# 金融AIGC音视频反欺诈白皮书

2024.12

# 版权说明



# 编写组

**主编：** 李肇宁

**副主编：** 钱菲、陈树华、田天

**参编人员：**

王光中、赵晗、艾国、高峰、魏恪、 王继科 、 史博、宋文利、李煜明、刘荔园、萧子豪、刘汉鲁、孙空军、杨金威

**参编单位：**

交通银行股份有限公司、北京顶象技术有限公司、北京瑞莱智慧科技有限公司

# 序

早在 2018 年，习近平总书记就强调要未雨绸缪，加强战略研判，确保人工 智能安全、可靠、可控。此后，习近平主席又在多个国际场合倡议“不断提升 人工智能技术的安全性、可靠性、可控性、公平性”“引领全球人工智能健康 发展”[1]。在此背景下，我国陆续出台了一系列法律法规与政策文件，以加强 AI 的安全监管和规范应用。2024 年 7 月，二十届三中全会通过的《中共中央关 于进一步全面深化改革、推进中国式现代化的决定》中，特别强调了“完善生 成式人工智能发展和管理机制。”“加强网络安全体制建设，建立人工智能安 全监管制度。”[2]这是党中央统筹发展与安全，积极应对人工智能安全风险作 出的重要部署。为此，国内发布了包括《国家新一代人工智能标准体系建设指 南》、《生成式人工智能服务管理暂行办法》和《关于依法惩治网络暴力违法 犯罪的指导意见》等多项政策，明确对利用深度合成技术发布违法信息的行为 从重处罚。

在金融领域，基于人工智能的 AIGC 技术的普及带来了显著的创新潜力，但同时也给金融机构的业务安全、客户信任以及系统稳定性构成了新的挑战。特别是音视频领域的 AIGC 欺诈手段，已经成为金融机构必须面对的重要风险之一。这些欺诈行为不仅损害了金融机构的声誉和利益，更对广大客户的财产安全构成了严重威胁。

AI 治理攸关全人类命运， 必须采取切实有效的措施， 贯彻人工智能安全理念，防范 AIGC 欺诈，保障金融业务安全。一方面，要加强技术研发和创新，提升 AIGC 技术的安全性和可控性。通过加强算法研究、优化模型设计、提高数据质量等手段，不断提升 AIGC 技术的准确性和稳定性，减少其被恶意利用的风险。另一方面，要加强监管和治理，建立健全人工智能安全监管制度。通过完善法律法规、加大执法力度、提高监管效能等手段，确保人工智能技术在金融领域的应用符合法律法规要求，保障金融业务的合规性和安全性。

基于此，交通银行、顶象技术、瑞莱智慧联合编写了《金融 AIGC 音视频反欺诈白皮书》，通过详实的数据、典型的案例和前瞻性的技术分析，系统介绍 AIGC 带来的欺诈风险，深入剖析金融机构面临的 AIGC 音视频风险挑战，并提出 AIGC 音视频反欺诈方案、技术实现路径及相关倡议，以期为金融机构提升 AIGC 欺诈识别和防范能力提供有益参考。

相信通过强化合规体系建设， 加强反欺诈技术创新， 构建全链条健康生态， 守正创新携手共进，必将推动人工智能的健康发展，赋能金融高质量发展。

交通银行副行长兼首席信息官：

# 目录

# 第一章 AIGC 带来的音视频欺诈风险

生成式人工智能（AIGC，Artificial Intelligence Generated Content）技术的迅猛发展，推动了内容生成领域的广泛应用，涵盖了文本、图像、音频、视频等多模态内容生成，为娱乐、教育、营销及各行各业的应用带来了前所未有的创新。然而，AIGC 的应用与普及也带来了新的风险挑战，亟需多方监管、加强技术检测与防范措施，确保其在商业应用的安全与透明性，同时加强用户教育以提升风险防范意识。

## 1.1 AIGC 驱动音视频技术创新的同时带来新风险

AIGC 已逐步渗透至各个应用场景中。其背后强大的技术支撑包括图像和视频的生成对抗网络（GAN）、扩散模型（Diffusion Model）、神经辐射场（NeRF）等一系列深度学习技术，以及音频合成中的文本到语音（TTS）和语音转换（VC）等技术。这些技术的进步不仅显著提升了 AIGC 内容的质量和生成效率，也带来了在娱乐、社交、金融等多个行业的广泛应用及新的风险。

### 1.1.1 图像和视频合成技术的发展

**生成对抗网络（GAN）。**生成对抗网络（GAN）是 AIGC 技术的基础之一，它通过生成器和判别器的对抗训练，不断优化生成内容的质量。生成器负责创造出新的图像或视频内容，而判别器则尝试辨别生成内容是否与真实内容相似，从而在不断对抗的过程中提升生成内容的真实性。GAN 技术已经实现了高度逼真的图像和视频生成效果，使得深度伪造成为可能。这一技术的应用场景包括人脸替换、虚拟化身创建等，但同时也为伪造视频的生成提供了可能。

**扩散模型（Diffusion Model）。**随着深度学习算法的进步，扩散模型逐渐成为 AI 视频伪造领域的新兴主流技术路径之一。扩散模型通过在噪声中不断增加与还原信号的过程，能够生成非常逼真的图像和视频序列。扩散模型不仅在生成效果上比 GAN 更为出色，且生成过程更为稳定，其在细节处理、光影效果等方面的表现尤为显著。这使得扩散模型在高保真视频和复杂场景的伪造方面具有巨大的潜力。

**神经辐射场（NeRF）。**神经辐射场（NeRF）技术的出现为 3D 重建与人脸伪造提供了新的方向。NeRF 通过学习光线在 3D 空间中的辐射强度分布，能够实现复杂的 3D 重建和高保真的人脸伪造。这种技术能够将 2D 图像数据重构为 3D 场景，并生成逼真的视觉效果，使得人脸伪造的真实感更高。与 GAN 和扩散模型相比，NeRF 更适用于 3D 场景的模拟与重建，因此其在元宇宙、虚拟现实等领域也具有广阔的应用前景。

当前，以 GAN、Diffusion 和 NeRF 为基础的技术路线在图像和视频伪造领域呈现出三足鼎立的趋势。这三种技术各有优势，分别在 2D 人脸伪造、复杂视频生成、3D 人脸重建等方面各显其长。这些技术的不断演进，使得 AI 视频伪造的质量、速度和逼真度不断提升，带来了更广泛的应用可能性。

### 1.1.2 音频合成技术的发展

**文本到语音（TTS）**。文本到语音（Text-to-Speech，TTS）技术通过将文本输入转化为自然语音，实现了较高质量的语音生成。这一技术的核心在于如何使合成语音听起来自然、流畅，并具有一定的情感表达能力。当前的 TTS 技术可以在短时间内生成高保真的语音，使得虚拟助手、虚拟主播等应用能够轻松模仿真人的语音风格。

**语音转换（VC）**。语音转换（Voice Conversion，VC）技术是另一种关键的音频伪造技术，通过将源语音的特定属性（如音色、语调）转换为目标语音的特征，从而生成与目标人物相似的语音内容。不同于 TTS，VC 技术在保留语音内容的前提下，能够改变语音的特征，使其听起来更接近目标人物。基于深度学习的 VC 技术相比早期的统计建模方法，生成效果显著提升，能够更真实地模拟目标语音风格。

**风格迁移和语音大模型。**在语音伪造领域，风格迁移技术进一步提升了合成语音的自然度和真实性。通过模拟目标语音的说话风格和情绪特征，风格迁移弥补了传统语音合成在情感表现上的不足。同时，语音大模型的出现进一步提高了语音合成的质量和效率。如今，仅需少量的音频样本便可生成高质量的语音合成内容，这使得高精度、低成本的语音伪造成为现实。

## 1.2 AIGC 带来的“换脸”风险

### 1.2.1 AIGC“换脸”的技术原理

AIGC“换脸”技术，是指利用 AIGC 技术，通过对目标视频或图像中的某个人的面部进行替换，将其变为另一个人的面部。此技术依托于深度学习框架，尤其是生成对抗网络（GAN）和大型预训练模型，通过大量人脸数据进行训练，以生成高度逼真的“换脸”效果。GAN 由生成器（Generator）和判别器（Discriminator）构成，生成器负责生成与真实数据难以区分的“假数据”，而判别器则负责判断生成的图像真假，二者不断对抗，优化生成效果，最终生成逼真的人脸替换效果。

