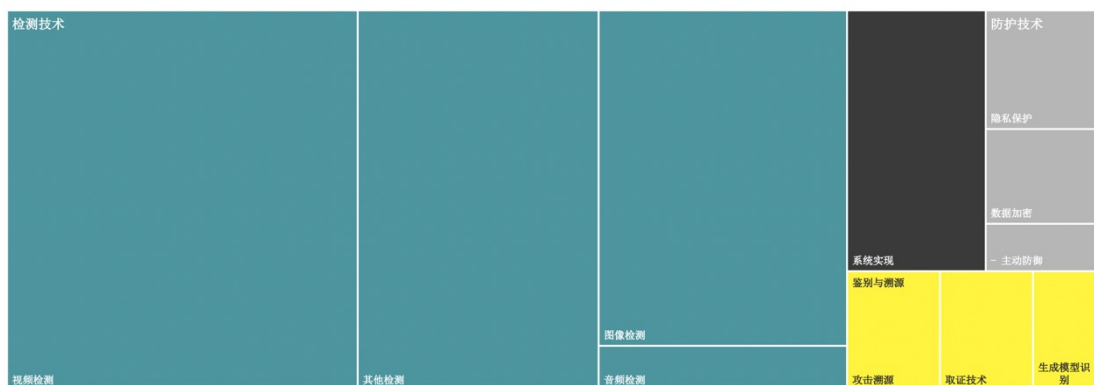


图 4.1.4 中国深度伪造专利申请案件状态分布图

安全牛公开调研收集的深度伪造专利相关数据显示，从深度伪造专利技术方向来看，检测类型是深度伪造发明专利的主要类型，占比达到 77%，其他大类，如防护类型、溯源和鉴别、系统实现占比一共 23%。在检测类型申请专利中，视频检测占比 42%，图像检测占比 26%，音频检测 3%，其他未能明确归结视音频和图像的检测类别占比达 29%。

深度伪造专利技术方向分布图



数据来源：安全牛根据公开调研信息自主绘制

图 4.1.5 中国深度伪造专利申请技术方向分布图

安全牛认为，深度伪造相关技术正处于快速发展阶段，当前的研究重点主要集中在检测技术上，未来在防护技术和系统实现方面还有较大发展空间。建议加强产学研合作，促进技术产业化，同时关注专利布局的均衡性，加强在防护技术和系统实现方面的创新。

4.1.2 国内深度伪造和合成媒体安全服务侧能力分析

安全牛认为，深度伪造和合成媒体安全服务侧厂商可以分为三类：AI 为主导的服务提供商、安全为主导的服务提供商和满足自我需求的内容运营商。每类厂商都有其独特的特点和优势，但也存在亟需提升的方面。

■ AI 为主导的服务提供商

安全牛认为，深度伪造和合成媒体安全服务侧厂商中，以 AI 为主导的服务提供商可以细化为两类：**第一类是以中科睿鉴和瑞莱智慧为代表的服务提供商**，他们主要依托学术起家，拥有浓厚的学术背景，并且是典型的国家队成员。中科睿鉴由中国科学院计算技术研究所的曹娟创立，其团队在国际顶会和权威学术期刊上发表了 70 余篇论文，并拥有多项伪造检测相关发明专利。瑞莱智慧则是依托清华大学人工智能研究院设立，由张钹院士和朱军教授担任首席科学家，专注于深度伪造内容检测平台的研发。**第二类是以“人工智能四小龙”为代表，如商汤科技、旷视科技等**，他们以人脸识别为主，主要客户为银行、金融等，但普遍存在安全能力相对有待提升的现实。

■ 安全为主导的服务提供商

安全牛认为，安全为主导的服务提供商，如国投智能、东方通、顶象、网易易盾、高维数据、亚信安全等，通常从攻防视角出发，目前主要依托客户现实需求提供定制化解决方案，如国投智能依托监管侧的需求研发的慧眼视频图像鉴真工作站，可为公安、司法、监委、第三方鉴定机构等行业客户在视频图像物证化处理中提供符合法庭科学的鉴定手段支撑；又如顶象依托保险客户研发的 Dinsight 实时风控引擎，帮助企业进行风险评估、反欺诈分析和实时监控，提高风控的效率和准确性。这些服务商在网络安全领域具有深厚的积累，能够提供全面的安全防护服务。然而，他们在 AI 能力方面存在一些欠缺，这不仅体现在技术储备和人才不足，而且在算法、算力等方面也缺乏后劲。

■ 满足自我需求的内容运营商

深度伪造和合成媒体技术的发展对内容运营商提出了新的挑战，尤其是在内容监管和自我审查方面。以抖音、爱奇艺、央视频等为代表的**内容运营商**，需要对输出的内容进行深度伪造检测，以确保内容的真实性和合规性，同时要接受监管侧的监管。通常他们的内审机制自我构建，设立多层次的内容审核体系，利用反深度伪造技术手段对视频内容进行自动筛查，排除不良信息。他们的需求很早就开始，以自我满足为主，这种满足分为自建和采购，但其较少对外输出安全检测能力。

深度伪造检测是一个多维度的问题，需要 AI 技术、安全防护和内容运营等多方面的合作。AI 为主导的服务商在技术上有优势，但需要加强安全能力；安全为主导的服务商在安全防护上有优势，但需要提升 AI 技术；内容运营商需要提升自身的技术能力和合作意识。三者之间的合作将有助于构建一个更加全面和有效的深度伪造检测体系。



图 4.1 6 中国深度伪造和合成媒体安全服务侧厂商分类及特点图

4.2 深度伪造和合成媒体安全防控体系分析

安全牛认为，当前针对实际场景中的深度伪造和合成媒体安全问题，需要构建一个关于深度伪造和合成媒体安全防控体系，可以分为三个主要层面，即治理层面、管理层面和技术层面。安全牛绘制了下图深度伪造和合成媒体安全防控体系分析图，以下是对这三个层面的详细解读：

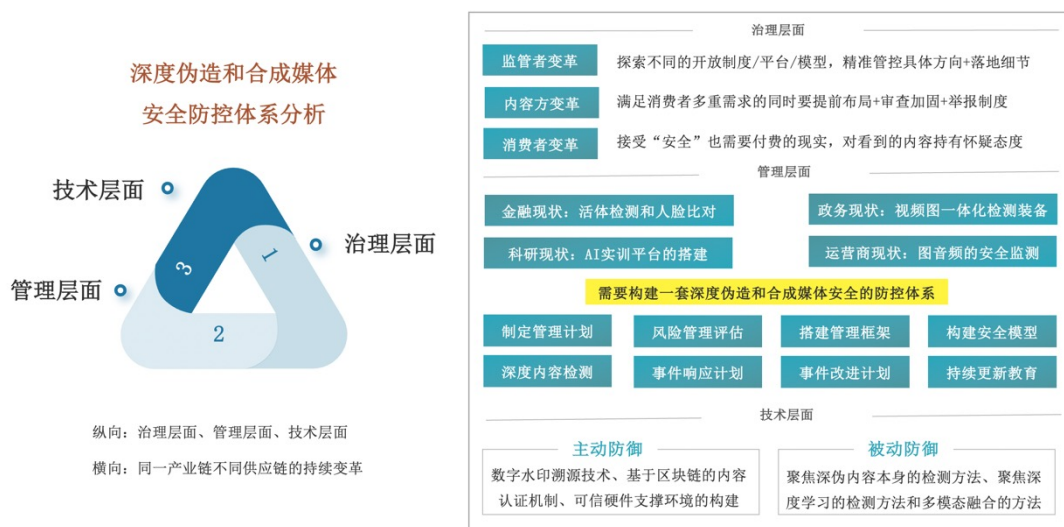


图 4.2.1 中国深度伪造和合成媒体安全防控体系图

4.2.1 深度伪造和合成媒体安全防控体系—治理层面

监管者变革：为应对深度伪造和合成媒体技术带来的挑战和问题，监管机构面临的任务愈发复杂且迫在眉睫。因此，监管机构需要采取全面、多维的策略，以确保能够有效识别、监管并防范深度伪造内容的不当使用。安全牛建议：监管机构需要探索不同的开放制度、平台和模型，以精准管控深度伪造和合成媒体的具体方向和落地细节。

内容方变革：随着生成式内容的指数级增长，内容方面临深刻变革。一方面，内容方需要顺应时代发展，满足大众日益提高的精神文化需要；另一方面，内容方需要建立专门的内容审查团队，负责监控上传的内容，快速响应并处理深度伪造内容的举报。安全牛建议：内容方需要满足消费者的多重需求，同时提前布局、审查加固和建立举报制度。

消费者变革：消费者需要“擦亮双眼”，在深度伪造和合成媒体内容高速发展的当下，眼见不一定为实，在肉眼分不清现实和虚拟的时候，借助科技的力量，也是不错的选择。安全牛建议：消费者需要接受“安全”也需要付费的现实，并对看到的内容持有怀疑态度。

4.2.2 深度伪造和合成媒体安全防控体系—管理层面

根据安全牛调研数据显示，监管方如各部委、公检法、网信办是深度伪造和合成媒体安全市场的最主要采购方，他们的采购占比超过 50%，其次是金融和运营商。安全牛认为，要想构建管理层面深度伪造和合成媒体安全防控体系，必须先了解主要采购方的安全现状，然后进行进一步系统梳理。

■ **政务现状：**深度伪造和合成媒体技术的普及让证据和数据面临多重风险，政府及监管方正采取措施应对这一挑战，并开发识别工具。执法部门急需能迅速鉴别多媒体真伪的装备，但现有产品在这方面能力不足，开发专业深度伪造鉴定装备任重道远。

■ **金融现状：**随着技术的发展和数字金融服务的普及，安全性和客户身份验证已成为银行、支付平台以及其他金融服务提供商最关心的问题之一。为了有效防止深度伪造带来的欺诈行为，保护客户资产安全，金融机构广泛采用了活体检测和人脸比对技术。

■ **运营商现状：**运营商在当前数字化、网络化快速发展的背景下，承担着对图文音视频内容进行安全监测的重要责任，旨在遏制和防止深度伪造内容的传播。运营商采取多种措施确保网络空间的清朗，如部署先进监测产品，与其他机构合作共同构建深度伪造样本库等。

■ **科研现状：**科研领域正积极搭建 AI 实训平台，以应对深度伪造技术带来的全球性挑战，旨在提升检测技术的准确性和响应速度，同时培养专业人才。这些平台促进了多模态数据分析、跨领域合作和技术的开放共享。

安全牛认为，虽然深度伪造和合成媒体安全的主要行业面临的现状各不相同，但均可按照制定管理计划—管理风险评估—管理框架搭建—构建安全模型—深度内容检测—事件响应计划—事件改进计划—持续更新教育的方式进行防御体系构建。

■ **制定管理计划：**包括风险管理评估、模型及内容深度检测、事件响应计划和事件改进计划，以构建一套深度伪造和合成媒体安全的防控体系。

■ **管理风险评估：**需要对现有业务流程中可能受到深度伪造影响的环节进行识别和评估。这包括但不限于用户认证、数据传输、通信安全等环节。针对每一个环节，评估深度伪造技术可能带来的风险，包括数据泄露、欺诈行为、声誉损害等，并制定相应的风险缓解措施。

■ **管理框架搭建：**建立一个全面的管理框架，以支持深度伪造和合成媒体的安全防控。

■ **构建安全模型：**利用大模型技术构建防伪模型，以实现伪造、假冒检测问题的复杂模型构建。防伪大模型因其参数量大、训练数据规模大，实际落地面临问题。

■ **深度内容检测：**利用现有的深度伪造检测技术，如中科睿鉴围绕数字内容伪造检测技术，布局“模型—

AI 深度伪造和合成媒体安全市场技术落地分析

数据一硬件”的 AI 技术设施，面向国家、行业、个人安全场景，提供音视图文全栈全类型鉴伪技术和产品服务；又如瑞莱智慧深度伪造内容检测平台 DeepReal 等，这些工具能够分析媒体内容并识别出是否为深度伪造。

■ **事件响应计划：**将深度伪造内容纳入安全事件响应计划，确保快速反应和降低损失。

■ **事件改进计划：**根据前面的深度伪造内容进行不断改进，确保防控体系能够适应技术进步和新出现的威胁。通过改进强化检测能力，提升服务意识，优化应急响应流程，以便更有效地应对深度伪造带来的挑战。

■ **持续更新教育：**由于深度伪造技术不断进步，需要定期更新检测工具和方法，以应对新的威胁。对相关从业人员进行培训，提高他们对深度伪造技术的认识，以及在面对潜在深度伪造内容时的应对能力。

4.2.3 深度伪造和合成媒体安全防控体系—技术层面

安全牛认为，深度伪造和合成媒体安全防控体系技术层面，包含主动防御和被动防御两方面。主动防御包括数字水印溯源技术、基于区块链的内容认证机制和可信硬件支撑环境的构建。被动防御包括聚焦于深度伪造内容本身的检测方法，以及深度学习和多模态融合的方法。

(1) 组织在主动防御技术方面的最佳实践

■ 数字水印和可追溯性

组织可将稳健的水印技术以及自动化的提取与验证能力集成到企业内容安全计划中，这是遏制未经授权分发并在复杂媒体供应链中追踪来源的关键措施。

为了有效保护企业生成的媒体内容，建议实施结合扩频和 DCT 技术的多层水印策略。水印的载荷应包括每个内容的唯一标识符，以及诸如创建时间戳和作者等元数据。嵌入过程应尽可能接近内容创建的时间点，以最大限度地减少未标记分发的风险。