通过 GAN 和其他模型的协同，AIGC“换脸”技术能够学习到人脸的细微表情、光线反射、纹理细节等因素，在面部表情变化、嘴唇与声音同步、光影调整等方面取得了极高的真实度。此外，AIGC“换脸”技术的生成过程也因其高度自动化而具备较强的泛化能力，无需过多人工干预便可以实现逼真且多样化的面部替换效果。

### 1.2.2 AIGC“换脸”的主要应用场景

**影视与娱乐行业。**AIGC“换脸”技术在影视制作中的应用广泛。例如，可以用明星的面孔替换替身演员的面孔，使表演更加真实且减少重复拍摄需求。

此类技术也被应用于影片复原或重拍，将已故演员的形象复现到影片中。此外，AIGC“换脸”技术在虚拟主播、数字偶像等新兴娱乐领域中同样受到关注。

**社交媒体和创作。** 在社交媒体上，“换脸”特效让用户能够体验角色扮演 的乐趣，迅速成为热门潮流。通过 AIGC“换脸”技术，用户可以在短时间内生 成内容，便捷地分享具有高度真实感的“换脸”视频。此技术不仅为用户提供 了极大的创作空间，也为个性化内容的生成和传播提供了可能性。

**虚拟现实与增强现实应用。** AIGC“换脸”技术在虚拟现实（VR）和增强现实（AR）领域也具有重要作用。例如，利用“换脸”技术可以使用户在虚拟场景中拥有不同的面孔，从而进一步提升沉浸感。无论是游戏角色的面貌定制，还是 AR 社交平台上的角色扮演，这类应用都因 AIGC“换脸”技术而变得更具吸引力和互动性。

### 1.2.3 AIGC“换脸”带来的安全挑战

尽管 AIGC“换脸”技术在多个领域展现出潜力，但其也带来了安全和道德上的挑战。黑灰产可能利用此类技术在未经授权的情况下非法使用他人肖像，甚至对当事人形象进行恶意篡改或丑化，存在侵犯肖像权、名誉权及隐私权的风险。

2024 年 1 月，美国知名歌手泰勒 ·斯威夫特的伪造“不雅照片”在 Facebook 等社交平台广泛传播，累计浏览量超过千万。尽管最初传播该照片的账号已被封禁，但照片的扩散仍未彻底遏制，严重侵犯了泰勒 ·斯威夫特的个 人权益。

2023 年 5 月 23 日，包头警方公布了一起利用 AI 实施电信诈骗的典型案例， 福州市某科技公司法人代表郭先生在短短 10 分钟内被骗走 430 万元人民币。

2024 年 2 月，一黑灰产通过“换脸”技术伪造跨国公司高层身份，参与视 频会议指挥分公司向指定账户汇款，成功骗取 2 亿港元。

利用 AIGC“换脸”技术带来的风险主要在身份伪造与深度伪造、隐私泄露和滥用、社会信任的破坏等三个方面。随着实施此类犯罪的技术门槛逐步降低，并预计将持续上升。

**身份伪造与深度伪造。** AIGC“换脸”技术的真实性使其成为一种极具隐患的身份伪造手段。

**隐私泄露和滥用。** 利用 AIGC“换脸”技术生成的假视频、假照片很可能在未经授权的情况下泄露他人隐私，甚至被用于恶意传播不实信息。这不仅侵犯了隐私权，还可能对受害人造成名誉损害。

**社会信任的破坏。**AIGC“换脸”技术的大范围应用可能削弱公众对视频和图像真实性的信任。例如，普通用户难以区分真假视频，进而对信息的真实性产生怀疑，甚至影响到司法调查、新闻报道等领域的公信力。

## 1.3 AIGC 带来的“拟声”风险

### 1.3.1 AIGC“拟声”的技术原理

AIGC“拟声”技术利用人工智能深度学习模型，通过对大量音频数据的学习和训练，生成高度逼真的合成语音，使其几乎与目标人物的声音完全相同。这一技术如今已经广泛渗透到智能语音助手、虚拟主播、金融身份认证等多个领域，显著提升了音频生成的质量与效率。尤其是小样本语音合成技术的突破，使得仅凭短短几秒或一分钟的音频样本，即可生成长时间、高质量的合成音频。这不仅提高了语音合成的便捷性，还增强了其多样化应用的可行性。

AIGC“拟声”技术主要通过深度学习模型实现，具体而言，模型首先需要大量目标声音的数据进行学习，以分析并捕捉声音特征。常见的方法包括生成对抗网络（GAN）和序列到序列模型（Seq2Seq）。生成对抗网络模型由两个部 分组成：生成器和判别器，生成器试图生成逼真的目标声音，而判别器则负责判断生成的音频是否与原声匹配。二者在对抗训练中不断优化，最终生成能够高度仿真的音频内容。

此外，小样本语音合成技术通过少量的目标音频样本，使用迁移学习或适应性建模，使 AI 在只获取少量数据的情况下便能够学习和模仿目标声音的个性 化特征。这意味着 AIGC 可以快速生成相似度极高的语音，并缩短了大量样本收 集的时间成本。

**小样本语音合成。**过去，音频合成往往需要数小时的目标音频数据，而如今小样本语音合成技术的出现，使得 AIGC“拟声”技术在获取少量音频样本后，即可生成高质量的语音。这一技术的关键在于迁移学习和适应性训练模型。通过仅一分钟左右的目标音频数据，AI 可以快速适应并生成具有高相似度的音频内容，大大降低了语音合成的时间和资源成本。

**情感与音调调节。**通过情感控制模型，AIGC“拟声”技术可以根据需求调节语音的情感色彩，生成更加自然的情绪表达效果。例如，在客服系统中，可以为合成语音添加温和、愉悦或严肃的情感，以改善用户体验。而在影视配音中，AIGC 技术可以根据剧中人物的情绪调节语音的抑扬顿挫，使语音更加贴合剧情需求。

**跨语言语音生成。**AIGC“拟声”技术还实现了跨语言语音生成，能够在不同语言间自动生成目标音色。这对于多语种语音助手、国际广告配音等应用场景具有重要意义。跨语言语音生成借助序列到序列模型的迁移学习功能，使得 AI 在不同语言间保持同一人的音色特征，从而实现自然语言间的声音转化。

### 1.3.2 AIGC“拟声”的主要应用场景

**智能语音助手与客服系统。**许多智能语音助手和客服系统都已应用了 AIGC“拟声”技术，使得用户能够与拥有自然人声的虚拟助手进行沟通。AI 语音助手不仅能够模仿不同年龄、性别的声音，还可以根据情境和用户需求自动调节音调、情感，从而实现个性化互动体验。

**虚拟主播和娱乐内容生成。**AIGC“拟声”技术在虚拟主播、短视频平台等娱乐领域得到广泛应用。借助该技术，虚拟主播可以拥有富有感染力的声音，而不需要真人配音。此外，视频制作人能够快速生成与内容相匹配的配音，不仅提升了内容生产的效率，也拓宽了音频内容创作的想象空间。

**金融身份验证与安全识别。**在金融行业，AIGC“拟声”技术逐渐成为身份认证的有效工具。例如，基于声纹识别的身份验证系统可以在不依赖于烦琐的

密码和指纹的情况下，通过声音对客户进行身份确认。然而，AIGC“拟声”技术的进步也对金融身份验证带来了新的挑战，需要进一步提高声纹识别系统的鲁棒性，以防止伪造音频的安全风险。

**影视、配音与广告制作。** AIGC“拟声”技术已广泛应用于影视、广告配音中，通过生成特定音色的语音，电影制作人可以将已故演员的声音“复活”，或者为国际观众定制化配音。同时，在广告中使用 AIGC“拟声”可以根据不同地域、年龄、性别的受众生成个性化声音，从而增强传播效果。

### 1.3.3 AIGC“拟声”带来的安全挑战

AIGC“拟声”技术使得伪造音频的成本降低，黑灰产可以模仿他人声音实施电信诈骗，甚至在虚假视频中伪造他人声音。这不仅可能损害个人声誉，威胁金融交易乃至人身的安全。此外，AIGC“拟声”技术生成的声音可用于不实信息的传播，如虚假新闻、政治操控等。

2023 年 12 月，一名留学生在境外被“绑架”，父母遭“绑匪”索要 500 万元赎金，还收到了“肉票”被控制、伤害的视频。通过现场调查及国际警务合作，5 个小时后发现，这是诈骗分子精心布设的一起骗局。