为了充分利用数字水印技术进行内容保护，开发高效的水印提取和验证工具至关重要。这些工具应能够在常见的信号处理操作和有意攻击的情况下，可靠地恢复嵌入的水印。水印提取通常需要应用与嵌入方法匹配的信号处理技术。例如，对于扩频水印，这可能包括计算内容与扩频序列之间的相关峰值；对于基于 DCT 的水印，则通常通过比较变换系数与预定阈值进行提取。为了确保提取水印的准确性，验证过程必须考虑潜在的比特错误和修改。这可以通过使用纠错编码、冗余水印嵌入以及对多个水印组件进行基于共识的验证来实现。

■ 基于区块链的内容认证

通过结合稳健的内容注册流程和在内​​容管理系统中实施的自动化区块链验证，我们可以创建一个强大的认证框架，有效防止篡改、确立内容溯源并增强对数字内容的信任。随着区块链技术在媒体行业的不断成熟和广泛应用，主动将这些能力集成到工作流中，可以使用组织在内容安全和真实性方面获取充分的保障能力。

区块链技术为保护和认证数字内容提供了一种强大的工具。通过利用区块链的不可篡改性和去中心化特性，组织可以创建防篡改的内容溯源和完整性记录。私有许可区块链（如 Hyperledger Fabric）尤其适用于企业内容认证的应用场景。Hyperledger Fabric 是一个用于构建支持可插拔共识协议、灵活访问控制和可定制智能合约的私有区块链的开源框架。其模块化架构使组织能够根据特定的安全和性能需求定制区块链。

为了利用区块链有效认证内容，必须建立一个全面的注册流程。这包括为每个内容创建区块链交易，并包含其加密哈希值、元数据和溯源信息。内容哈希值作为唯一的指纹，可以在内容生命周期的任何阶段验证其完整性。通过将此哈希值以及元数据（如创建者、时间戳和版本）存储在区块链上，可以创建内容来源和历史的不可更改记录。

为了充分发挥基于区块链的内容认证的优势，必须将区块链验证功能无缝集成到现有的内容管理和分发系统中。这种集成可以在内容生命周期的各个阶段实现内容完整性和溯源的自动化验证。

根据世界经济论坛 2021 的一份报告，将区块链集成到媒体和内容管理工作流中有助于打击虚假信息和假内容的传播。通过将区块链验证直接构建到内容系统中，我们可以实现认证过程的自动化，并为内容完整性提供端到端的保证。

■ 利用可信硬件环境

通过结合可信硬件环境和基于硬件的密钥管理技术，我们可以为主动内容保护建立稳健且具有弹性的基础。这些技术在面对复杂对手和内部威胁时，能够为媒体认证和水印工作流的完整性、机密性和真实性提供强有力的保障。

通过将可信执行环境（Trusted Execution Environments, TEEs）集成到内容安全工作流中，可以显著提高敏感操作的安全性，保护数字资产的完整性和隐私，为应对日益复杂的安全挑战奠定基础。TEE 是一种基于硬件的解决方案，用于安全且隔离地执行敏感操作，例如水印嵌入、内容哈希计算和人工智能模型训练。TEE 通过创建安全的隔离区（enclaves），保护代码和数据免受未经授权的访问，即使是特权系统进程或物理篡改尝试也无法干扰。

通过将基于硬件的密钥管理集成到组织的内容保护工作流中，可以显著提高密钥的安全性，降低泄露风险，为数字资产的保护提供强有力的支持。有效的密钥管理是维护加密操作（如水印嵌入和内容哈希计算）安全性的关键。基于硬件的密钥管理解决方案，例如硬件安全模块（HSM）和可信平台模块（TPM），为敏感的加密密钥提供额外的保护层。探索安全多方计算

安全多方计算（Secure Multi-Party Computation, MPC）是一种加密技术，允许多个参与方在不暴露各自输入的情况下共同计算一个函数。MPC 使得在不信任的各方之间进行隐私保护的协作和计算成为可能，非常适合用于协作式媒体认证和水印嵌入等场景。在基于 MPC 的媒体认证方案中，内容创作者和分发者可以安

全地贡献各自的水印密钥、哈希算法和内容元数据,而无需向其他方暴露其专有信息。MPC 协议确保计算输出(如带水印的内容或认证结果)对所有参与方可用,同时保持每个参与方输入的机密性。

(2) 组织在被动防御技术方面的最佳实践

■ 针对伪造内容的检测

针对深度伪造视频中的伪造特征和不一致性,例如异常的眼睛眨动模式、唇同步不匹配以及肤色不均匀等问题进行检测。检测时可利用卷积神经网络(CNN)和循环神经网络(RNN)分析视频帧中的空间和时间特征,识别出操控的迹象。

尽管伪造特征检测模型专注于深度伪造视频中的视觉不一致性,但更全面的检测方法需要分析多个模态,包括视觉、音频和语言线索。结合计算机视觉、自然语言处理(NLP)和音频分析技术的集成模型相比单一模态方法在深度伪造检测中表现出了更高的性能。通过实施多模态集成模型,组织可以显著提高深度伪造检测的准确性和鲁棒性,为应对深度伪造技术的快速发展提供强有力的保障。

■ 基于深度学习的检测

通过利用先进预训练模型的迁移学习、探索快速适应的小样本学习方法,以及集成 Grad-CAM 等可解释 AI 技术,我们可以显著提升基于深度学习的深度伪造检测能力的性能、适应性和可解释性。在深度伪造生成与检测的对抗中,保持在深度学习研究与创新的前沿对维护这一威胁的强大防御至关重要。

采用迁移学习可以有效地利用已有的模型和技术,加快深度伪造检测的开发周期,同时确保检测性能和效率在不断演变的深度伪造环境中保持领先。利用预训练模型的知识,能够更高效地解决相关任务,同时提升性能。通过将最新的模型(如用于图像分析的 EfficientNet 和用于语音处理的 Wav2Vec)适配于深度伪造检测,我们可以显著增强检测能力,同时减少对大规模训练数据和计算资源的需求。

小样本学习是深度学习中的新兴范式,旨在使模型能够通过有限的训练样本快速适应新任务和类别。在深度伪造检测领域,小样本学习可以帮助我们快速开发针对新型深度伪造生成方法和伪造特征的检测能力,即使仅有少量示例可供学习。

可解释人工智能(Explainable AI, XAI)技术旨在为深度学习模型的预测提供人类可理解的解释,从而提高 AI 驱动决策的透明度、信任度和责任感。在深度伪造检测中,集成如梯度加权类激活映射(Grad-CAM)的 XAI 技术,可以帮助我们更好地理解 and 解释模型检测决策背后的因素,从而提升模型的鲁棒性和可靠性。

■ 多模态检测融合

即使是最先进的检测模型在面对新型和未知的深度伪造类型时,性能也可能显著下降。这突显了建立多模

态数据集扩展和优化的稳健流程与管道的重要性，以应对快速变化的深度伪造威胁环境。通过开发有效的多模态深度伪造检测架构和流程，包括视听交叉验证、集成融合、公众人物知识图谱以及数据集扩展和优化，可以创建一个全面且可适应的防御系统，充分利用不同模态和技术的互补优势。面对日益增长和演变的深度伪造威胁，保持警惕并主动应对对于检测和对抗这些恶意媒体操控至关重要。

视听交叉验证是一种深度伪造检测技术，通过利用真实视频内容中视觉和听觉组件之间的固有同步性，检测伪造视频。通过开发能够识别视频中人物唇部动作、面部表情和说话风格之间不匹配和不一致性的模型，可以有效标记未能保持自然同步的深度伪造视频。

集成方法是另一种多模态检测技术。通过结合多个模型或模态的输出，可以提高深度伪造检测系统的整体性能和鲁棒性。通过使用如堆叠（Stacking）、提升（Boosting）和装袋（Bagging）等技术，可以有效融合视觉、听觉及其他模态的互补检测信号，从而实现更准确和可靠的预测。

知识图谱是一种强大的工具，可以用结构化和机器可读的方式表示和推理实体（如公众人物）的关系和属性。通过将公众人物的知识图谱集成到多模态深度伪造检测管道中，我们可以利用这些先验知识来识别可疑深度伪造视频中人物身份、外貌和行为的非一致性和异常情况。

为了有效训练和评估多模态深度伪造检测模型，需要访问大规模、多样化且具有代表性的数据集，以覆盖各种深度伪造类型、质量水平和操控技术。然而，随着深度伪造方法的快速发展和传播，维护和扩展这些数据集以跟上不断变化的威胁环境是一项重要挑战。为了解决这一挑战，需要建立系统的流程和管道，持续从各种来源收集、注释和优化多模态深度伪造数据集，以应对不断变化的深度伪造威胁。

4.3 深度伪造和合成媒体安全技术落地难点分析

安全牛认为，当前深度伪造和合成媒体安全技术落地难点主要体现在以下方面：

1. 难以平衡深度伪造智能生成与场景适应性

随着深度学习技术的快速发展，Deepfake 技术可以以较低成本生成高度真实的人脸伪造内容。然而，由于不同场景需求的差异，单一的伪造框架难以适应不断变化的需求。因此，结合人工干预的处理方法，可能是实现更高级深度伪造的关键。

一种潜在的解决方案是开发模块化的深度伪造框架，允许用户根据特定用例轻松定制和微调。这可能涉及创建包含不同面部特征、表情和属性的预训练模型库，用户可以根据需要自由组合。此外，将用户反馈和人工调整整合到生成过程中，可以帮助优化输出结果，确保其符合期望的标准。

2. 亟需构建多维度的主动防御体系

在防御深度伪造技术方面，事后的检测技术往往难以抵消已经造成的损害。因此，主动防御技术，即在伪造内容产生之前就进行干预和防护，成为一个重要的研究方向。尽管该领域的研究相对较少（例如，目前实践主要是将安全模块内置到手机等终端设备中，以提升私域安全性），但它代表了未来防御技术发展的重要趋势。

为了构建一个有效的多维度主动防御体系，安全人员需要探索多种策略，从伪造生成的各个阶段进行预防或干扰。这包括开发算法以识别并标记潜在的深度伪造源材料，以及设计嵌入水印或其他识别特征的技术，使真实媒体更难被篡改。另一个有前景的研究方向是利用区块链技术创建不可篡改的媒体来源和真实性记录，从而增强可信度和透明性。

3. 真实环境中的深度伪造检测能力不足

在真实场景下，深度伪造检测技术面临鲁棒性和实用性方面的重大挑战。传播场景的复杂性和多样化的对抗手段使现有检测模型的泛化能力尚未达到可靠水平。这一点在 2020 年 Facebook 主导的 DFDC (Deepfake Detection Challenge) 比赛中得到了验证。为了提高检测技术的泛化能力，关键在于开发多样化且高度真实的数据集，并不断探索反伪造方法。

一种提升真实环境中深度伪造检测能力的方法是专注于开发多模态检测方法，这些方法可以分析视频或图像的多个方面，如音频、视觉和上下文线索。通过整合来自不同模态的信息，可能更容易发现深度伪造的迹象。此外，另一个重要的研究领域是开发更复杂的机器学习模型，能够适应新型深度伪造的出现。这可能包括使用迁移学习或小样本学习等技术，快速训练检测模型以应对新出现的数据集。

4. 深度伪造安全技术的社会与伦理问题

深度伪造安全技术的开发与应用虽然带来了技术上的突破，但其潜在的社会和伦理问题同样不容忽视。深度伪造技术易于获取和操作，这使得不法分子能够利用它生成虚假信息、伪造身份或传播不实言论，从而对社会、政治和个人安全构成威胁。

为了规范深度伪造技术的使用，需要尽快制定法律法规，并明确违法行为的界定与处罚措施，包括但不限于：明确立法禁止恶意生成和传播伪造内容；对生成的深度伪造内容嵌入溯源标记，使得任何伪造内容都能追踪到其源头；规定社交媒体和内容传播平台的审核义务，对于恶意深度伪造内容应迅速检测并移除，并向相关部门报告。

其次，技术防护与法规建设的有效性，离不开公众的广泛参与。应过科普宣传让公众了解深度伪造的特点及识别方法，例如观察视频的不自然细节、背景异常或声音延迟等；提升个人隐私保护意识，例如避免随意上传敏感照片或音视频素材，以减少深度伪造素材来源；建立便捷的举报渠道，让公众能够迅速报告可疑的伪造

内容，从而形成社会共同防控的网络。

另外，技术的发展需要伦理与道德的规范以保证其正当应用。这要求技术开发者、政策制定者与社会各界通力合作，在深度伪造技术开发阶段，应引入伦理审查机制，确保技术用途与道德价值一致；借鉴人工智能治理的成功经验，吸纳学术机构、技术企业、政府部门及民间团体参与深度伪造的伦理监管；深度伪造的传播无国界特性，要求各国合作制定统一的标准与法规，形成国际共识，共同打击跨境伪造行为。

深度伪造技术的社会和伦理问题具有复杂性与多维性，单靠技术防护无法彻底解决。通过法律、教育、伦理、技术四位一体的综合治理模式，可以在规避技术风险的同时，确保深度伪造技术的负责任使用，为社会带来更多积极价值。

第五章 国外深度伪造和合成媒体安全工具 / 平台介绍（按首字母排序）

安全牛分析师筛选了国外专注于检测和防范深度伪造内容的四个工具 / 平台，从工具 / 平台介绍及具体特点给国内读者进行相关举例展示，他们分别是：Deepware、Illuminarty、Sentinel 和 WeVerify。

5.1 Deepware

【平台介绍】

Deepware 是一款由人工智能驱动的创新解决方案，专门用来对抗深度伪造视频的威胁。它通过使用先进的 AI 算法分析数字媒体内容，识别操纵迹象并提供概率得分，指示内容被合成修改的可能性。Deepware 的主要功能包括：基于人工智能的深度伪造检测、详细的报告、用户友好的界面、多功能性以及多平台支持。它能够检查图像和视频，确保全面识别深度伪造，并且可以直接扫描来自 YouTube、Facebook 和 Twitter 等平台的视频链接。Deepware 在新闻、社交媒体审核和企业安全等领域有广泛的应用，帮助新闻机构、社交平台等免受恶意深度伪造的侵害。

【工具特点】

- 为所有用户提供实时深度伪造检测
- 可跨多个平台分析视频，可直接对 YouTube、Facebook 和 Twitter 等视频链接进行检测

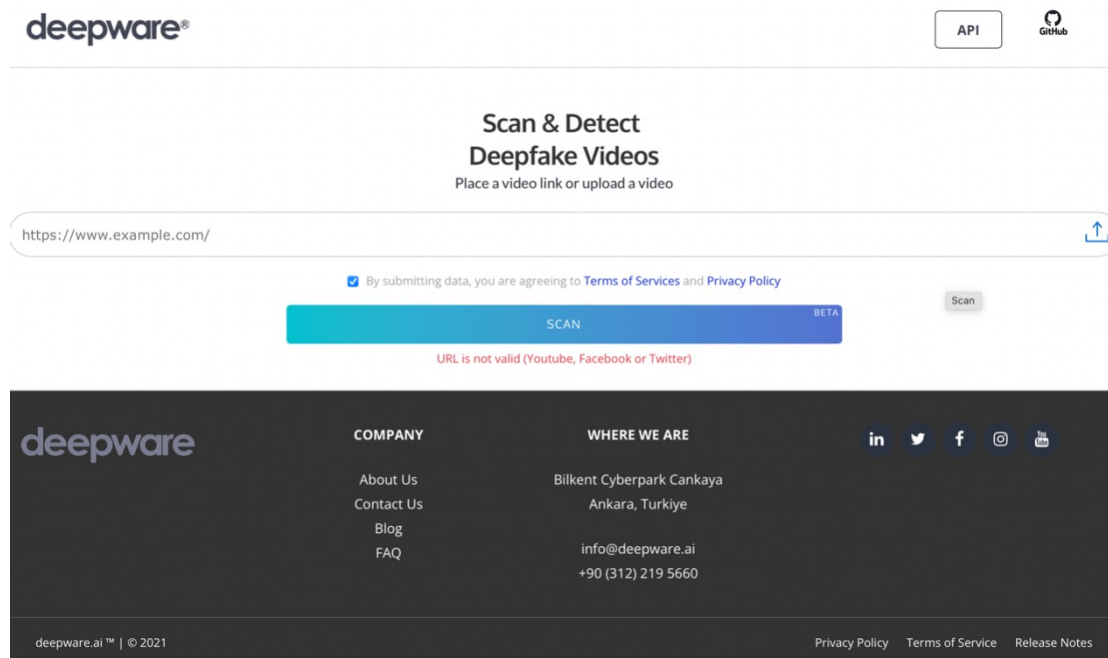


图 5.1 1 Deepware 验证页图

5.2 Illuminarty

【工具介绍】

Illuminarty 是一款在线工具，专门用于检测 AI 生成的图像、合成图像、篡改图像和深度伪造图像。它运用多种计算机视觉算法来评估图像是否由公共 AI 模型生成，并提供图像中 AI 生成区域的局部检测功能。此外，Illuminarty 还提供自动检测 API，以及识别生成器的详细信息，帮助用户进一步判别。

Illuminarty 提供免费的基础服务，用户可以通过月度订阅计划来获取更多高级功能，例如更高的每日请求限制和 API 集成。这使得 Illuminarty 不仅适用于个人用户，也适合需要更专业服务的企业用户。下图罗列 Illuminarty 的收费标准，分为免费版、基础收费（10 美元 / 月）、进阶收费（30 美元 / 月）和企业级按需定制。

【工具特点】

- **快速检测：**只需把照片放进去，几秒就可以显现检测结果，检测结果为 AI 可能性比例数据。
- **准确率低：**根据安全牛分析师多次测试，检测结果准确率低，如分析师放入本人照片截图，AI 可能性比例会高达 74%，如果放入同样的照片原图，则 AI 可能性比例只有 17%。

Free	Basic	Pro	Enterprise
We will always provide the basic AI detection functionalities for free.	Basic functionalities of Illuminarty, made for individuals.	Full functionalities of Illuminarty, made for professionals.	For the integration of Illuminarty into custom services.
\$0.00 /mo	\$10.00 /mo	\$30.00 /mo	Price on Request
Open Webapp	Use Basic Plan	Use Pro Plan	Contact Us
Included Features ✓ AI image classification ✓ AI text classification	Included Features ✓ AI image classification ✓ AI image localized detection ✓ AI model identification for image generators ✓ AI text classification ✓ AI text localized detection ✓ API usage (10,000 requests/day)	Included Features ✓ All Basic plan functionalities, in nightly build ✓ Increased API usage (40,000 requests/day) ✓ API integration support	Included Features ✓ User-based custom detection ✓ Unlimited API Usage ✓ Custom API server ✓ SLA

图 5.2 1 Illuminarty 收费标准图

5.3 Sentinel

【平台介绍】

Sentinel 是一个基于人工智能的保护平台，被政府和企业用于打击深度伪造。它允许用户上传媒体进行自动分析，以检测和防范深度伪造内容的传播。Sentinel 的操作步骤：**1) 媒体上传：**用户可以通过 Sentinel 的网站或 API 上传数字媒体文件，这些文件可以是视频或音频样本。**2) AI 欺诈的自动检查：**Sentinel AI 在媒体提交后立即开始工作，自动对其进行分析，寻找 AI 伪造的证据，这些证据往往是肉眼难以察觉的细节。**3) 确定深度伪造的真实性：**Sentinel AI 使用的技术能够区分人为更改的内容和真实内容，帮助揭露可能削弱信心和传播虚假信息的深度伪造。**4) 操作可视化：**Sentinel AI 不仅能够识别深度伪造，还能提供操纵的可视化展示。

【工具特点】

- 适合分析多种形式的数字媒体内容
- 提供媒体操作的视觉表示
- 简化上传和识别虚假媒体的过程
- 安全牛分析师进行测试，没有直接开放的测试入口。需要预定演示，预定时需要填写姓名、企业名称、职位、请求细节等信息。

[About](#) [Contact](#) [Careers](#) [Book a demo](#)

Defending Against Deepfakes and Information Warfare

We Help Democratic Governments,
Defence Agencies and Enterprises Stop
the Threat of Deepfakes through a
Leading AI-Based Protection Platform.

[Book a demo](#)

图 5.3 1 Sentinel 官网首页图

5.4 WeVerify

【平台介绍】

WeVerify 是一个欧盟资助的研究项目，旨在开发智能人机交互内容验证和虚假信息分析的方法和工具。该项目旨在通过分析和情境化社交媒体和网络内容，揭露虚假信息。它采用跨模式内容验证、社交网络分析、微目标揭穿和基于区块链的虚假信息数据库等方法。WeVerify 项目还包含一个名为 InVID WeVerify 的验证插件，这是一个开源工具箱，为记者、事实核查者、人权捍卫者、研究人员和公众提供组件，帮助他们验证跨模态信息并揭穿虚假信息。目前，该插件已经被下载和使用了超过 57000 次。

【平台特点】

- 跨模式内容验证
- 社交网络分析
- 微目标揭穿
- 开源平台

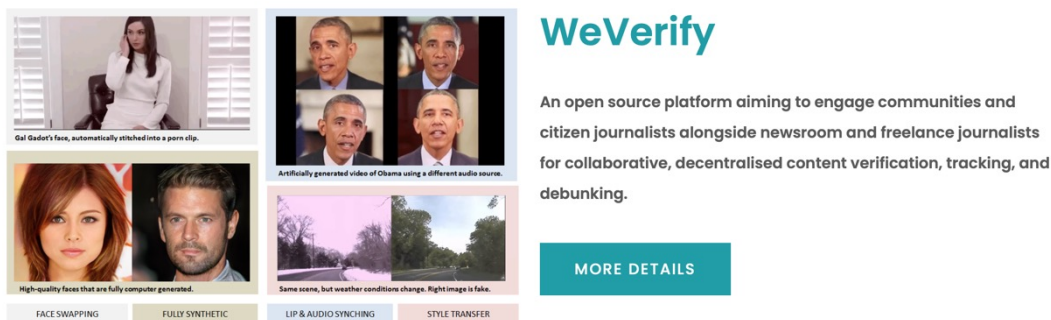


图 5.4 1 WeVerify 官网首页图

第六章 国内深度伪造和合成媒体安全厂商典型案例 （按简称字母排序）

6.1 顶象典型案例分析

本案例深度伪造和合成媒体服务提供商：顶象

顶象作为国内业务安全领域的领先企业，专注于帮助企业建立自主可控的安全体系，应对伪造、篡改等业务欺诈，抵御网络黑灰产风险，保障业务健康稳定，推动企业创新增长。公司自主研发的业务安全感知防御云涵盖设备指纹、智能验证等产品，服务于多个行业，积累了丰富的实战经验，提供全生命周期安全服务。顶象总部位于北京，并在多地设有分部，是国家级技术支撑单位和信创工委会会员，获得多家资本投资。

深度伪造和合成媒体典型案例：顶象助力某保险集团防范虚假人脸考勤

【背景介绍】

随着保险公司考勤系统的数字化转型，考勤作弊现象不断增加。2023—2024 年，顶象防御云业务安全情报中心监测发现，不法分子破解多家公司考勤系统，并制作打卡作弊工具，向员工兜售“代打卡服务”，使得员工不出门即可轻松领取全勤奖。

虚假考勤打卡严重损耗保险公司支出，造成投入浪费。2019 年某保险企业分公司有代理人实名爆料，其所在分公司为了完成增员人数任务，在内部系统中窃取客户身份证等个人资料办理虚假入司，10 多年来平均每年有 200 多人的虚假增员，套取公司奖金、绩效和队伍建设费几百万元等。

【安全需求】

不法分子通过伪造人脸视频和考勤 App 等手段，实现虚假打卡，主要方法为：①伪造人脸视频代打卡。个人提供真实的人脸视频，不法分子上传视频至考勤系统，骗过考勤系统的人脸识别，帮助购买者完成每日的考勤打卡。如果个人只提供照片，不法分子就利用 AIGC 和其他视频编辑软件，制作出具有高真实性的人脸视频，精确地模拟目标员工的面部表情和动作，系统将错误地认为是员工本人，从而完成考勤打卡。②伪造考勤 App。不法分子通过逆向工程或利用系统漏洞，获取保险公司考勤系统的源代码或协议文档，破解考勤 App，制作出山寨版考勤 App。该 App 能够屏蔽真实的摄像头影像采集，拦截蓝牙和无线网络并伪造 GPS 定位。当个人在家中使用该 App 打卡时，会模拟出公司场景、位置等信息，系统将错误地识别为正常的考勤打卡。

保险公司亟须强化人脸识别技术的安全性，通过整合行为识别、设备指纹等多重验证手段来增强防御。同时，必须提升考勤 App 的安全防护，定期开展渗透测试，以便快速定位并修补潜在的安全漏洞。此外，实施动态验证码和实时监控机制，将有助于显著防止作弊事件的发生。

【解决方案】



图 6.1.1 顶象防范深度伪造虚假人脸考勤案例解决方案图

■ **识别虚假的人脸视频。**顶象全链路全景式人脸安全威胁感知方案通过在设备环境、人脸信息、图像鉴伪、用户行为、交互状态等多维度信息进行智能核验，快速识别注入攻击、活体伪造、图像伪造、摄像头劫持、调试风险、内存篡改、Root/越狱、恶意 Rom、模拟器等运行和系统类等 30 多类恶意攻击行为，及时发现伪造视频、虚假人脸图片、异常交互行为后，可自动阻断操作。同时能够灵活配置视频核验强度与友好度，实现对正常用户无感验证，对异常用户加强验证的动态机制。

■ **识别异常的设备。**顶象设备指纹通过对设备指纹的记录和比对，可以辨别合法用户和潜在的欺诈行为。其对每个设备进行唯一标识，识别出虚拟机、代理服务器、模拟器等被恶意操控的设备，分析设备是否存在多账号登录、是否频繁更换 IP 地址、频繁更换设备属性等出现异常或不符合用户习惯的行为，帮助追踪和识别欺诈者的活动。

■ **识别异常操作的账号。**异地登录、更换设备、更换手机号、休眠账户突然活跃等，需要加强频繁验证；此外，保持持久性检查以确保用户的身份在使用期间保持一致。顶象无感验证基于 AIGC 技术，能够防止 AI 的暴力破解、自动化攻击和钓鱼攻击等威胁，有效防止未经授权的访问、账户被盗用和恶意操作，从而保护账户的稳定性。

国内深度伪造和合成媒体安全厂商典型案例（按简称字母排序）

■ **挖掘潜在欺诈威胁。**顶象 Dinsight 实时风控引擎帮助企业进行风险评估、反欺诈分析和实时监控，提高风控的效率和准确性。Dinsight 的日常风控策略的平均处理速度在 100 毫秒以内，支持多方数据的配置化接入与沉淀，能够基于成熟指标、策略、模型的经验储备，以及深度学习技术，实现风控自我性能监控与自我迭代的机制。与 Dinsight 搭配的 Xintell 智能模型平台，能够对已知风险进行安全策略自动优化，基于风控日志和数据挖掘潜在风险，一键配置不同场景支持风控策略。其基于关联网络和深度学习技术，将复杂的数据处理、挖掘、机器学习过程标准化，提供从数据处理、特征衍生、模型构建到最终模型上线的一站式建模服务。

【案例价值】

此方案在特大型金融保险企业公司省级分公司部署后，当月发现 10000+ 代理人虚假考勤打卡，当月拦截阻止超过 15 万次虚假考勤打卡操作，为保险公司挽回 500 万+ 的代理费用。该特大型金融保险企业公司，名列《财富》世界 500 强，业务范围全面涵盖寿险、财险、企业和职业年金、银行、基金、资产管理、财富管理、实业投资、海外业务等多个领域。2023 年，营业收入近 9900 亿元，合并保费收入近 7900 亿元，员工规模近 18 万。