2024 年 2 月，刘德华公司发表声明，称有人利用 AGIC 技术合成刘德华的声音以吸引流量并销售商品获利。

2024 年 4 月，密码管理工具公司 LastPass 披露一起 AIGC 伪造声音的诈骗。骗子冒充 LastPass 首席执行官卡里姆 · 图巴创建了一个 WhatsApp 账户，然后向码管理工具公司 LastPass 员工接员发送一系列消息，包括使用 AI 的伪造的卡里姆 · 图巴声音的语音消息，多条未接的音频通话等。不过，该员工很快就 识别出这是骗子伪造的电话。

2024 年 7 月，公安部公布十大高发电信网络诈骗类型，其中有一个“冒充领导、冒充熟人”诈骗格外引人注目。诈骗分子利用受害人领导、熟人的照片、姓名包装社交账号，通过添加受害人为好友或将其拉入微信聊天群等方式，冒用领导、熟人身份对其嘘寒问暖表示关心，或模仿领导、老师等语气发出指令，从而骗取受害人信任。

AIGC“拟声”技术让电信网络诈骗更加复杂，黑灰产利用社交媒体、社交工具、电话、远程会议等发动各类攻击。

**社交媒体诈骗。** 黑灰产利用克隆声音在社交媒体上冒充公众人物或受害者 亲友，进行诈骗活动。

**社交工具或电话诈骗。** 通过社交工具发送伪造的语音，或直接拨打电话，黑灰产诱使受害者透露敏感信息或直接转账。

**远程会议攻击。**在远程会议中，黑灰产通过克隆参会者的声音进行干扰或误导，窃取商业机密或个人数据。

**伪造新闻资讯。** 黑灰产在伪造新闻资讯中使用克隆声音，以增加信息的可 信度，诱导受害者上当。

## 1.4 AIGC“换脸”“拟声”风险的特征

### 1.4.1 生成内容的高仿真性

随着 AIGC 生成技术的快速发展，尤其是深度学习算法中的生成对抗网络，已经能够创造出极其逼真的图像和视频，它们通过学习大量数据样本，生成与真实样本难以区分的视觉内容。在音频领域，这项技术能够模拟和复制特定人的声音特征，包括语调、节奏和情感色彩，使得合成音频与真实录音难以区分。AIGC 生成内容的高仿真性是此类风险的典型特征之一。

### 1.4.2 内容生成的低成本和高效率

随着 AIGC 技术的普及，内容生成的成本显著下降，易用性大幅提升，这使得普通用户无需专业知识即可生成高质量内容。AIGC 工具的易用性使得生成逼真的虚假图像、音频和视频变得更加简单，成为 AIGC 风险增加的重要因素。此外，AIGC 技术的自动化生成能力也意味着虚假内容的生产门槛降低，可以大规模自动化地生成内容，为黑灰产开展规模化攻击提供了技术基础，降低了黑灰产攻击的难度。

### 1.4.3 难以溯源的隐匿性

随着 AIGC 技术的快速发展，内容生成的随机性和复杂性显著增加，这不仅使得追踪虚假内容的来源变得异常困难，也对虚假内容的识别和溯源提出了新的挑战。利用生成对抗网络（GANs）和变分自编码器（VAEs）等先进模型，可以根据文本描述或随机种子生成高度逼真的图像和视频内容，这种随机性的特性让虚假内容的来源难以捉摸。同时， 匿名发布的特性和匿名化处理技术，如 实时匿名化处理，允许用户在不暴露身份的情况下发布内容，这虽然在一定程 度上保护了用户的隐私，但也为虚假信息的传播提供了可乘之机。因此，面对 AIGC 技术不断进步带来的挑战，我们需要开发新的检测工具和技术，以保护信 息的真实性和网络环境的安全。

### 1.4.4 跨模态内容生成与融合

AIGC 技术通过学习大量训练数据中的模式， 能够自主创建包括文本、图像或音乐在内的各种原创内容，其应用范围广泛，从文本生成、图像生成到音频生成和视频生成都能覆盖。跨模态融合内容，例如带语音的视频或图文并茂的虚假报道，由于结合了多种感官信息， 更具欺骗性，增加了识别难度。这种跨 模态的生成能力不仅拓宽了艺术创作的表现形式，但也为 AIGC 内容检测带来了 新的挑战，尤其是在虚假信息的识别和防范方面，如通过输入文本描述生成视 觉内容，或者将文章自动转换成视频， 使得虚假内容更难被识破。跨模态内容 生成技术的出现和普及为黑灰产攻击提供了更加具有欺骗性的技术攻击，进一 步加大了防御的难度。

# 第二章 AIGC 音视频欺诈典型攻击方法

AIGC“换脸”攻击利用生成对抗网络（GAN）或扩散模型等技术，将攻击目 标的面部与他人面部替换，生成极具欺骗性的伪造图像或视频。此类攻击已被广泛用于身份冒充、金融诈骗等场景， 增加黑灰产获取非法收益的手段的复杂 性和隐蔽性。 AIGC “拟声”攻击通过生成和模仿目标高度相似的语音内容进行 欺诈。在电话诈骗、语音识别系统攻击中常被使用，通过假冒声音进行身份验 证绕过和财产欺诈，为金融和企业安全带来严峻挑战。

面对 AIGC 带来的新型攻击，现有防护手段难以有效检测伪造的内容，亟需加强技术研发、制定应对策略，并提升公众对 AIGC 攻击手段的风险防范意识。

## 2.1 AIGC“换脸”攻击分析

### 2.1.1 AIGC“换脸”攻击目标

当前人脸识别系统作为一种重要的身份核验方式被广泛应用于金融行业，金融账户远程开户、账户解锁、消费金融申请、信用卡申领、业务签约、银行卡业务、核保理赔等金融业务，均可利用远程人脸识别进行身份的核验。人脸识别系统作为客户身份认证的关键环节，也成为黑灰产的主要攻击目标。

远程人脸识别系统由客户端、服务器端、安全传输通道组成。系统由客户端实现人脸的采集，经安全传输通道传输，在服务器端远程进行比对。


图 2-1 远程人脸识别系统参考模型[3]

基于远程人脸识别的互联网账户刷脸流程如图所示，具体流程如下：

（1）在移动应用中采集用户人脸数据，并存储在服务端，作为人脸的比对源。

（2）当一个业务场景需要核实用户的身份， 则需要再次采集用户人脸数据。

（3）系统将获取到的人脸数据，通过活体检测和质量检测后，与用户原有的比对源数据进行人脸比对，如比对通过则用户可以进行后续操作。如果未通过活体检测、质量检测或者人脸比对失败，用户需要进行重试。

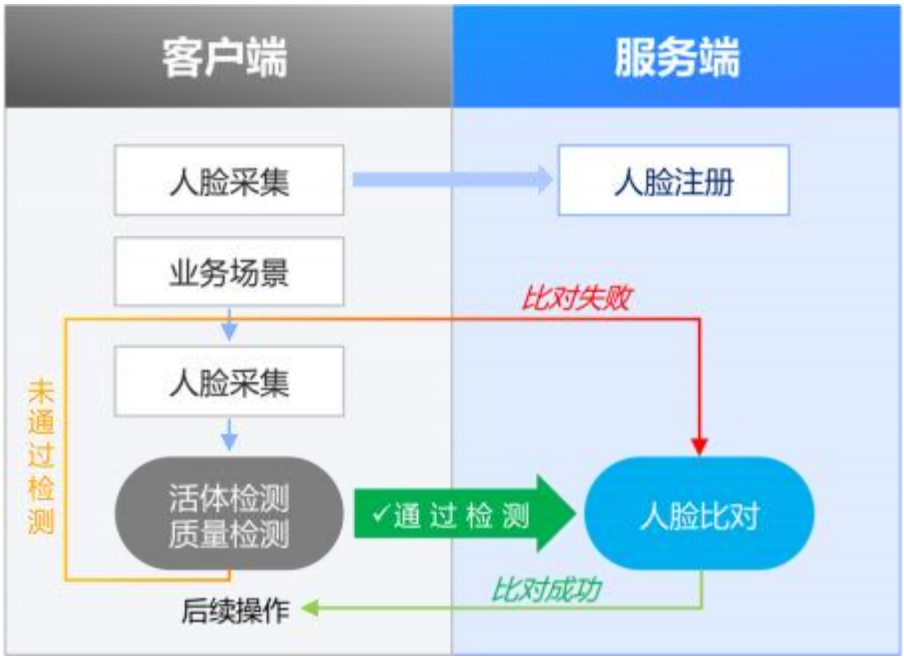


图 2-2 基于远程人脸识别的互联网账户刷脸流程[4]