6.2 东方通典型案例分析

本案例深度伪造和合成媒体服务提供商：**东方通**

东方通（股票代码：300379），是国内 A 股上市的基础软件厂商，以“安全+”、“数据+”和“智慧+”三大产品体系为基础，为客户提供综合解决方案及服务，是大安全及行业信息化产品、解决方案提供商。东方通的产品及解决方案广泛应用于国内数千个行业，服务电信、金融、政府、能源、交通等行业领域 10000 多家企业级用户，与 3000 多家合作伙伴携手打造合作共赢的产业生态。东方通连续十二年被认定为“国家规划布局重点软件企业”，承担多项国家重大科技专项的研制任务，是北京软件和信息服务业综合实力百强企业，曾荣获国家科技进步二等奖、北京市科学技术进步二等奖等多项荣誉。

深度伪造和合成媒体典型案例：**东方通为某单位构建深度伪造检测系统**

【背景介绍】

随着 AI 技术和大型模型的高速发展，生产效率、教育普及受到极大促进。但与此同时，不法分子利用深度伪造技术，制造虚假新闻误导公众、伪造私人音视频侵犯隐私、影响公众舆论等风险也构成严重安全隐患，亟需有效检测和防范。因此，某单位拟开展深度伪造合成检测工作，构建一套高效、可靠的深度伪造检测系统，实现对四类内容的深度伪造检测。首先通过收集和整理真实与伪造的音视频、图像数据，构建高质量、覆盖广泛的数据集，为检测模型的训练和评估提供基础。数据集将涵盖各种场景和类型，确保模型在不同应用场景下的泛化能力。然后开发端到端的检测系统，从数据输入到结果输出全流程自动化处理。支持实时检测和批量处理，满足不同用户需求。

【建设目标】

■ **技术突破与创新能力提升：**某单位开展音视图文鉴伪技术研究，对多种主流生成内容进行真伪检测，针对新出现生成算法，仅依赖少量数据样本便可实现鉴伪算法快速迭代，通过自主研发和创新，提升系统的检测精度和效率。

■ **满足行业监管与合规要求：**深度伪造检测系统符合国家和行业的监管要求，确保在检测过程中遵循相关法律法规和标准。通过系统的实施，助力监管机构有效打击网络虚假信息，维护网络空间的清朗和秩序，保障公众利益。

■ **保障用户信息安全与隐私：**在深度伪造检测系统建设过程中，某单位将用户信息安全和隐私保护放在首位。通过采用先进的加密技术和安全措施，确保用户数据在检测过程中的安全性和保密性，防止数据泄露和滥用，提升用户对系统的信任度和满意度。

国内深度伪造和合成媒体安全厂商典型案例（按简称字母排序）

■ **推动行业标准化与规范化发展：**通过系统的建设和应用，促进深度伪造检测技术的规范化发展，提升整个行业的技术水平和竞争力，为行业的可持续发展贡献力量。

【解决方案】

部署方案：部署深度伪造检测平台，以 MySQL 集群、Redis 集群、存储服务器等为底层支持，根据业务量的变化扩展核心计算单元 GPU 服务器。同时，可对不同的服务器设置任务调度的机制，以保障服务的高可用性。对外通过防火墙的安全保障，为用户封装成 HTTPS 等形式。

■ **针对小样本导致的识别欠拟合：**本项目不仅依赖于通过爬虫从互联网上收集的数据，还利用部署在 IDC 数据机房的 DPI 设备来获取互联网上真实会被用户访问到的数据。这两种数据源提供了多样性和丰富性，有助于增加训练数据的数量和多样性；对于负样本采用现代流行的各种伪造算法生成大量的负样本数据。这些数据不仅能够帮助模型学习如何识别现有的深度伪造技术，还能够扩展模型的训练空间，提高其对新型伪造技术的适应能力。开发有效的多模态数据融合策略，确保模型能够同时处理和学习来自不同模态的信息，提高识别的准确性。

■ **针对伪造图像视频检测模型的泛化性问题：**融合新型的特征提取、模型训练、模型设计方法。本项目将建立面向伪造图像和视频的跨模型集成学习与多任务学习机制，在模型设计阶段，通过添加注意力机制，使得网络能更好地关注在具有重要信息的特征空间；在模型训练阶段，采用数据和知识协同驱动的跨模型集成学习方法，研究多任务的目标优化模型更新策略，结合伪造内容生成机理，挖掘伪造共性痕迹，突破内容理解层面的视频图像鉴伪技术，提升伪造内容的检测能力。

■ **针对跨模态深度伪造线索建模能力不足技术难点和创新性：**基于多模态特征语义对齐的视音混合伪造检测技术，通过捕捉面部区域与音频的时空、语义一致性以提高多模态检测准确性和鲁棒性。聚焦现有伪造方法生成多模态伪造数据存在面部—音频细粒度对齐不准确的问题，提取视频音频全局和局部特征，对时空不一致、语义不一致进行建模，实现高精度检测。

【案例价值】

■ **提升深度伪造检测算法先进性。**智融伪影识别：本项目首次结合频域特征、时空域特征等特征提取技术，采用注意力机制、递归神经网络等模型结构设计，提升伪造图像视频检测模型在高折损场景下的识别准确率，提升检测系统在互联网流量下的可靠性。模态智融检伪：本项目基于现有生成技术无法精确实现音频和视频细粒度对齐线索，研究多模态特征融合、特征语义对齐等技术方案，利用多模态信息互补验证，提高多模态检测的准确性和可解释性。

■ **驱动技术创新与发展。**提升信息真实性与安全性：通过高精度检测算法，有效识别并过滤掉伪造的文件、图像、视频，确保信息的真实性和准确性，保障受众免受虚假信息的误导和欺骗。优化内容管理与审核流程：

系统自动化处理大量内容，提高企业审核效率，减少人工干预，帮助受众在海量信息中快速识别并处理伪造内容，优化企业内容管理流程。促进业务合规与法律保护：深度伪造检测产品符合相关法律法规要求，帮助客户遵守信息监管政策，降低法律风险，保护客户及用户免受法律纠纷的困扰。提升用户体验与满意度：通过提供可靠、高效的内容审核服务，确保用户接收到真实、有价值的信息，提升用户体验和满意度，增强客户与用户之间的互信关系。促进技术创新与业务发展：深度伪造检测技术的应用为客户提供了技术创新的机会，助力客户在数据安全、内容审核等领域取得突破，推动业务发展和市场竞争力的进一步提升。

6.3 国投智能典型案例分析

本案例深度伪造和合成媒体服务提供商：国投智能

作为网络空间安全与社会治理领域国家队，国投智能（厦门）信息股份有限公司 2017 年成立 AI 研发中心，深入开展人工智能技术研究。为了应对利用人工智能技术可能带来的安全问题，2019 年针对深度合成技术特别成立专项研究团队。公司通过对深度合成和生成式 AI 技术的深入研究以及在人工智能安全方向的经验积累，自主研发深度伪造合成视频图像检测鉴定的核心引擎，打造一系列视频图像检测鉴定的一体化智能装备，如“AI-3300 慧眼视频图像鉴真工作站”等。

未来，公司将聚焦生成式人工智能和 AI 生成内容的检测识别技术，按照《中华人民共和国网络安全法》《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》的管理要求，针对深度合成生成文本视频、图像、音频的检测鉴定技术、生成式人工智能技术、人工智能安全取证技术、人工智能大模型安全风险评估等方面的应用需求，不断打磨推出更为安全、可信、可靠的人工智能技术产品，为打击利用人工智能的新型涉网犯罪提供有力的技术武器，为维护和保障人工智能的健康发展和规范应用贡献力量。

深度伪造和合成媒体典型案例：国投智能为公安司法建立视频图像鉴真工作站

【背景介绍】

AI 技术的广泛应用使得物证、书证、视听资料和电子数据面临被“深度伪造”的风险。我国已意识到这一潜在危险，并开始采取措施，同时技术上也在开发识别伪造的工具。执法部门急需能快速判断多媒体信息真伪的智能装备，以打击使用 DeepFake 技术制作的伪造信息。市场上现有的产品在深度伪造鉴定能力上不足，国内司法机关和网安部门在视频图像深伪鉴定技术工作中面临困难。因此，开发专业的视频图像深度伪造检验鉴定装备具有重要的现实意义。

【建设目标】

此方案旨在建立一套视频图像鉴真工作站系统，提供能够对视频图像传统伪造、深度伪造以及 AI 生成手段进行检验鉴定的完整解决方案。支持新建卷宗、智能鉴定、专业鉴定、综合评断、报告导出功能。利用各种方法对视频图像进行原始性、真实性鉴定分析；能够对传统伪造、深度伪造以及 AIGC 生成的视频和图像进行全面分析。覆盖司法鉴定、视频侦查技术实验室、视频图像内容审核等应用场景。

【解决方案】

AI-3300 慧眼视频图像鉴真工作站是一款以人工智能技术为核心的视频图像真伪检验鉴定智能装备。该系统能够识别和分析多种篡改手段，包括传统的 PS 篡改、图像美化、视频帧的抽删或重复等。此外，它还能检

测深度伪造技术，如 AI 换脸、美颜滤镜、伪造溯源以及由 AI 生成的图像和视频。该系统集成了多种合规的视频图像检验鉴定算法，可以有效鉴别数字图像和视频是否符合原始性、真实性特征，为行业客户在视频图像物证化处理中提供符合科学的鉴定支撑。

此方案落地后，可解决公安、司法等技术部门视频图像检验鉴定装备缺乏、视频图像深伪鉴定技术欠缺等问题，助力打击新型网络犯罪。该方案以软硬件一体化的形式进行落地，为用户提供一站式的检验鉴定服务。



图 6.3 1 国投智能慧眼视频图像鉴真工作站产品图

【案例价值】

■ **有利于支撑国家和行业政策落地。**本方案通过提升国家和行业在人工智能和网络安全领域的技术实力，加强了互联网内容的检测与识别能力。利用集中资源，提高了从复杂网络数据中识别不良信息的效率，增强了打击网络犯罪的能力，有效对抗深度伪造技术，减少了其对国家安全的威胁。同时，通过识别虚假信息，增强了视频图像证据在司法中的可信度。

■ **降低深伪技术给国家安全带来的潜在风险。**深度伪造技术可能成为国家安全的新威胁，用于误导国家决策，影响世界秩序。通过 AI 深度伪造形成的“制伪产物”，需要借助于一定媒介进行传播。在这些内容被广泛传播前，如果可以事先筑起一道大坝和筛网，对涉嫌虚假内容进行技术审核，对篡改的图片、音视频进行技术检测鉴定，将有助于大大降低 AI “深度伪造”给国家安全带来的潜在风险。

■ **提高国家对互联网深伪犯罪的打击能力和手段。**深度伪造技术虽能丰富用户体验，但其伪造信息的高可信度和互联网的快速传播能力，使得负面影响急剧扩大。尤其是制作虚假色情视频，不仅严重损害受害者名誉和隐私，还可能引发新型网络犯罪。因此，通过引入深伪鉴定技术，有利于更好地寻找信息的最初来源，研究信息的模式和规律，查证信息的准确性，从而从源头上打击利用深伪技术的网络犯罪活动。深化深伪鉴定技术的研究，有利于提高国家对互联网深伪犯罪的打击能力和手段，对深伪技术进行有效对抗与防御，以专治专，以快制快，让利用深度伪造的网络犯罪从“事后防御”、“事中拦截”，扩展到“事先预警”，从而降低犯罪发生率。

6.4 瑞莱智慧典型案例分析

本案例深度伪造和合成媒体服务提供商：瑞莱智慧

瑞莱智慧 2018 年依托清华大学人工智能研究院发起设立，聚焦人工智能安全和垂类大模型应用，深耕安全、可信、可靠和可拓展的第三代人工智能，通过构建大模型安全基座、AI 伪造内容检测基座等赋能新质生产力建设。在产品研发方面，瑞莱智慧于 2020 年底正式推出深度伪造内容检测平台 DeepReal；2023 年 6 月 DeepReal 2.0 版本正式发布，并升级为生成式人工智能内容检测平台；2024 年 8 月、10 月，DeepReal 一体机、移动版两种全新产品形态先后推出。2022 年 9 月，针对人脸识别系统安全性问题，推出人脸 AI 安全防火墙 RealGuard，可有效防范对抗样本攻击、深度伪造攻击等。2024 年 6 月，瑞莱智慧还推出了全球首个可实时检测 AI 合成内容产品 RealBelieve，中文名“尊嘟假嘟”。

瑞莱智慧长期深耕 AI 生成内容伪造检测算法研究，团队曾荣获 2019、2020GeekPwn AI 安全竞赛、NIPS 2017 人工智能对抗攻击竞赛等多个竞赛冠军，并在此方面获得发明专利共计 10 余项，在 NIPS、ICML 等机器学习顶级学术会议和期刊上发表学术论文数十篇，相关研究成果被诺贝尔奖得主 Hinton 教授、图灵奖得主 Hopcroft 教授在 ICLR2019 等论文中引用，作为对抗攻击的代表性方法，并被主流开源软件收录为标准的对抗攻击算法。

深度伪造和合成媒体典型案例：瑞莱智慧赋能某大型银行人脸识别 AI 欺诈威胁态势感知系统

【背景介绍】

金融和政务等行业面临严峻的新型攻击威胁，已有多起因应对不当而遭受重大损失的案件。中国人民银行、中央网信办等监管机构已在进一步加大监管力度。中国人民银行在《关于加强人脸识别技术应用安全管理的通知》中明确表示：“要强化人脸识别技术的风险评估和管控”，提升深伪检测 and 对抗样本检测能力，并及时更新系统引擎和防攻击算法。因此，进一步加强人脸识别系统的安全防护，才能应对不断增长的深度伪造等攻击威胁，以保护用户资产和敏感信息。

某大型银行人脸识别系统存在安全漏洞，用户手机银行曾出现过被黑灰产业链攻击现象，不法分子利用泄漏的人脸数据，通过深度伪造视频绕过前端活体检测及人脸比对模块，已给用户造成实际经济损失。

【解决方案】

瑞莱智慧人脸识别 AI 欺诈威胁态势感知系统旨在通过“查、打、防”三位一体的 AI 安全体系，提升银行系统的安全性和稳定性。具体解决方案包括：

■ **全面的安全检测：**通过多种手段检测和预防活体攻击和注入攻击，确保人脸识别系统的安全性。

- **AI 安全威胁感知**：实时监控和判断 AI 系统的安全性，及时发现并处理潜在的安全威胁。
- **二次验证机制**：在活体检测通过后，进行二次验证，确保交易的安全性。
- **数据完整性和保密性**：确保数据在传输和存储过程中的完整性和保密性，符合国家相关标准。
- **智能监控和运维管理**：通过智能监控系统和运维管理区，实现对银行业务的全面监控和管理。



图 6.4 1 瑞莱智慧为某银行建立人脸识别 AI 欺诈威胁态势感知系统图

通过这些措施，瑞莱智慧人脸识别 AI 欺诈威胁态势感知系统有效提升银行系统的安全性，保护客户信息，防止金融欺诈，为银行提供稳定可靠的服务。

【方案价值】

根据瑞莱智慧测算，人脸识别 AI 欺诈威胁态势感知系统在该行应用以来，总防护次数超过 38,000 万次，月平均防护次数超过 3,000 万次；总攻击检出次数超过 28,000 万次，月平均检出次数超过 2,300 万次。这些数据表明瑞莱智慧人脸识别 AI 欺诈威胁态势感知系统在安全防护方面有着显著的效果，能够实时发现并拦截对抗样本攻击、伪造视频攻击等新型安全风险。经测算，该系统为该行累计避免直接经济损失达到亿级，在保护数字资产安全方面具有重要作用。

6.5 网易易盾典型案例分析

本案例深度伪造和合成媒体服务提供商：[网易易盾](#)

网易易盾是网易数智旗下一站式数字内容风控品牌，依托网易二十余年的先进技术沉淀和一线实践经验，作为国内领先的数字内容风控服务商，为面向开展数字化业务的客户提供专业可靠的安全服务，涵盖内容安全、业务安全、应用安全、安全专家服务四大领域，全方位保障客户业务合规、稳健和安全运营。

目前，网易易盾已服务数千家行业头部企业，其中不乏 B 站、知乎、网易云音乐、汽车之家、招商银行等知名企事业单位，服务间接覆盖 99% 以上网民。网易易盾将持续挑战科技极限，让科技更安全，用最好的技术和专业服务，共创美好网络空间。

深度伪造和合成媒体典型案例：网易易盾为某城商行提供智能活体检测解决方案

【背景介绍】

国内某城商行金融理财子公司为客户提供一系列的金融产品和服务，包括但不限于股票、债券、基金、保险、养老金、税务筹划等。这些服务通常根据客户的财务状况、风险偏好、时间和财务目标进行个性化定制。国家金融监督管理总局发布的《个人贷款管理办法》正式开始实施，对于不超过 20 万元额度的贷款业务，可以通过远程的数字化途径完成签约；超过 20 万元的贷款业务，借款人必须亲自前往银行进行面对面签约，整个签约过程必须全程录音录像。无论是线上远程完成贷款申请还是线下面对面签约，背后是风险防控与合规需求的双重考量，都需进行全程录音和录像，即完成所谓的“双录”。

【客户挑战】

金融作为高净值行业，黑灰产通过深度伪造的虚假人脸视频，可以登录用户的银行账号，拦截银行发送的手机验证短信，进而进行转账、消费、修改账号密码等。客户在人脸识别方面面临以下挑战：

■ **高度复杂的伪造技术：**不法分子可能使用高级的人工智能技术和深度学习生成对抗网络（GAN）等手段制作逼真的虚假人脸，以欺骗传统的人脸识别系统。

■ **视频攻击：**攻击者可能使用合成的视频素材来模拟真实的人脸动作和表情，以绕过静态图像识别的防护。

■ **活体检测：**部分攻击者可能利用面具、照片或视频等手段来绕过活体检测技术，使得传统的人脸识别技术难以应对。

■ **对抗性攻击：**攻击者可能会使用特制的眼镜、面具或其他物品，以对抗人脸识别系统，使得系统难以准确识别真实的人脸。

为应对这些挑战，需要金融机构不断改进其人脸识别技术，采用先进的活体检测技术、三维人脸识别技术、基于行为的认证等手段，以及建立反欺诈的全面策略，确保人脸识别的准确性和安全性。

【解决方案】

■ 通过 SaaS 服务接口调用的方式快速对接，传入待审核的人脸视频等，实时返回人脸伪造检测结果。

产品能力：双录视频截帧能力，结合视频大小、时长、拍摄时机以及是否有人脸等维度，抽样截取人像照。视频正脸照提取算法，结合人脸旋转角度、人像模糊程度、图片清晰度、光线是否昏暗 / 背光等多维度，选取高清晰的人像照。

■ 网易易盾提出了基于人脸深度评估和注意力机制的活体检测算法（专利名：基于大规模人脸自监督和非对称度量学习的静默活体），通过创新的“光线深度分析”估计真假人脸深度信息并结合注意力机制捕捉假脸细微痕迹进行识别，通过领先的“真人认知人脸模型”，将各种假体攻击进行精准防御。伪造人脸检测能力全覆盖，具体包含：疑似深度伪造、疑似合成攻击、疑似黑产攻击、疑似生成人脸、疑似翻拍人脸等。

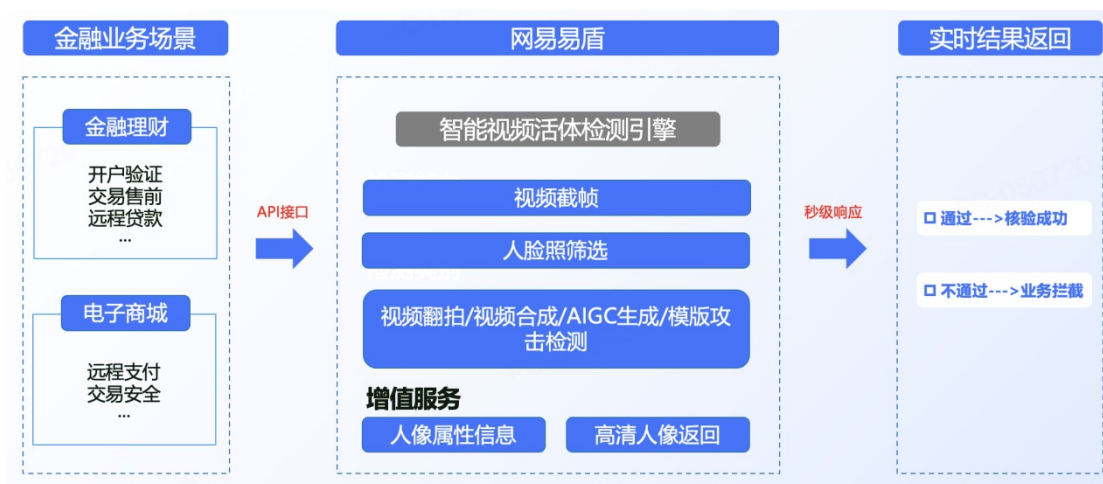


图 6.5 1 网易易盾智能活体检测解决方案架构图

【方案价值】

部署前：对于新型 AIGC 生产的合成脸，常规活体检测算法漏报率较高，对于同背景不同人的情况，只能通过人审的方式去发现，效率相当低下，而且对于高阶的合成脸对抗，人审也存在一定的漏判。视频活体，因为是纯 SaaS 服务接口调用的方式对接，没有 SDK 集成，所以也较难通过设备环境信息来检测端侧的风险，这就对活体检测算法提出了更高的挑战。

● 国内深度伪造和合成媒体安全厂商典型案例（按简称字母排序）

部署后：

- **安全性提升：**减少身份盗用和欺诈风险，保护客户和用户的个人信息安全。
- **信任度提高：**增强客户对平台或产品的信任，促进用户参与交易的积极性。
- **遵守法规：**满足相关法规和合规要求，降低因安全漏洞而引发的法律责任和罚款风险。
- **用户体验改善：**提高身份验证的准确性和效率，简化注册和登录流程，提升用户体验。
- **品牌价值提升：**对于企业来说，这种技术的应用可以提升品牌形象，增加市场竞争力。

6.6 中科睿鉴典型案例分析

本案例深度伪造和合成媒体服务提供商：中科睿鉴

中科睿鉴科技有限公司是伪造检测领域的国家队，致力于运用 AI 技术赋能数字内容安全。公司以“用技术让世界更可信”为使命，深耕 AI 生成检测、虚假信息识别、深度伪造检测、深度伪造溯源、大模型内容安全审核等技术的研发，在公共安全监管、媒体内容审核、金融风控管理、教育出版诚信审核等关键业务场景中，率先实现文本、图像、视频、音频生成和伪造检测全技术栈布局，有效推动人工智能时代数字内容安全技术的落地与应用。公司的技术团队迄今拥有 20 年的内容安全科研积累，在行业内处于领先地位。

中科睿鉴基于超过 200TB 的伪造数据训练，自研国内首个多模态伪造检测基模型，有效应对新型 AIGC 技术发展挑战，检测性能行业领先；2020 年，公司获得了北京市科技进步奖一等奖，是国内首个伪造检测领域的省部级奖项；2021 年，公司研发的国内首款软硬一体深伪检测专用设备“睿安”，在重点行业大规模部署，调用超千亿次，成为国内最早经过大规模验证的伪造检测产品。2022 年，公司获得世界互联网大会领先科技成果奖全球 TOP15，同年亦获得中国人工智能大赛“创新之星”等多项科技奖励；2023 年公司发布国内首个伪造检测领域大模型—睿鉴图灵，入选《北京市人工智能行业大模型创新应用白皮书（2023 年）》大模型创新应用案例；2024 年，公司的深伪内容检测技术成功应用在运营商的流量端口和消费者的终端上，并发布了国内首个终端 AI 鉴伪大师，为行业和个人提供 AI 时代的可信通信保障。

深度伪造和合成媒体典型案例：中科睿鉴联手终端厂商共创可信通信方案

【背景介绍】

AI 生成合成技术的工具化和普及化使得用户面临大量来自公域和私域的伪造内容挑战。技术的恶意使用给用户的工作和生活带来财产、隐私等各方面的安全威胁。手机、电脑和平板这类终端设备作为工作、生活信息最终的展示平台，目前已在多种场景中产生了深伪内容检测的诉求，如线上视频会议人员识别、AI 换脸视频通信防范、互联网信息鉴真、婚恋交友平台身份辨别、虚假招聘信息核实等。

为应对这些问题，中科睿鉴同终端头部厂商建立合作，专注于为用户提供创新、安全、可靠的智能设备体验。与此同时，公司在隐私保护和内容安全方面持续投入，致力于为用户提供高水准的信息安全保障，建设手机端 AI 生成、AI 合成内容检测能力，在系统层和应用层共同保障用户安全。

【解决方案】

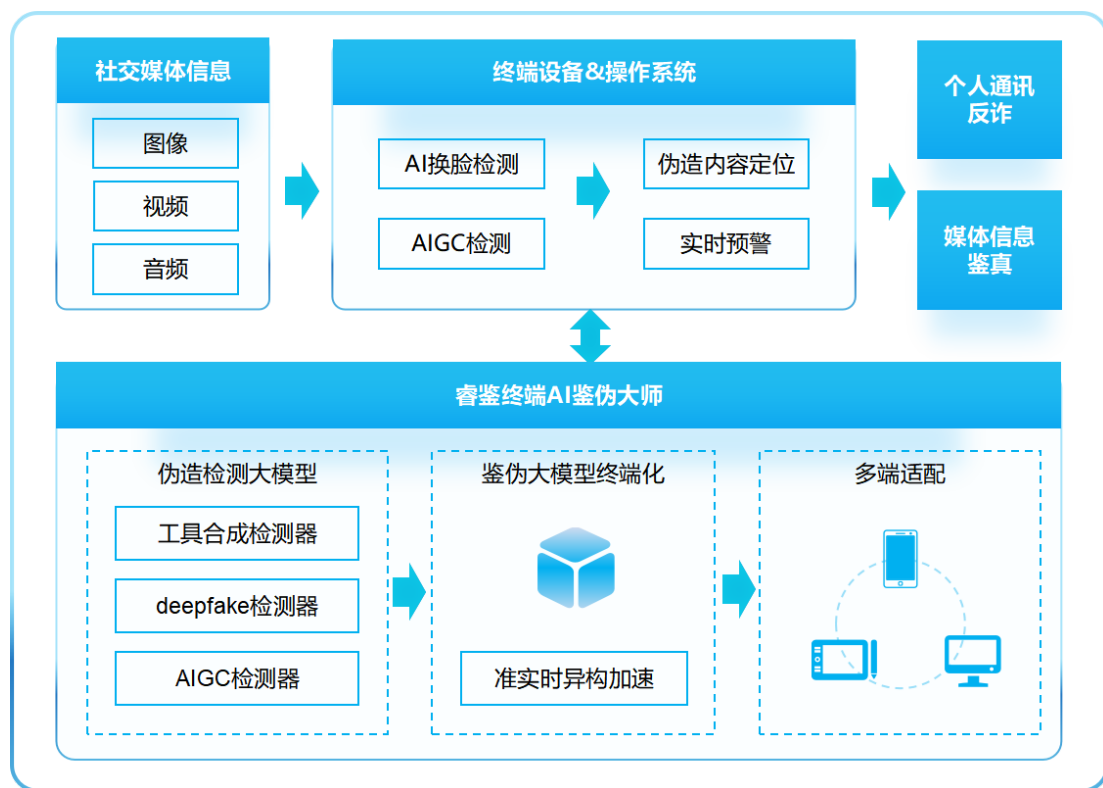


图 6.6 1 中科睿鉴解决方案架构图

终端场景对于检测算法的响应实时性和检测准确度有极高要求，除此之外，部分私域通信场景需要采用本地化运行的终端检测算法来保证用户的隐私安全。中科睿鉴构建的 AI 鉴伪底座模型经过千亿次实战调用打磨，实现了复杂工况下快速、高精度的检出效果。同时针对终端设备内存及算力资源有限的特点，中科睿鉴采用多芯异构、量化加速等核心技术，将鉴伪大模型装进终端设备。当前的终端综合方案具有高度的可扩展性，支持用户基于业务需求灵活选择部署方案，能够根据不同的设备环境与应用场景进行定制优化。