### 2.1.2 AIGC“换脸”攻击过程

利用远程人脸识别系统进行身份认证需要经过人脸采集、活体检测、人脸比对等多个环节，黑灰产攻破其中任意一个环节都有可能攻破人脸识别系统。随着 AIGC 工具的普及，利用 AIGC 工具可以实现人脸替换和表情操纵。黑灰产可通过定制客户端 ROM 或者劫持客户端摄像头的方式，在人脸采集环节将受害者的伪造的受害者视频注入客户端，非法通过活体检测和人脸比对环节，成功实现攻击。

黑灰产对人脸识别系统实施攻击的流程如下图所示。首先，黑灰产可通过购买身份信息等方式获取受害人的高清身份证人脸图像；其次，利用 PS 工具将 人脸图片设置为带背景的图片；然后，利用 AIGC 工具对受害者的照片进行活化 处理，生成“眨眼”“摇头”的伪造视频；最后，将伪造视频注入 APP 中对身 份认证系统实施攻击。


图 2-3 AIGC“换脸”攻击过程

### 2.1.3 AIGC“换脸”攻击技术

黑灰产在实施 AIGC“换脸”攻击时的核心目标是将攻击目标的人脸“活化”，驱动攻击目标实现“点头”、“摇头”、“张嘴”等动作，来通过人脸识别系统的活体检测验证。在这个过程中，利用的核心技术是人脸表情驱动技术，随着实时人脸替换技术的出现，实时人脸识别技术也可以作为 AIGC “换脸”攻击的重要技术手段。

人脸表情驱动技术是指利用深度合成技术实现对图像或视频中的人脸表情进行分析、编辑和修改的技术。这种技术能够操纵原始图像、视频中的人脸，使其做出指定的表情和口型，合成指定的讲话音视频。该技术一般通过提取人脸图像中与表情相关的特征，如眼睛、嘴巴、眉毛等部位的形状、位置及运动信息，将提取的特征与预先训练的表情模型进行匹配，以确定人脸所表达的情感状态。基于分析结果，对目标表情进行编辑或操纵，如改变表情类型、强度或合成全新的表情。

人脸替换技术是指利用深度合成技术将原始图像中的人物的面部，替换成其他人物的面部，完成人脸的“裁剪”和“嫁接”。该技术首先提取人脸的关键特征，如眼睛、鼻子、嘴巴等部位的位置；其次将源人脸与目标人脸进行对齐，确保替换后的面部特征与原视频中的人物动作保持一致；最后将提取的源人脸特征合成到目标人脸的位置上，实现自然过渡和逼真效果。在此基础上，实时人脸替换技术可以实现人脸表情的实时变换，实现 AIGC 人脸攻击。

目前已经出现 Muglife、CrazyTalk 等伪造 APP 实现眨眼、点头、摇头等人 脸表情的操纵，降低了 AIGC 人脸攻击的难度。Muglife、CrazyTalk 等伪造 APP 可以实现眨眼、点头、摇头等人脸表情的操纵。Roop、DeepFaceLive 等开源应 用实时的人脸替换。这些 APP 的出现降低了 AIGC 人脸攻击的难度。

## 2.2 AIGC“拟声”攻击分析

### 2.2.1 AIGC“拟声”攻击目标

除了人脸识别系统外，声纹识别系统也被作为一种常用的身份认证方式被应用于金融服务中，也是黑灰产开展 AIGC“拟声”攻击的目标。声纹识别是根据待识别语音的声纹特征鉴别该段语音所对应的说话人的身份过程，在移动金融服务中基于声纹识别的应用流程如图 2-4 所示，用户通过拾音设备进行语音采集，经移动金融客户端加密传输至服务器端。客户端前置服务器进行必要的业务处理后将语音信息传输至声纹服务器。声纹服务器完成声纹的注册、验证、变更或注销，并将相应的结果（接受或拒绝）经客户端前置服务器反馈至移动金融客户端。


图 2-4 声纹识别应用流程示意图[5]

### 2.2.2 AIGC“拟声”攻击过程

利用远程声纹识别系统进行身份认证需要经过语音采集、声像攻击检测、声纹比对等多个环节，黑灰产攻破任意一个环节都有可能攻破声纹识别系统。随着 AIGC 工具的普及，利用 AIGC 工具可以实现语音的合成及转换。黑灰产可以在语音采集环节，通过播放受害者的合成音频对声纹识别系统实施攻击，通过身份验证。

黑灰产对声纹识别系统实施攻击的流程如下图所示。首先，黑灰产可通过电话诈骗录音等方式获取受害人的语音素材；其次，利用 AIGC 工具生成攻击目标的伪造音频；最后，利用呈现攻击或者注入式攻击方式对身份认证系统实施攻击。


图 2-5 AIGC“拟声”攻击过程

### 2.2.3 AIGC“拟声”攻击技术

黑灰产在实施 AIGC“拟声”攻击时的核心目标是利用 AIGC“拟声”技术合成目标任务的伪造音频，通过播放伪造音频来通过声纹识别系统的检测验证。在这个过程中，利用的核心技术是文本转语音的合成技术。

自动语音合成（TTS）技术可根据指定的语言文本生成目标说话人声音，实现文本到语音的映射。典型的语音合成系统包括前端文本分析和后端语音波形生成两部分。文本分析将输入文本通过规范化、分词、词性标注等步骤生成对应的因素序列、时长预测等信息；语音波形生成根据文本分析生成的语言规范合成目标说话人的语音波形。通过对合成声音进行微调，可以让合成语音听起来更自然，更容易理解。随着技术的发展，目前已经出现 ChatTTS、Seed—TTS 等开源语音合成应用服务，可以快速合成高质量的伪造语音，降低了 AIGC“拟声”攻击的难度。

# 第三章 AIGC 音视频欺诈对金融业务的影响

## 3.1 增加金融业务风险

随着数字化进程的加速，金融业务全面线上化已成为行业趋势，用户的生物特征如人脸与声音，成为身份认证和交易的重要凭证。金融业务线上化提升了金融服务的便利性，但也让金融机构和消费者面临全新的安全挑战。AIGC 技术的快速发展，使得金融业务的风险面进一步扩大，尤其是在换脸技术和拟声技术被恶意利用时，金融业务面临前所未有的威胁。

随着 AIGC 技术的快速发展，金融机构的线上业务风险结构正在经历显著的变化。传统的金融身份验证依赖于生物特征，如人脸识别和语音认证，因其基于个人的独特特征，这些方法曾被认为高度安全。然而，黑灰产利用 AIGC 技术轻松复制个人特征，生成高精度的深度伪造内容。例如，换脸技术能根据获取到的受害者照片生成难以鉴别的图像或视频，拟声技术则能根据受害者少量声音片段生成高质量伪造音频，这使得攻击者能够绕过生物认证系统，冒充用户进行高风险操作，如资金转账、贷款申请等。相较于传统手段，AIGC 技术的应用提高了虚假内容的生成质量，使得欺诈行为更加隐蔽和难以分辨，依靠 AIGC 技术生成的虚假音频、视频可达到以假乱真的程度，使金融机构难以通过常规手段识别，直接影响线上业务的安全性。

相较传统方法，AIGC 工具大幅提高了攻击效率和影响范围。借助开源或商用的 AIGC 工具和自动化脚本，黑灰产可以快速实现对金融系统的批量攻击，造成重大经济损失。例如，黑灰产可以利用开源工具生成大量伪造身份信息，在金融系统中实施“批量开户”或“多头借贷”，扰乱正常的业务运转。这类攻击成本更低且速度更快，能够在短时间内造成巨大的经济损失，对金融安全构成严重威胁。

## 3.2 给黑灰产攻击提供新手段

AIGC 的应用让欺诈变得愈加隐蔽和高效。这些技术不仅提高了欺诈内容的真实感，还极大地降低了欺诈操作的门槛，使得欺诈行为更加复杂多样。

AIGC 生成的音频、视频和图像内容具有高度真实的深度伪造能力，能够轻松欺骗普通用户甚至部分现有的验证系统。例如，语音模拟技术仅需少量语音片段即可生成高度逼真的声音，精确模仿语调、语速和情感特征，用于实施电话诈骗或身份冒充。此外，深度伪造技术可生成精准的换脸视频，真实还原受害者的面部特征与表情，甚至在通话中欺骗验证。AIGC 还可制作高分辨率的伪造身份文件，如身份证、合同等，使欺诈手段更加难以察觉。这种技术上的突破不仅增加了伪装的真实性，也使得传统防护手段面对新型欺诈时难以奏效。

此外，相较传统方法，AIGC 生成内容的隐蔽性和多样性显著提升，进一步加剧了金融欺诈的复杂性，传统检测系统难以快速识别伪造性质。例如，伪造视频在光影一致性、面部纹理和表情细节上高度逼真，使得识别伪造更加困难。