【方案亮点】

- **高精度检测能力：**覆盖主流超 200 种生成合成算法及应用，支持 AI 生成、AI 换脸、AI 换声等多种伪造内容检测，实战准确率 90% 以上；
- **精准实时计算响应：**准实时检测，最快达 20ms 响应，确保用户无感检测；
- **低开销资源占用：**采用蒸馏 + 量化 + 多芯异构 + 算子优化技术实现资源最优使用；
- **跨平台软硬件适配和集成开发语言支持：**1) 操作系统: Android、IOS、Windows、Linux、MacOS 等; 2) 芯片支持: arm64、armeabi、x86、x86_64 等指令集架构; 高通、华为、联发科、苹果、英特尔等厂家芯片;

3) 开发语言：C/C++、Java/Kotlin 等。

【案例价值】

中科睿鉴和荣耀终端共同打造的 AI 生成 / 合成内容检测功能已在其最新推出的操作系统 MagicOS 9.0 中成功上线，针对人物换脸、AI 生成内容实现检测算法系统级快速调用，这标志着荣耀成为国内首批为用户提供此类服务的终端厂商。作为中国前五的智能手机厂商，荣耀具有庞大的用户基础，树立了智能手机行业的新标杆，加速了全类型终端设备鉴伪技术的应用。

未来，具备伪造内容检测能力的终端设备将逐渐成为业内普遍采用的标准。除此之外，在一些对于伪造内容危害更为敏感的领域如金融、保险行业，能够运行在终端设备上的鉴伪算法亦可以为相关业务场景进行赋能，为身份识别场景融入新的安全保障。中科睿鉴将继续联合广大企业，持续进行检测技术的更新迭代以及鉴伪算法终端化的能力构建，为广大民众营造一个内容真实、隐私安全的网络环境。

第七章 AI 深度伪造和合成媒体安全市场未来五大趋势分析

7.1 深度伪造和合成媒体安全潜在市场规模巨大

安全牛认为，深度伪造和合成媒体安全潜在市场规模巨大，短期来看（3 年），B 端（主要是监管方、企业和机构）和 C 端（消费者）的联合市场规模可达到 30 亿元。虽然目前市场正处于初期阶段，但产业发展迅速，预计 2026 年将会迎来“爆发”。根据安全牛公开调研和访谈推算，2023 年深度伪造和合成媒体安全市场规模为 1.2 亿元，2024 年为 4.1 亿元，同比增长超过 200%，预计 2025 年，随着深度伪造和合成媒体技术合成的内容指数级增长，安全市场会在滞后一段时间后有所反应，预计在 2026 年迎来“爆发”点。**需要注意的是，深度伪造和合成媒体市场不同于其他市场，它的增长是超越常规的，各方需抓紧布局。**

深度伪造和合成媒体安全市场规模预测 2023-2027E

单位：亿元

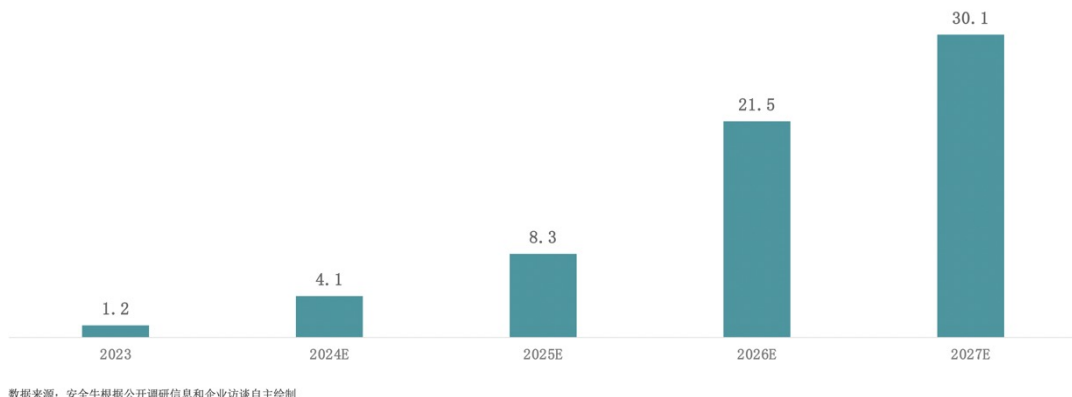


图 7.1.1 深度伪造和合成媒体安全市场规模预测图（2023-2027E）

随着深度伪造技术的快速发展，B 端市场对相关安全解决方案的需求日益增长。监管机构加强了对深度伪造技术的监管，要求企业遵守法律法规，这推动了市场对安全解决方案的需求。金融机构、社交媒体平台和在线服务提供商等都需要有效手段来防范身份盗用和虚假信息传播。企业应用场景包括金融欺诈检测、版权保护、法律合规、网络安全和媒体检测等。安全牛调研显示，B 端市场的需求主要集中在监管侧，尤其是各部委、公检法等执法机关。这些机构需要先进的检测工具来识别和打击深度伪造犯罪活动，目前监管侧的检测技术还在不断提升完善中。

在 C 端市场，消费者对个人信息安全和隐私保护的意识正在提高，对深度伪造技术的风险认知逐步深刻。但消费者为潜在风险支付费用的意愿仍需培养。市场上对个人身份验证、内容真实性检测的需求正在上升，但 C 端用户付费的商业模式还需关键推动因素。目前已经出现的案例如荣耀手机，在其更新的操作系统中加入内

容检测功能，但消费者是否因此原因进行付费购买还需进一步调研。

7.2 未来三年，企业和大众将会深刻感受到深度伪造和合成媒体安全带来的挑战

安全牛认为，未来三年，企业和大众将会更加深刻地感受到深度伪造和合成媒体安全带来的挑战，这种判断主要基于以下几方面：

■ **数据量激增与伪造内容的飞速扩散：**根据 Gartner 的预测，到 2025 年，生成式人工智能（Generative AI）生成的数据将占有所有数据的 10%。这意味着，随着技术的进步和应用的普及，基于人工智能生成的伪造内容将大幅增加，这种体量的爆发是指数级的增长，伴随而来的安全问题将层出不穷。虽然我国属于强监管，但不难预计，指数级的增长将带来深刻变革。

■ **社会影响和经济损失加重：**2024 年，深度伪造技术的消极应用已经导致公民损失的增加，名人名誉受损、普通百姓遭遇诈骗以及企业被骗等事件层出不穷。这些事件不仅对个人和企业造成了直接的经济损失，还引发更广泛的社会信任危机，影响社会稳定和经济平衡。

■ **技术发展存在滞后性，带来越来越多的挑战：**技术的发展相对生成内容的层出不穷是滞后的。深度伪造技术的发展带来了一系列法律和伦理问题，包括版权、隐私权、言论自由等。

综上所述，企业和大众必须提高对深度伪造和合成媒体安全的重视程度，防患于未然，以积极思考的态度应对这些技术带来的挑战。

7.3 技术创新和应用创新双轮驱动，推动深度伪造和合成媒体安全行业快速发展

安全牛认为，当前深度伪造和合成媒体的安全检测领域正展现出技术创新与应用创新双轮驱动的积极性发展态势。通过分析中国知网“深度伪造”和“深度合成”等关键词的年度论文数量分布情况，安全牛发现，自 2018 年至 2023 年，深度伪造相关研究论文数量持续增长，并在 2023 年达到 234 篇的高峰。尽管 2024 年的数据仅涵盖前 10 个月，但已记录有 147 篇论文，这表明该领域的理论研究基础日益坚实，开始从学术更多转向具体应用。

AI 深度伪造和合成媒体安全市场未来五大趋势分析

在技术实践方面，以人工智能为主导的安全服务提供商正致力于将深度伪造和合成媒体的安全技术转化为实际应用。他们采用贝叶斯深度学习、多特征融合以及多任务学习等先进方法进行系统设计和优化，同时紧密结合客户端的实际需求，推动技术的落地实施。与此同时，以网络安全为主导的安全服务提供商则更加专注于特定领域的需求，针对这些垂直领域提供定制化的解决方案，以满足市场的需求。

中国知网搜索“深度伪造”、“深度合成”等关键词论文数量年度分布图 2018-2024（1-10月）

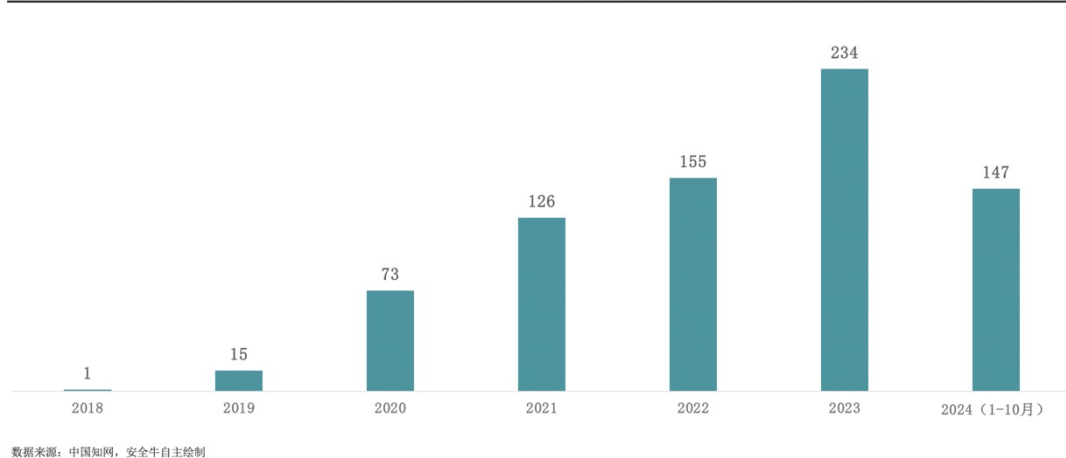


图 7.3 1 中国知网搜索“深度伪造”、“深度合成”等相关关键词论文数量年度分布图

7.4 国内深度伪造相关技术短期领先，长期发展需全生态协同推进

安全牛认为，短期来看，国内在深度伪造和合成媒体检测技术方面的优势主要得益于几个关键因素。首先，监管侧的较早布局及力度加强为技术研发和应用提供了明确的指导和规范，这有助于技术的健康发展。其次，相对充足的经费支持使得研究机构和企业能够投入更多资源进行技术开发和创新。此外，国内丰富的数据资源为大类识别和多任务学习提供了良好的基础，这在一定程度上推动了检测技术的进步。国外在学术研究方面产出比国内多、质量高，但从具体落地检测结果来看，国外的一些检测技术在实际应用时检测准确率相对较低。这表明在将研究成果转化为实际应用方面，国外可能还需要进一步加强。

然而，从长期来看，要维持这一优势，还需要整个生态系统的持续努力。这包括但不限于加强国际合作，吸收全球智慧，共同应对这一全球性的挑战。同时，需要不断推动技术创新，提高检测技术的准确性和效率。跨学科合作也是关键，不同领域知识的融合可以促进技术的发展。政策支持和人才培养同样不可忽视，这是保证技术持续发展的重要因素。

综上所述，国内在深度伪造和合成媒体检测技术方面虽然短期内具有优势，但长期发展仍需不断努力，包括技术创新、跨学科合作、政策支持、人才培养以及国际合作等多个方面。同时，国外在学术研究方面的优势也值得国内借鉴和学习，以实现技术的持续进步和应用的广泛落地。

7.5 “融合”成深度伪造和合成媒体安全关键词

安全牛认为，“融合”是未来深度伪造和合成媒体安全发展的关键词，这不仅涉及技术层面的融合，也包括产业链和各组织之间的生态融合。在技术层面，多模态融合是一个重要的方向，它涉及从文本、图像、语音、视频等多个领域获取信息，并实现信息的转换和融合，以提升模型性能。这种融合技术的发展，不仅可以提高深度伪造的检测能力，也能增强合成媒体的安全性。在生态融合方面，产业链的积极构建和政府、企业、学术机构和民间组织之间的合作也是关键。值得一提的是，在构建融合共治的生态系统中，创新能力、学习能力、适应能力和组织能力至关重要。



图 7.5 1 融合与深度伪造和媒体合成安全生态与技术的关系

第八章 附录

8.1 本报告主要参考文献

- [1] 杨睿, 胡心如, 黄卓超, 张玉书, 蓝如师, 邓珍荣, 罗笑南. 深度网络生成式伪造人脸检测方法研究综述 [J/OL]. 计算机辅助设计与图形学学报.
- [2] 杨宏宇, 李星航, 胡泽. 深度伪造人脸生成与检测技术综述 [J/OL]. 华中科技大学学报 (自然科学版).
- [3] 董佳乐, 邓正杰, 张宝, 李喜艳. 基于人脸关键区域的深度伪造视频检测方法 [J/OL]. 计算机技术与发展.
- [4] 周文柏, 张卫明, 俞能海, 等. 人脸视频深度伪造与防御技术综述 [J]. 信号处理, 2021, 37(12): 2338-2355.
- [5] 王禄生. 论“深度伪造”智能技术的一体化规制 [J]. 东方法学, 2019(06): 58-68.
- [6] 段伟文. 人工智能时代的价值审度与伦理调适 [J]. 中国人民大学学报, 2017, 31(06): 98-108.
- [7] 邓悦, 许弘楷, 王诗菲. 人工智能风险治理: 模式、工具与策略 [J]. 改革, 2024(1): 144-158.
- [8] 赖志茂, 章云, 李东. 基于 Transformer 的人脸深度伪造检测技术综述 [J]. 广东工业大学学报, 2023, 40(6): 155-167.
- [9] 池珠, 王永超, 刘丽, 等. “目标—议题—工具”框架下人工智能国家政策解析及发展建议: 基于近十年政策文本镜像扫描 [J]. 山东师范大学学报 (自然科学版), 2023, 38(3): 221-234.
- [10] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[EB/OL]. [2023-08-30]. <http://arxiv.org/abs/1406.2661>.
- [11] Karras T, Aila T, Laine S, et al. Progressive growing of GANs for improved quality, stability, and variation[EB/OL]. [2023-08-03]. <http://arxiv.org/abs/1710.10196>.
- [12] Karras T, Laine S, Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(12): 4217-4228.
- [13] Rossler A, Cozzolino D, Verdoliva L, et al. FaceForensics++: Learning to Detect Manipulated Facial Images[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019: 1-11.
- [14] Dolhansky B, Bitton J, Pflaum B, et al. The DeepFake Detection Challenge (DFDC) Dataset[EB/OL]. [2023-08-07]. <http://arxiv.org/abs/2006.07397>.

- [15]Li Y, Chang M C, Lyu S. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking[C]//2018 IEEE International Workshop on Information Forensics and Security (WIFS). Hong Kong: IEEE, 2018: 1-7.
- [16]Huang G B, Mattar M, Berg T, et al. Labeled faces in the wild: A database for studying face recognition in unconstrained environments[C]//Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition. 2008.
- [17]Liu Z, Luo P, Wang X, et al. Deep Learning Face Attributes in the Wild[C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015: 3730-3738.
- [18]CHOI Y, CHOI M, KIM M et al. StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT: IEEE, 2018: 8789-8797.
- [19]Du C, Chen Q, He T, et al. Dae-talker: High fidelity speech-driven talking face generation with diffusion autoencoder[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023: 4281- 4289.
- [20]Deepfakes. Deepfakes github [EB/OL]. <http://github.com/Deepfakes/faceswap>, 2017. Accessed 2020-08-18. 2, 3, 5, 6.
- [21]Facebook, Kaggle Deepfake detection challenge [EB / OL]. Retrieved from <https://www.kaggle.com/c/Deepfake-detection-challenge>, 2020-04-01.

8.2 国内深度伪造相关专利信息汇总表

表 8.2 1 中国深度伪造相关专利信息汇总表 2021-2024

申请号/专利号	发明名称	申请人	申请日	发明专利申请公布号	法律状态
202411060517X	基于可分离感知哈希增强的深度伪造主动取证方法	齐鲁工业大学（山东省科学院）	2024/8/5	CN118587568A	案件状态：等年登印费
2024108549788	一种基于重建学习的深度伪造人脸取证方法	北京交通大学	2024/6/28	CN118781665A	案件状态：等待实审提案
2024108343145	一种基于自适应学习的高鲁棒性深度伪造检测方法	支付宝（杭州）信息技术有限公司	2024/6/26	CN118379608A	专利权有效
2024107729872	深度伪造鉴别的人工智能治理方法和机器人	华南师范大学	2024/6/17	CN118797404A	案件状态：等待实审提案
2024107576773	一种基于语义解纠缠的深度伪造图像检测方法	南昌大学	2024/6/13	CN118334473A	专利权有效
2024107148304	基于相机指纹和自注意力机制融合的深度伪造检测方法	电子科技大学	2024/6/4	CN118537651A	案件状态：等待实审提案
2024106823097	人脸深度伪造取证方法、系统、设备及存储介质	中国科学技术大学	2024/5/29	CN118279995A	案件状态：撤回专利申请
2024106639614	一种基于多尺度多注意力的深度伪造视频鉴别方法	北京国信鸿思科技有限公司	2024/5/27	CN118485944A	案件状态：等待实审提案
2024106368491	基于伪影域对抗学习的深度伪造图像检测方法及装置	华侨大学	2024/5/22	CN118469968A	案件状态：等待实审提案
2024105820960	一种训练方法、深度伪造鉴别方法及电子设备	北京智慧易科技有限公司	2024/5/11	CN118427755A	案件状态：等待实审提案
2024105669330	基于无监督域自适应的深度伪造检测方法和系统	中国科学院计算技术研究所	2024/5/9	CN118397440A	案件状态：等待实审提案
202410486972X	深度伪造音频检测模型训练方法、电子设备和存储介质	思必驰科技股份有限公司	2024/4/22	CN118380015A	案件状态：等待实审提案
2024104338504	移动端深度伪造内容攻击防护方法	杭州中科普鉴科技有限公司	2024/4/11	CN118230134A	案件状态：等待实审提案
2024104323937	一种基于鲁棒身份感知水印的深度伪造检测检测方法	齐鲁工业大学（山东省科学院）	2024/4/11	CN118691452A	案件状态：等待实审提案
2024104014026	一种基于多域融合的深度伪造检测方法	西北大学	2024/4/3	CN118470585A	案件状态：等待实审提案
2024103594880	基于频率增强自注意力的深度伪造图像处理方法和系统	中国信息通信研究院	2024/3/27	CN118115481A	案件状态：等待实审提案
2024103459643	基于噪声感知的可泛化深度伪造图像检测方法及系统	齐鲁工业大学（山东省科学院）	2024/3/26	CN118196865A	案件状态：等待实审提案
2024103325704	一种基于设备指纹的深度伪造视频检测方法	杭州电子科技大学	2024/3/22	CN118230141A	案件状态：等待实审提案
2024103001122	一种防御深度伪造的人脸隐私保护方法	中国科学院信息工程研究所	2024/3/15	CN118194342A	案件状态：等待实审提案
2024102991337	语音转换深度伪造中的源说话人溯源方法、设备及存储介质	中国人民解放军陆军工程大学	2024/3/15	CN118197330A	案件状态：等待实审提案
2024102799099	一种针对深度伪造的可迁移对抗样本生成方法	重庆邮电大学	2024/3/12	CN118298272A	案件状态：等待实审提案
2024102794220	基于深度金字塔聚合网络的人脸深度伪造检测方法	重庆移通学院	2024/3/12	CN118196912A	案件状态：等待实审提案
2024102307179	一种基于深度学习的社交网络深度伪造视频检测方法及系统	西安理工大学	2024/2/29	CN118097502A	案件状态：等待实审提案
2024102172218	一种深度伪造模型的比对方法、装置及存储介质	厦门市美亚柏科信息安全研究所有限公司	2024/2/28	CN118133037A	案件状态：等待实审提案
202410176933X	远程身份认证场景下的细粒度深度伪造人脸身份检测方法	复旦大学	2024/2/8	CN118172810A	案件状态：等待实审提案
2024101608732	一种基于人脸特征及其关联水印的深度伪造主动检测方法	北京理工大学	2024/2/5	CN117975578A	案件状态：一通出案待答复
2024101448273	一种基于面部动态集成的深度伪造检测方法及系统	南京邮电大学	2024/1/31	CN117975577A	案件状态：等待实审提案
2024101117020	一种基于图结构的深度伪造模型特征提取方法及其系统	厦门市美亚柏科信息安全研究所有限公司	2024/1/26	CN118035701A	案件状态：等待实审提案
2023118486070	一种基于伪影检测的深度伪造视频检测方法	四川新网银行股份有限公司	2023/12/29	CN118015713A	案件状态：等待实审提案
2023118412062	一种可学习局部差异的深度伪造人脸检测定位方法	暨南大学	2023/12/29	CN117496583A	专利权有效
2023118279808	深度伪造视频检测方法、装置、设备及存储介质	中邮消费金融有限公司	2023/12/27	CN117690061A	专利权有效
2023118159825	一种基于视听联合学习的深度伪造检测方法	中国电子科技集团公司第十研究所	2023/12/26	CN117789084A	案件状态：等待实审提案
2023118131321	基于信息瓶颈的深度伪造图像检测方法、系统及装置	浙江大学杭州国际科创中心	2023/12/26	CN117877129A	案件状态：等待实审提案
2023118049149	一种基于自增强的深度伪造图像鉴别方法和系统	中国信息通信研究院	2023/12/26	CN117765620A	专利权有效
2023117971100	基于频域-时域知识蒸馏的低质量压缩语音深度伪造检测方法	大连理工大学	2023/12/26	CN118016046A	案件状态：等待实审提案
2023117617135	一种基于像素级别人脸修正预训练的深度伪造检测方法	北京邮电大学	2023/12/20	CN117953563A	案件状态：等待实审提案
2023117599476	一种基于三维时空轨迹的深度伪造视频检测方法	湖南大学	2023/12/20	CN117727083A	案件状态：等待实审提案
2023117573762	一种基于面部关键点光流的深度伪造视频检测方法	上海应用技术大学	2023/12/20	CN117746514A	案件状态：等待实审提案
2023116985888	一种基于深度伪造产物检测模型的检测方法及其系统	南昌大学	2023/12/12	CN117690118A	
2023117005887	一种基于生成对抗网络的深度伪造人脸检测方法	天翼云科技有限公司	2023/12/12	CN117831137A	案件状态：等待实审提案
2023116415649	一种基于多特征融合的深度伪造图像检测模型	新型显示与视觉感知石城实验室	2023/12/4	CN117726912A	案件状态：等待实审提案
2023116297568	开集场景下防深度伪造的讲话人认证方法及系统	上海交通大学	2023/11/30	CN117558064A	案件状态：等待实审提案
202311546911X	一种基于身份型特征的深度伪造检测方法	齐鲁工业大学（山东省科学院）	2023/11/20	CN117315798A	专利权有效
2023115019476	基于安全码的深度伪造视频检测方法、装置和电子设备	杭州中奥科技有限公司	2023/11/10	CN117557440A	案件状态：等待实审提案
2023114369025	基于时空特征的社交网络深度伪造视频检测方法及系统	西安理工大学	2023/10/31	CN117523439A	案件状态：等待实审提案
2023114351568	一种音频辅助的深度伪造人脸视频检测方法、系统及设备	重庆邮电大学	2023/10/31	CN117437935A	案件状态：等待实审提案
2023114349981	面向低质量人脸图像的深度伪造检测方法、系统和设备	重庆邮电大学	2023/10/31	CN117423148A	案件状态：等待实审提案