与此同时，AIGC 赋予了欺诈行为多样化的能力，从语音冒充到视频伪造，再到文本生成，每种手段均可独立实施，或联合形成多维度攻击的“组合拳”。在对真实性要求高的场景中，如身份验证或在线交易，AIGC 生成的语音模拟和深度伪造视频足以假冒客户身份，通过银行验证流程并非法窃取资金。高度拟真的内容让受害者更难辨别真伪，直接威胁金融交易的安全性。

同时，AIGC 技术的普及大幅降低了欺诈操作的技术门槛，传统金融欺诈通常需要技术团队支持和复杂设备，成本较高。而 AIGC 工具以其易用性和普及性，使得没有专业技术背景的攻击者也能生成高度逼真的音频、视频或文本内容，轻松实施复杂的欺诈行为。例如，攻击者可以利用成熟的开源技术伪造身份文件突破支付 APP 身份认证环境实施诈骗。这种技术的普及使黑灰产组织能够以更低成本实现更高效率的欺诈操作。这样的大规模欺诈行为突破了人工操作的局限，为黑灰产的扩张提供了便利。

## 3.3 为防御带来新挑战

AIGC 技术赋予黑灰产以更高的效率和更大规模生成复杂攻击样本的能力，极大地增加了欺诈手段的多样性和隐蔽性，带来巨大信息威胁，使得金融业务的防御体系面临严峻考验。

**欺诈内容的真实性与难辨性。**AIGC 技术的核心优势在于其能够生成极具真实性的伪造内容，这些内容往往在视觉和听觉上都能以假乱真，令受害者难以辨别真伪。在金融业务场景中，尤其是身份验证和在线交易等对真实性要求极高的环节，AIGC 生成的伪造内容能轻松冒充客户身份进行欺诈。黑灰产能够利用伪造的声音模拟技术，模仿客户的语音指令进行转账操作；或通过伪造客户的面部图像、视频进行远程身份验证， 从而获取非法资金。这种高度拟真的内 容直接威胁到金融交易的安全性，给金融机构带来了巨大的安全隐患。

**攻击样本的多样化与复杂化。** 传统的欺诈手段往往是针对某一特定漏洞或 手段的攻击，而基于 AIGC 的欺诈行为却能够突破原有防御体系，采用多重伪装 技术进行攻击。AIGC 的迅速普及意味着黑灰产能够以极低的成本和快速的迭代 速度，不断创新攻击方式。这对防御系统的多层次性和应变能力提出了更高要 求。现有的防御系统往往难以应对这种快速变化的复杂威胁，特别是在面对智 能化攻击时，传统的基于规则的防御方式显得捉襟见肘。金融机构必须提升对 AIGC 带来新型风险的敏感度，及时识别复杂的欺诈模式，并做好相应的风险应 对措施。

**快速迭代的攻击与防御的滞后性。** 与传统攻击方式相比，AIGC 工具的易用性和高效性为攻击者提供了更多的时间优势。黑灰产不再依赖传统的复杂编程或长时间的攻击准备，而是可以通过现成的 AIGC 工具快速生成攻击样本并开展攻击。这使得攻击策略的更新速度远远快于金融机构现有防御系统的响应速度，防御系统的应对能力往往滞后于攻击者的创新。在这种背景下，金融机构的防御系统必须具备快速迭代的能力，借助机器学习和人工智能技术，建立具有自适应能力的动态防御系统，通过利用大数据分析、行为识别技术以及实时监控系统，防御系统可以自动识别异常行为并及时响应。

**动态学习与预测潜在威胁。** 随着攻击手段的不断演化， 金融机构必须确保 防御体系能够及时学习和适应新的威胁模式， 不仅能够实时识别当前的攻击，

还能基于历史数据和攻击模式进行预测，并预防未来可能的攻击。例如，通过训练深度学习模型，系统可以从历史欺诈数据中学习不同的攻击路径，并预测潜在的风险区域，从而提前做出防范。

## 3.4 对金融反欺诈提出新要求

AIGC 对金融机构的风险管理体系提出了更高的要求。传统的基于规则的风控系统，通常依赖固定的模型和人工设置的规则，已无法有效应对 AIGC 驱动的新型欺诈方式。因此，金融机构亟须升级其风险管理体系，采用更加智能化、动态化的防控手段，以提升对 AIGC 欺诈行为的识别、应对和防范能力。

**使用 AI 技术作为反欺诈重要工具。**AI 技术本身可以成为金融机构防控欺诈的有力工具。通过分析客户的行为数据，AI 可以帮助识别潜在的欺诈行为。例如，通过分析客户的使用习惯数据，系统能够检测到任何异常的操作模式。这种基于行为分析的风控机制，可以帮助识别伪造身份、非法登录等欺诈行为。

同时，AI 可以帮助金融机构开发更高效的伪造检测工具。传统的验证方法，如密码、验证码等，已经无法完全防范由 AI 生成的伪造视频、语音等内容。金融机构应加强深度伪造检测技术的应用，实时识别和验证换脸视频、语音模拟等伪造内容。AIGC 的实时分析能力可让风控系统在检测到异常行为时，迅速响应。例如，发现大额转账或可疑登录行为时， 系统可以自动冻结账户或要求额外身份验证，从而及时制止欺诈行为，减少损失。

**建立动态化的智能风控体系。** 金融机构需要建立更加动态和自适应的风控 体系。随着 AIGC 欺诈手段的不断演化，传统的静态风控系统已无法满足需求。 金融机构应通过机器学习等技术，实时学习和适应新的欺诈模式，动态调整防 御策略。基于人工智能的风控系统可以快速识别新型欺诈行为，并及时更新防 御规则。对于交易数据的存证也是提升风控系统可信度的重要手段。通过采用 区块链等技术，金融机构可以确保交易数据的不可篡改性，增强数据的可靠性 和透明度。这有助于在发生欺诈行为时，能够追溯到原始数据并提供可靠的证 据。

**加强技术创新和行业协作。** 应对 AIGC 带来的欺诈风险，不仅需要技术创新，还需加强行业间的协作。金融机构应与技术公司、监管部门紧密合作，共享威胁情报和技术创新成果，共同开发应对 AIGC 欺诈的新技术和方法。此外，金融机构还需积极参与全球或区域性的 AIGC 技术和安全规范的制定，为反欺诈技术的标准化提供支持。

**提升公众教育与安全意识。** 公众教育是反欺诈工作中不可忽视的一环。金 融机构应通过科普活动向消费者普及 AIGC 欺诈的风险，提升公众对深度伪造内 容的辨识能力。提供权威的伪造检测工具，帮助用户验证可疑内容的真实性， 增强其自我保护意识。这不仅能够提高用户对金融机构的信任度，还能有效减 少社会层面上 AIGC 欺诈的传播。

面对 AIGC 技术引发的安全挑战，金融机构必须不断创新和完善其风险管理体系，以适应这一新兴威胁。

# 第四章 AIGC 音视频反欺诈方案

为了有效应对 AIGC 带来的威胁，金融机构必须建立一个覆盖全周期、全场景、全链条的防御体系。技术层面，金融机构应引入多模态 AIGC 音视频欺诈检测与鉴定技术，通过结合图像、声音、行为等多维数据，精准识别伪造内容。同时，借助 AIGC 特征的欺诈团伙识别技术和融合智能决策引擎，实时监测并应对欺诈行为。从业者方面，金融从业人员应不断提升对 AIGC 深度伪造欺诈的识别与应对能力，帮助机构更好地管理风险，减少攻击影响。管理体系方面，通过引入前沿技术、加强数据整合与共享、提升员工培训，并优化法律与合规管理，提升金融机构风险应对能力。此外，持续的演练和模拟也有助于强化防御，确保金融业务的安全和稳定。

## 4.1 构建全面防御体系

金融机构面临的 AIGC 欺诈风险日益加剧，必须建立覆盖全周期、全场景、全链条的防御体系，以应对 AIGC 带来的复杂威胁。


图 4—1 AIGC 反欺诈技术体系

**构建全周期防御。** 在全周期防御方面，事前预防是关键。金融机构需通过 严格的数据访问控制和加密存储技术保护敏感信息，并利用多因子验证手段增 强身份认证的安全性。此外，基于机器学习和行为分析的动态风控模型可帮助 机构提前预测潜在欺诈行为。在事中监测阶段，AIGC 技术为实时监控交易和操 作行为提供了技术支持，能够快速识别异常操作并采取响应措施。同时，深度 伪造检测技术可验证音视频内容的真实性，有效遏制换脸视频和模拟语音欺诈 行为。此外，通过 TLS 协议和端到端加密技术，可进一步防范数据传输过程中 的信息篡改与泄露。事后响应则聚焦于日志追溯与快速冻结，利用区块链技术 确保数据不可篡改性，为欺诈事件的后续追踪提供可靠依据，并及时冻结涉事

账户以控制损失。

**全场景防护是应对多样化 AIGC 威胁的重要手段。**在用户身份验证场景中，活体检测技术（如眼球追踪、光影变化捕捉）能够有效防范深度伪造攻击。同时，AI 辅助审核技术可识别身份文件的细微伪造特征，提升审核精度。在交易场景下，动态监测用户行为特征（如交易频率、地理位置）并建立异常警报机制，可以及时发现高风险操作；对于高额或敏感交易，还可通过分级验证策略强化安全保障。在数据流转环节，依托零信任架构限制敏感数据的传播范围，确保数据仅在授权环境下可见，从而有效防范内部数据泄露。

**全链条防控的构建需要内外协作共同发力。**内部防控强调最小权限原则和员工安全教育，减少未经授权访问和内部人员成为攻击突破点的风险。外部防控则依赖于行业协作，与其他金融机构、监管部门和技术企业建立威胁情报共享机制，快速获取新型欺诈手段的信息并调整防御策略。

**技术和管理的协同发展是构建全面防御体系的关键。**一方面，通过开发深度伪造检测和区块链存证技术，可从技术层面遏制欺诈行为；另一方面，完善 AIGC 监管规则与问责机制，并加强公众教育，能提升全行业和用户的安全意识。只有从全周期、全场景、全链条的维度构建这一综合防御体系，金融机构才能在 AIGC 欺诈日益复杂的环境中确保安全与稳定。

## 4.2 技术解决思路

AIGC 音视频欺诈的防御过程包括异常检测和鉴定欺诈两个关键环节，通过系统化流程实现对欺诈行为的快速识别与应对。

在异常检测环节，系统通过监测用户行为特征，如登录设备的变化、操作频率的异常、跨区域访问等情况，识别可能的风险信号。同时，对接收的音视频数据进行结构性分析，标记潜在的异常内容，例如音频中不自然的语调或视频中的细微拼接痕迹。并利用机器学习模型不断优化检测能力，识别新型伪造手段，提高异常检测的精准性。

在鉴定欺诈环节，针对确认的异常内容，系统进一步验证其与真实数据的匹配程度。例如，通过比对用户的历史行为模式、生物特征以及系统存储的原始数据，判定音视频是否为伪造。一旦确认欺诈行为，系统将自动触发警报，并采取相应措施，如冻结账户、阻止交易或要求额外验证，确保风险能够得到及时控制。

### 4.2.1 多模态 AIGC 音视频欺诈的检测技术

在音频伪造中，黑灰产可以利用 AIGC 生成几乎完美模仿目标对象声音的伪造语音，并拼接不同音频段以制造欺骗效果。图像伪造则可能涉及对人脸、物体或场景的精细修改，例如细微的纹理变化、边缘处理等，极大增加了检测难度。视频伪造则更加复杂，伪造者可以篡改特定片段中的面部表情或肢体动作，从而影响整体语义一致性。多模态 AIGC 伪造内容，结合了音频、图像和视频等领域，通过生成对抗网络（GAN）等深度学习技术生成。这些伪造内容看似真实，但实际上包含许多细微的伪造痕迹。针对多模态 AIGC 伪造内容的检测技术，通

过结合音频、图像和视频等不同模态的信息，利用深度学习和专家知识，可以构建更为精准的伪造内容检测体系。

#### 4.2.1.1 音频伪造检测技术

针对音频伪造，构建基于声纹拼接痕迹的检测体系是目前的关键技术路径之一。通过深入分析音频的波形、频谱图极其微小的拼接痕迹，结合深度学习技术与专家知识，可以检测出伪造痕迹。同时，结合不确定性估计技术，开发出更具鲁棒性的伪造语音鉴别方法。例如，基于声纹细节的拼接点异常特征，可以识别出人工合成或拼接的语音内容。这种方法不仅可以用于检测合成语音，还可以识别出经过编辑的录音，确保其真实性。

此外，声音的频率成分和语音波形中的异常变化也是关键的伪造线索。通过机器学习算法自动提取这些细微的差异，能够有效区分伪造语音和真实语音。这种技术还可以与基于人类声音特点的分析方法相结合，提升对语音伪造的识别精度。

#### 4.2.1.2 图像伪造检测技术

图像伪造检测技术的难点在于伪造痕迹往往细微且多样。通过基于局部区域特征的检测方法，侧重于对图像中的纹理、边缘和结构信息的分析，能够有效发现伪造内容中的异常。细粒度特征检测技术利用深度学习模型，自动识别图像中肉眼难以察觉的伪造痕迹，例如图像局部的颜色不一致性、模糊的边缘处理或不自然的纹理变化。

在 GAN 伪造图像的检测中，纹理分析是一个有效的手段。GAN 生成的图像通常在局部区域存在纹理的微妙差异， 特别是在细节处理上与真实图像存在不 一致。这些差异可以通过卷积神经网络 (CNN) 等深度学习算法加以捕捉，进而 鉴定图像的真实性。

此外，另一种基于光照不一致性的方法也被用于图像伪造检测。伪造图像中的光源往往与真实场景中的光照不符，特别是在多张图像合成的伪造场景中，不同物体的光照方向不一致是常见的伪造痕迹。通过计算图像中的光照方向和阴影分布，可以检测出伪造痕迹，为鉴别伪造图像提供可靠的辅助依据。

#### 4.2.1.3 视频伪造检测技术

视频伪造比图像和音频更为复杂，因为它不仅涉及图像处理，还涉及时间维度上的语义一致性问题。在伪造视频中，常见的技术手段是对视频中的特定帧进行修改，如更改人物面部表情、语音同步或动作，这些改动导致视频整体语义上的不一致。

视频伪造检测的一个重要方向是基于语义一致性的鉴伪技术。通过分析视频中的面部表情、人物情绪和动作流畅性，可以识别出伪造的部分。伪造视频通常会在面部表情、口型与语音的匹配上出现偏差，而这些偏差是人工合成视频的常见漏洞。此外，伪造视频在帧与帧之间的过渡不够自然，动作显得生硬，通过分析这些时序特征，可以有效检测出视频的伪造痕迹。

深度学习技术同样在视频伪造检测中发挥着关键作用。例如，基于循环神经网络 (RNN) 和长短期记忆网络 (LSTM) 的时序分析模型可以跟踪视频帧的时

间序列特征，从而检测出帧间不连续性和异常变化。同时，利用 3D 卷积神经网 络 (3D CNN) 捕捉视频中的空间和时间信息， 可以更精确地识别视频伪造的迹 象。

#### 4.2.1.4 多模态融合检测技术

单一模态的检测方法在面对多模态伪造内容时往往不够充分，因此，融合多模态信息的检测技术逐渐成为研究的热点。通过将音频、图像和视频等多模态信息进行融合分析，可以提高伪造内容的检测精度。例如，将图像的纹理特征与音频的频率特征相结合，通过跨模态的伪造痕迹分析，可以发现单模态检测难以察觉的伪造行为。

多模态融合技术不仅可以提高检测的准确率， 还能够有效减少误报率。例如，在视频伪造检测中，结合声音的语义分析与视频帧的图像分析，可以更精确地检测出音画不同步、情绪不一致等问题， 从而提高整体伪造检测的鲁棒性。

### 4.2.2 多模态 AIGC 音视频欺诈的鉴定技术

为了有效应对 AIGC 生成工具的欺诈风险，针对金融场景的 AIGC 鉴定技术框架需要覆盖从内容检测到预警再到系统迭代的全流程。这一框架可以通过表征学习、半监督学习、度量学习和迁移学习等关键技术，建立完整的风险内容识别和标记机制。

#### 4.2.2.1 多模态 AIGC 音视频欺诈鉴定技术的框架

**提取 AIGC 生成内容的深层特征。**表征学习是 AIGC 生成工具鉴定的核心技术之一。它通过深度学习模型，提取伪造内容中的深层特征，从而识别出 AIGC 生成的痕迹。例如，在处理伪造的财务报表或合同时，表征学习可以分析文本的语法结构、内容风格、数据分布等隐含特征。AIGC 生成的文本尽管在表面上与人工编写的文档相似，但在语言模式、句子结构上往往存在不一致性，表征学习通过深度模型捕捉这些微妙差异，能够有效识别出伪造内容。