2023108042035	面向深度伪造音频的多模态融合检测方法	南京理工大学	2023/7/3	CN116797896A	案件状态：等待实审提案
2023107452422	基于空频特征融合的知识蒸馏深度伪造检测方法	大连理工大学	2023/6/21	CN116895100A	案件状态：等待实审提案
2023106454632	深度伪造图像检测方法和装置	马上消费金融股份有限公司	2023/6/1	CN117495766A	案件状态：等待实审提案
2023106261998	基于生物信号和时空信号的深度伪造视频检测方法	南京理工大学	2023/5/30	CN116630858A	案件状态：等待实审提案
2023106182835	深度伪造图像检测方法和装置	马上消费金融股份有限公司	2023/5/29	CN117496212A	案件状态：等待实审提案
2023106086185	一种深度伪造人脸图像检测方法	支付宝（杭州）信息技术有限公司	2023/5/26	CN116758643A	案件状态：等待实审提案
2023105098857	基于三维时空网络的深度伪造视频检测方法	广州大学	2023/5/8	CN116665089A	专利权有效
2023104645851	一种基于图像重构的人脸深度伪造检测系统	北京航空航天大学	2023/4/26	CN116503927A	案件状态：等待实审提案
2023104587273	基于深度伪造检测的深度学习推理加速方法、装置	中国科学院计算技术研究所	2023/4/25	CN116645733A	案件状态：等待实审提案
2023104655726	基于聚类决策的深度伪造图像检测方法及相关设备	哈尔滨工业大学（深圳）	2023/4/20	CN116580462A	案件状态：等待实审提案
2023104188131	一种基于人脸几何关系推理的深度伪造检测方法	四川大学	2023/4/18	CN116758604A	案件状态：等待实审提案
202310240293X	一种视频深度伪造检测方法	中国电子科技集团公司第三十研究所	2023/3/13	CN116778545A	案件状态：等待实审提案
2023101964183	一种基于多尺度特征解耦的深度伪造视频检测方法和装置	巧智绘科技（浙江）有限公司	2023/3/3	CN116310969A	案件状态：等待实审提案
2023101948091	一种深度伪造指纹检测模型的版权保护方法、装置及介质	南京信息工程大学	2023/3/3	CN115879072A	专利权有效
202310172633X	防深度伪造的人脸隐私保护方法及设备	武汉大学	2023/2/23	CN116305258A	案件状态：等待实审提案
2023101502818	基于图像多样化特征深度伪造检测方法	华南理工大学	2023/2/22	CN116311430A	案件状态：等待实审提案
2023100937738	度伪造人脸鉴别的方法、系统、装置、处理器及其计算机可读存储介质	华东理工大学	2023/2/10	CN116012958A	案件状态：等待实审提案
2023100472779	一种深度伪造人脸图像检测的方法及相关设备	湖南大学	2023/1/31	CN116188956A	案件状态：等待实审提案
2023100891949	跨域的深度伪造检测模型训练方法、深度伪造检测方法和系统	支付宝（杭州）信息技术有限公司	2023/1/17	CN116385817A	案件状态：等待实审提案
2022116789981	一种深度伪造视频检测方法、装置、设备及介质	中国科学院信息工程研究所	2022/12/26	CN116310937A	案件状态：等待实审提案
2022116704339	一种基于面部肌肉运动的压缩深度伪造视频检测方法	湖南大学	2022/12/25	CN115984740A	案件状态：等待实审提案
202211639132X	深度伪造视频检测分割识别系统、终端及存储介质	浙江大学	2022/12/20	CN115909172A	案件状态：等待实审提案
2022115597128	深度伪造换脸图像生成方法、系统、设备及存储介质	西安交通大学	2022/12/6	CN116092153A	案件状态：等待实审提案
2022115339855	基于跨领域自适应的深度伪造语音检测方法	浙江大学嘉兴研究院	2022/12/1	CN116153331A	案件状态：等待实审提案
2022115097979	一种基于多重拉姆理论的人脸深度伪造检测装置及方法	北京航空航天大学	2022/11/29	CN115909459A	案件状态：等待实审提案
2022113579623	一种基于细粒度特征深度伪造图像检测方法和系统	厦门市美亚柏科信息股份有限公司	2022/11/1	CN115690501A	案件状态：等待实审提案
2022113579394	双注意力机制高频信息感知的深度伪造检测方法和系统	厦门市美亚柏科信息股份有限公司	2022/11/1	CN116051941A	案件状态：等待实审提案
2022113540093	一种高鲁棒性的深度伪造人脸检测方法	重庆邮电大学	2022/11/1	CN115588226A	案件状态：等待实审提案
2022112754663	基于频域变换的深度伪造图像检测方法和系统	支付宝（杭州）信息技术有限公司	2022/10/18	CN115546598A	案件状态：等待实审提案
2022112245682	一种基于面部语义内容分解领域泛化框架的深度伪造检测算法	四川大学	2022/10/9	CN116343279A	案件状态：等待实审提案
2022112201171	深度伪造检测方法	阿里巴巴（中国）有限公司	2022/10/8	CN115311525A	专利权有效
2023113049388	基于小波变换和时序特征提取的深度伪造视频检测方法	中山大学	2023/10/10	CN117496393A	案件状态：等待实审提案
2023113027181	基于时序差异的深度伪造视频检测方法	中山大学	2023/10/10	CN117496392A	案件状态：等待实审提案
2023112731030	一种基于异常模拟和噪声抑制的异常深度伪造检测方法	华东师范大学	2023/9/28	CN117351541A	案件状态：等待实审提案
2023112685314	人脸视频深度伪造检测方法	中国科学院自动化研究所	2023/9/27	CN117542124A	案件状态：等待实审提案
2023112531316	一种深度伪造音频鉴别方法和系统	中国信息通信研究院	2023/9/27	CN116994590A	专利权有效
2023112207929	一种引入无监督约束的深度伪造图像检测方法和系统	厦门市美亚柏科信息安全研究所有限公司	2023/9/21	CN117333760A	案件状态：等待实审提案
202311211903X	基于人脸时序信息的人脸视频深度伪造检测方法	中国科学院计算技术研究所	2023/9/19	CN117275064A	案件状态：等待实审提案
202311211893X	基于三维信息时序一致性的人脸深度伪造检测方法	中国科学院计算技术研究所	2023/9/19	CN117275063A	案件状态：等待实审提案
2023112066691	深度伪造视频检测方法、装置、电子设备	中国电信股份有限公司技术创新中心	2023/9/18	CN117274861A	案件状态：等待实审提案
2023111677857	人脸图像的深度伪造检测方法、系统、设备及介质	西安交通大学	2023/9/11	CN117238041A	案件状态：等待实审提案
2023111398280	深度伪造视频检测方法、装置、电子设备	中国电信股份有限公司技术创新中心	2023/9/5	CN117173611A	案件状态：等待实审提案
2023110840457	一种基于生成模型的通用深度伪造检测方法	浙江大学	2023/8/28	CN117238015A	专利权有效
2023110703977	一种面向音频多模态深度伪造内容的时序伪造定位方法	四川大学	2023/8/24	CN117079115A	案件状态：等待实审提案
2023110413476	一种基于适配器的特定身份人脸深度伪造检测方法	北京浩瀚深度信息技术股份有限公司	2023/8/17	CN117058107A	专利权有效
2023109500134	深度伪造反取证图像的生成方法	南昌大学	2023/7/31	CN116664880A	专利权有效
2023109293007	一种基于属性引导的深度伪造人脸图像检测方法和系统	厦门市美亚柏科信息股份有限公司	2023/7/27	CN117173761A	案件状态：等待实审提案
2023109272299	基于新型神经网络架构的深度伪造视频检测方法	中国人民解放军国防科技大学	2023/7/26	CN117115627A	案件状态：等待实审提案
2023109237115	一种基于时空注意力引导融合的深度伪造检测方法	华南农业大学	2023/7/26	CN117238011A	案件状态：等待实审提案
2023109214024	一种深度伪造视频数据增强方法、装置及计算机设备	中国人民解放军国防科技大学	2023/7/26	CN116645299A	专利权有效
2023108654807	深度伪造视频的时空结合检测方法、装置及设备	中国科学院软件研究所	2023/7/14	CN116994175A	案件状态：等待实审提案
2023108371704	一种基于噪声不一致性的深度伪造检测分类和定位方法	华中科技大学	2023/7/10	CN117079354A	案件状态：等待实审提案

2022112114811	一种基于多尺度监督的深度伪造人脸图像检测方法和系统	厦门华厦学院	2022/9/30	CN115909510A	案件状态：等待实审提案
2022112060173	深度伪造视频检测方法、装置及存储介质	北京瑞莱智慧科技有限公司	2022/9/30	CN116645710A	案件状态：等待实审提案
202211185641X	一种基于音视频多模态的特定人物深度伪造检测方法	北京邮电大学	2022/9/27	CN115661889A	案件状态：等待实审提案
2022111593965	一种基于多模态仿影的人脸深度伪造检测方法及系统	云南大学	2022/9/22	CN115984917A	案件状态：等待实审提案
2022111536229	基于像素级注意力机制的深度伪造图像检测方法和系统	厦门市美亚柏科信息股份有限公司	2022/9/21	CN115424036A	案件状态：等待实审提案
2022109613138	对抗条件下深度伪造视频检测方法	中国人民解放军战略支援部队信息工程大学	2022/8/11	CN115410118A	案件状态：等待实审提案
202210931430X	基于身份不变特征深度伪造人脸检测方法、系统及设备	武汉大学	2022/8/4	CN115457622A	案件状态：等待实审提案
2022108874766	基于高阶张量型神经元的深度伪造溯源检测方法和系统	厦门市美亚柏科信息安全研究所有限公司	2022/7/26	CN115711194A	案件状态：等待实审提案
2022108458455	一种针对深度伪造的主动防御方法、系统	北京大学	2022/7/19	CN115273247A	案件状态：等待实审提案
2022108436070	基于图像特征融合的深度伪造人脸视频检测方法	中国人民警察大学	2022/7/18	CN115273186A	案件状态：驳回失效
202210796847X	针对深度伪造的双向人脸数据保护方法、系统及设备	中国科学技术大学	2022/7/7	CN114881838A	专利权有效
2022107753641	一种基于深度学习的人脸深度伪造鉴定方法及相关设备	公安部第三研究所	2022/7/1	CN115063868A	案件状态：一通回案实审
2022107189733	基于帧内帧间特征差异化的深度伪造视频检测方法	山东大学	2022/6/23	CN115147758A	案件状态：等待实审提案
2022107040521	一种基于双细粒度伪影的深度伪造视频检测方法	深圳大学	2022/6/21	CN115019370A	案件状态：等年登印费
2022106712610	一种基于伪影噪声的深度伪造检测方法	人民网股份有限公司	2022/6/15	CN115100128A	案件状态：等待实审提案
2022105982716	一种基于局部异常的强泛化深度伪造人脸检测方法	南京信息工程大学	2022/5/30	CN114926885A	案件状态：等待实审提案
2022105864442	一种基于时频域视觉伪影特征自适应融合的深度伪造检测方法	大连理工大学	2022/5/27	CN114898438A	专利权有效
2022105862292	一种基于图像频域信息的深度伪造视频溯源方法	国家计算机网络与信息安全管理中心	2022/5/27	CN115188039A	案件状态：等待实审提案
2022105856408	一种基于频域信息与多任务学习的深度伪造视频鉴别方法	国家计算机网络与信息安全管理中心	2022/5/27	CN115187891A	案件状态：等待实审提案
2022105770163	一种基于频率学习的深度伪造人脸检测方法	北京邮电大学	2022/5/25	CN114898437A	案件状态：等待实审提案
2022105489606	一种人脸特点实现深度伪造融合检测的系统、方法、装置、处理装置	公安部第三研究所	2022/5/20	CN114898269A	专利权有效
2022104862469	一种融合噪声感知的深度伪造图像检测方法	山东工商学院	2022/5/6	CN114677372A	案件状态：一通回案实审
2022104318516	深度伪造算法的识别方法、装置、电子设备及存储介质	马上消费金融股份有限公司	2022/4/22	CN114841340A	专利权有效
2022103807661	一种联合多尺度特征深度伪造图像检测方法	中国人民解放军国防科技大学	2022/4/12	CN114724008A	专利权有效
2022103768934	一种支持隐私保护的深度伪造人脸图像检测方法	南京信息工程大学	2022/4/11	CN114758393A	案件状态：等待实审提案
2022102698830	基于三胞胎网络的深度伪造人脸检测方法、系统及设备	武汉大学	2022/3/18	CN114360034A	案件状态：等待合议组成立
2022102295717	一种基于频域滤波残差的深度伪造检测方法	大连理工大学	2022/3/9	CN114757877A	案件状态：等待实审提案
2022102225502	一种无监督深度伪造视频检测处理的方法、装置、处理器及其介质	公安部第三研究所	2022/3/7	CN114627412A	案件状态：等待实审提案
2022101945417	一种利用时空特征深度伪造视频检测方法	北京寒思信安科技股份有限公司	2022/3/1	CN114550268A	案件状态：驳回失效
202210177248X	一种融合深度和宽度学习的深度伪造图像检测方法	河北工业大学	2022/2/25	CN114549481A	案件状态：等待实审提案
2022100585912	一种深度伪造人脸逆向溯源方法及系统	武汉大学	2022/1/19	CN114093013A	专利权有效
2022100401779	一种基于自适应MDS分数的深度伪造视频检测方法	北京邮电大学	2022/1/14	CN114387675A	案件状态：等待实审提案
2021116337797	深度伪造检测方法、装置、存储介质及计算机设备	北京瑞莱智慧科技有限公司	2021/12/28	CN114463805A	专利权有效
202111510599X	联邦协作下的隐私型深度伪造检测方法	大连理工大学	2021/12/10	CN114639174A	案件状态：等年登印费
2021801046496	音频通话的音频深度伪造和非人类说话者的检测	雷德本德有限公司	2021/12/2	CN118369714A	案件状态：等待实审请求
2021114354552	基于插件式部署的深度伪造及检测模型应用系统及方法	合肥高维数据技术有限公司	2021/11/29	CN114201151A	案件状态：一通回案实审
2021113720176	一种基于光流法的深度伪造视频检测方法和系统	厦门市美亚柏科信息股份有限公司	2021/11/18	CN114078119A	案件状态：一通出案待答复
2021112650161	基于视频帧序预测的深度伪造检测方法	杭州中科睿鉴科技有限公司	2021/10/28	CN113989713A	专利权有效
2021111373161	基于空域和频域双特征的深度伪造视频鉴定方法及系统	华南农业大学	2021/9/27	CN113935365A	专利权有效
2021109634943	深度伪造检测模型的评测方法、系统及设备	西安交通大学	2021/8/20	CN113822160A	专利权有效
2021109194727	基于大数据联邦学习架构下的深度伪造溯源系统	电子科技大学	2021/8/11	CN113723220A	专利权有效
2021108745173	基于密集卷积神经网络的深度伪造视频频域检测方法	中国人民解放军战略支援部队信息工程大学	2021/7/30	CN113609952A	专利权有效
2021107768534	基于面部划分的人脸深度伪造检测方法	中国科学院计算技术研究所	2021/7/9	CN113537027A	专利权有效
2021107557533	一种基于人脸关键点的深度伪造人脸视频检测方法、系统及介质	中山大学	2021/7/5	CN113449657A	专利权有效
2021107301226	深度伪造人脸数据鉴别方法	复旦大学	2021/6/29	CN113536990A	案件状态：逾期视撤失效
2021107288698	基于深度伪造视频检测的模型的评价方法及系统	合肥高维数据技术有限公司	2021/6/29	CN113326400A	专利权有效
2021104823012	一种基于表征对比预测学习的深度伪造视频检测方法	湖南大学	2021/4/30	CN113205044A	专利权有效
2021104734324	一种基于多特征融合的深度伪造识别方法和装置	中国科学院信息工程研究所	2021/4/29	CN114067381A	案件状态：一通出案待答复
202110325432X	基于时空融合的深度伪造人脸视频定位方法	西安电子科技大学	2021/3/26	CN113011357A	专利权有效
2021102075750	基于3DMM软生物特征深度伪造视频检测方法	华南理工大学	2021/2/25	CN112818915A	案件状态：驳回失效
2021101100967	一种对深度伪造人脸图像和视频的鉴别方法	四川宏杰麒科技有限公司	2021/1/27	CN112861671A	专利权有效
202110097198X	深度伪造检测方法、装置、存储介质及电子设备	北京有竹居网络技术有限公司	2021/1/25	CN112766189A	专利权有效



扫码开启AI报告互动

零时代·安全牛

网络安全旗舰智库



欢迎关注安全牛