同样，在图像和视频检测中，表征学习可以帮助模型提取纹理、边缘和局部结构的特征。例如，针对生成图像或视频中的视觉伪造痕迹，表征学习可以识别出 AIGC 生成图像在细节处理上的不自然之处，如边缘模糊、纹理不一致等。

**在有限标注数据下进行有效学习。** 金融领域的 AIGC 欺诈场景往往缺乏大量标注的训练数据，这使得传统的监督学习方法难以适用。半监督学习可以在有限标注数据的情况下，结合大量未标注的数据，自动学习生成工具的鉴定特征。通过这种方式，系统可以逐步识别并标记出伪造内容，提高模型的检测能力。

例如，针对少量已知的 AIGC 伪造财务数据，半监督学习可以利用这些标注样本作为“种子”，用于训练模型提取伪造模式的共同特征。随后，模型可以将这些特征应用于大量的未标注数据中，从而识别出更多潜在的伪造风险内容。通过这种方式，半监督学习不仅可以显著提高数据利用效率，还能够加速金融场景中 AIGC 欺诈内容的发现与预警。

**精准计算伪造内容与真实内容的差异。** 度量学习是建立在相似性分析基础 上的技术，旨在根据内容的内在特征度量出伪造内容与真实内容的差异。在金

融业务中，度量学习可以帮助模型区分正常交易记录与 AIGC 生成的伪造交易，计算出其相似度并评估风险等级。

例如，度量学习可以对生成的伪造文档或图像进行嵌入表示，将其转换为高维向量空间，并通过计算伪造内容与真实内容之间的“距离”来判断其真实性。如果两者之间的度量距离较大，则模型可以判定为潜在的伪造内容。这种基于度量学习的鉴定技术可以有效降低误报率，提升对复杂 AIGC 生成内容的识别精度。

**应对 AIGC 生成工具的快速迭代。**AIGC 生成工具的快速迭代使得传统的模型训练和更新周期难以跟上技术变化，而迁移学习为此提供了有效的解决方案。迁移学习通过将已经训练好的模型知识迁移到新的欺诈场景中，无需从零开始训练模型，大大提高了检测效率。

在金融场景中，迁移学习可以帮助系统快速适应新的 AIGC 生成工具。例如，针对不同类型的 AIGC 生成合约或伪造交易记录，迁移学习可以利用已有的语音识别、图像处理技术，迁移至新的伪造模式中，快速鉴定新型风险。这种方法不仅提升了模型的泛化能力，还确保了系统在面对未知威胁时的快速响应能力

#### 4.2.2.2 全流程 AIGC 生成工具鉴定体系

通过表征学习、半监督学习、度量学习和迁移学习等技术，可以构建出一个全流程的 AIGC 生成工具鉴定体系。该体系包括以下几个核心环节：

**检测。** 自动化检测系统基于表征学习和度量学习，实时监控金融业务中的 各类 AIGC 生成内容，提取伪造特征并进行风险评估。

**预警。** 系统一旦检测到伪造内容，将及时发出风险预警，标记内容来源并生成相应的风险报告。预警机制可以根据不同的风险等级提供相应的处理建议，帮助金融机构采取适当的防范措施。

**迭代优化。** 系统通过迁移学习不断自我优化， 确保其检测能力能够跟上 AIGC 生成工具的快速演变。随着新型 AIGC 生成工具的出现，模型可以通过迁 移已有的知识，快速适应并更新检测策略。

AIGC 生成工具的应用已经深入到金融行业的多个领域，其潜在的安全隐患不容忽视。通过开发针对 AIGC 生成工具的专用鉴定技术，金融机构可以更好地识别和标记潜在的风险内容来源，从而有效防范 AIGC 欺诈行为。利用表征学习、半监督学习、度量学习和迁移学习等核心技术，构建完整的检测、预警和迭代的鉴定体系，可以为金融业务的安全提供有力保障。

### 4.2.3 AIGC 特征的欺诈团伙识别技术

AIGC 特征的欺诈团伙识别技术，如基于多维特征关联的智能反欺诈系统，通过多维特征融合、知识图谱推理与图挖掘技术，为打击网络欺诈提供了全新的解决方案。该技术不仅可以识别出当前的欺诈行为，还能够持续优化系统模型，提升反欺诈能力，为金融行业提供坚实的安全保障。

#### 4.2.3.1 多维特征融合构建反欺诈关联网络

AIGC 特征识别是打击 AIGC 生成的虚假信息和欺诈行为的核心技术之一。通过融合以下几类多维特征，能够构建出更加精准的关联网络：

**AIGC 技术特征。**通过 AIGC 鉴定技术提取 AIGC 生成内容（如虚假视频、音频等）的独特特征，包括生成模型的指纹、虚拟人物或声音的合成特征等。这些特征能够有效区分人工生成的内容和真实数据，为后续的图计算提供基础。

**设备维度关联特征。**通过设备指纹技术，可以捕捉设备的独有信息（如 IP 地址、硬件特性、操作系统信息等）。这些信息可以识别出是否存在设备被代 理、虚拟机使用、模拟器运行等欺诈行为。

**环境维度关联特征。**涉及操作环境的相关数据，例如用户操作的地理位置、 时区、操作习惯等。当环境特征与用户的历史记录不符时，这种偏差可以用作 异常行为的标志。

通过融合上述多维特征，构建出反欺诈的关联网络。该网络不仅能够揭示潜在的欺诈团伙，还能深入分析其行为模式， 为金融行业提供强有力的风险防 控支持。

#### 4.2.3.2 用关联图谱挖掘欺诈团伙

关联知识图谱在多维数据关联中扮演了核心角色。通过构建基于 AIGC 特征的关联知识图谱，系统能够自动捕捉并分析欺诈团伙的关系网络：

**关联关系构建。**通过 AIGC 技术、设备特征、操作环境等多源数据节点之间的关联，使用图计算算法进行推理，揭示团伙成员之间的协同作案关系。这种技术特别适用于识别那些利用虚拟身份、复杂网络环境和高级 AIGC 工具掩饰其真实身份的团伙。

**图挖掘算法。**采用先进的图挖掘技术（如社区发现算法、节点重要性排序 等），从复杂的多维数据中提取潜在的团伙特征。这些算法通过分析节点之间 的频繁互动、合作行为以及共用的设备和环境特征，识别出异常团伙活动。例 如，如果多个账户使用相同的设备进行登录， 或频繁在不同时区出现，系统可 以将这些异常行为标记为潜在欺诈活动。

#### 4.2.3.3 基于团伙特征的模型训练与优化

在识别出欺诈团伙之后，这些团伙的行为特征可以进一步用于 AI 模型的优 化与再训练。

**智能决策支持。**提取的团伙特征可以作为智能决策系统的重要依据， 帮助 反欺诈系统在未来迅速识别类似的欺诈行为。同时，这些特征还可以通过 AIGC 模型的再训练过程，提升整个系统的反欺诈能力。

**模型迭代与标注。**在识别出团伙之后， 系统会将其特征作为训练数据输入 AIGC 模型中，进行标注并优化模型的检测精度。通过不断迭代，AIGC 系统能够 逐步增强对新型欺诈手段的识别能力，并应对不断变化的欺诈模式。

#### 4.2.3.4 AIGC 欺诈内容的深度识别

对于那些未知的、可能涉及新技术或工具的 AIGC 内容，特征往往隐藏得极深，传统的反欺诈检测手段往往难以应对。基于知识图谱的欺诈团伙挖掘技术，通过对 AIGC 内容的特征提取，包括 AIGC 生成特征、设备信息、操作环境等能够有效识别隐藏的欺诈行为。

基于银行等金融机构的内部业务数据，结合图计算技术、深度学习和 AIGC 算法，能够深入挖掘复杂的数据关联， 并从中提取出欺诈团伙的行为特征。这 不仅有助于识别是否存在新的欺诈手段和技术，还能够深入追踪欺诈行为的来 源及其演变过程，从而更有效地应对复杂多变的欺诈风险。

### 4.2.4 融合 AIGC 欺诈的多模态智能决策引擎技术

融合 AIGC 欺诈的多模态智能决策引擎技术，通过集成音视频伪造检测、设备维度关联特征、用户行为分析和欺诈团伙特征等多模态数据，能够为银行等业务场景提供实时的、高效流计算功能，同时具备灵活配置、可验证、可溯源等优势，实现对 AIGC 欺诈行为的全面识别和辅助决策支持。

#### 4.2.4.1 融合多模态特征的智能决策引擎

多模态智能决策引擎是通过整合不同类型的数据源和技术模块，为欺诈检测和防范提供综合分析和决策支持的系统。其核心架构包括以下几个关键部分：

**多模态数据采集层。** 包括 AIGC 检测技术（如音视频伪造特征）、设备维度关联数据、用户行为数据、环境数据等。数据采集层通过多种渠道获取各类信息，确保系统能广泛收集相关特征进行实时处理。

**特征融合与关联分析层。** 将多模态数据进行特征提取、处理与融合， 尤其 是将不同维度的数据通过特征向量化、图模型等方式进行关联分析。这一层通 过深度学习、图计算等技术，挖掘不同维度之间的潜在关联性，识别出欺诈团 伙的特征及其协同作案模式。

**流计算与实时决策层。** 流计算技术的核心是提供实时、高效的计算能力， 尤其是在欺诈检测中的低延迟需求至关重要。流计算引擎支持对大量并发数据 进行高效处理，能够在交易或行为发生的瞬间做出实时的风险评估和决策。

**智能决策与反馈优化层。** 在做出判断后， 智能决策引擎能够输出直观的风 险分析报告，并将判断结果返回业务系统。同时，该层还提供决策反馈机制， 通过持续学习和自我优化，不断提高系统的判断精度。

**可验证与可溯源机制**。针对每一次判断，系统会记录相关数据和决策流程，确保所有操作都可以追溯，方便银行在出现风险或纠纷时进行审核和验证。

#### 4.2.4.2 多维特征融合与智能决策

融合多模态特征的智能决策引擎主要具备以下核心功能，确保其在不同场景中有效识别和防范 AIGC 欺诈行为：

**AIGC 伪造内容检测。**通过先进的 AI 鉴定技术，识别音视频伪造内容，尤其是利用深度伪造技术生成的虚假信息。检测过程涵盖音频的语调、音频频谱异常、视频中的光影不一致等特征。

**设备维度关联分析。**系统会跟踪设备指纹、IP 地址、地理位置等信息，通 过关联多个设备的登录、操作历史，识别出是否存在欺诈行为。例如， 一个欺 诈团伙可能通过多个设备进行分布式操作，系统能够通过设备关联性分析发现 异常模式。

**用户行为特征识别。**该引擎通过行为分析算法， 识别用户操作习惯， 构建 用户的行为画像。 一旦某个账户的行为与其历史记录明显不符，例如异常交易 频率、异常登录地点等，系统可以及时发出风险警报。

**欺诈团伙行为挖掘。**结合知识图谱技术和图计算算法， 系统能够发现潜在 的欺诈团伙及其协同作案模式。例如， 系统可以通过分析多个关联账户的操作 行为，识别出某些账户背后可能存在的团伙作案迹象，并提取其团伙特征，帮 助制定防控策略。

#### 4.2.4.3 多模态特征的集成与优化

多模态智能决策引擎通过融合不同特征，构建了一个多维度、全方位的欺诈识别体系。以下是几个关键特征的融合与优化方式：

**音视频伪造特征与设备特征融合。**AIGC 技术的欺诈行为往往通过伪造的音视频内容进行，系统可以将伪造内容的特征与设备的使用情况结合起来分析。例如，某个设备频繁登录多个账户，并生成大量虚假音视频，这种情况可能表明该设备是欺诈团伙的重要工具。

**用户行为特征与环境特征融合。**用户的操作习惯和设备的环境信息（如网 络位置、设备类型等）结合，可以帮助系统更精确地判断异常行为。若某用户 突然在不同国家的 IP 地址上快速切换设备登录，这样的环境特征与行为特征结 合的异常模式可以快速触发风险预警。

**跨场景数据融合。**通过集成银行内部的交易数据、设备信息、客户行为记录，智能决策引擎可以通过跨场景数据分析， 发现某些复杂的欺诈模式。例如， 某账户在一天之内进行异常频繁的交易操作， 并使用不同设备登录，这些多维 度信息结合可以判断该账户可能受到了控制或是恶意操作。

#### 4.2.4.4 多模态引擎在金融业务反欺诈中的作用

多模态智能决策引擎技术在银行及金融机构的多个业务场景中，能够精准地识别欺诈行为、提高风控效率，并为金融机构提供更智能、动态的反欺诈解决方案。

**信贷业务。**在信贷业务中，欺诈行为通常表现为虚假信息提交、伪造身份 或不真实的收入证明等。多模态智能决策引擎通过整合客户的身份验证数据、
信用历史、社交行为和生物特征信息， 实时检测和分析客户的申请信息，验证客户提交的各类资料的真实性，如通过面部识别技术验证身份证照片的真实性，或通过行为分析识别客户是否为操控账户的恶意行为者，减少不良贷款的发生，降低信贷欺诈风险，提高贷款审批的效率和准确性，同时确保合规性和客户资金安全。

**信用卡业务。** 信用卡欺诈通常表现为身份盗用、卡片信息盗取和未经授权 的交易。多模态智能决策引擎结合客户的行为模式、交易历史、设备指纹、位 置数据等多种信息，对客户的消费习惯、地理位置以及交易频率进行分析，判 断是否存在异常活动有效防止信用卡盗刷和身份盗用，减少因欺诈行为而产生 的经济损失，提升客户对信用卡产品的信任。

**支付结算业务。** 支付结算环节是金融服务中最易受到欺诈攻击的部分， 常见的欺诈形式包括支付信息篡改、跨境欺诈和洗钱等。多模态智能决策引擎技术能够通过监控支付交易的多维度数据，结合交易金额、支付方式、设备指纹、用户身份等多种信息，实时判断是否存在风险。如客户在短时间内进行频繁小额支付或进行高风险的跨境支付时，系统会自动触发警报并要求二次身份验证或冻结交易，从而提高支付安全性，防止支付过程中的欺诈行为，保障金融机构和用户的资金安全。

**客户营销推广**。客户营销活动中，欺诈通常表现为恶意注册、虚假客户信息、滥用优惠活动等。多模态智能决策引擎通过分析用户的行为特征、购买习惯、社交网络互动等进行全面分析，识别异常客户，如通过分析用户注册过程中的设备信息、IP 地址和位置等，判断是否为虚假注册账户。此外，基于历史交易和互动数据，该技术可以检测是否有滥用促销活动或不正常的优惠申请行为，减少因虚假注册、滥用优惠等行为导致的营销损失，提升客户营销活动的准确性和 ROI。

**智能客服。** 智能客服在提供金融服务时，通常需要处理大量的用户信息与 请求。恶意用户可能通过伪造身份或恶意行为试图通过客服渠道进行欺诈。多 模态智能决策引擎可以通过分析用户的语音、文本和行为数据，识别客户声音 中的情绪波动、语言特征等，判断其是否存在欺诈意图。例如，如果用户的声 音与其身份不符，或者语音内容与系统记录的客户信息不匹配，系统将自动提 示进行二次验证，有效阻止通过客服渠道进行的身份冒用和欺诈行为，提高客 户服务的安全性和可信度。

**反洗钱。** 反洗钱是金融机构必须遵守的重要合规要求。通过监控客户的交易行为和资金流向，多模态智能决策引擎能够实时发现洗钱活动中的异常行为，如资金流动异常、频繁的大额交易、交易记录、客户身份、地理位置等数据综合分析，实时监测并识别是否存在洗钱行为，加强金融机构对洗钱行为的实时监控和早期发现，确保合规性，并防止金融机构因洗钱行为而遭受监管处罚或声誉损害。

多模态智能决策引擎技术能够帮助金融机构实时识别欺诈行为、降低风险并提高效率。无论是在信贷审批、信用卡监控、支付结算、客户营销推广、智能客服还是反洗钱等多个业务场景中， 它都能发挥出强大的安全保障作用，为 金融机构提供更加智能化、精准化的反欺诈解决方案。

## 4.3 从业人员能力的提升

加强金融从业人员的培训，不仅有助于提高他们对 AIGC 深度伪造欺诈的识别和应对能力，还能够帮助金融机构建立更加完善的风险管理体系，减少潜在的风险和损失。

**提升风险应对能力。** 通过系统培训，金融从业人员可以学习如何辨别深度

伪造的音视频内容。例如，了解深度伪造技术的基本原理，掌握如何识别面部识别中的细微伪造迹象（如面部表情不自然、眼部反应迟钝等）或语音中的非自然语气（如合成声音的节奏和语调问题）。此外，培训还可以帮助员工掌握如何使用反伪造技术（如换脸检测工具、深度伪造音频检测工具），及时发现并标记可疑交易或操作，避免通过 AIGC 欺诈手段导致身份盗用或资金损失。

**强化风险与决策。** 通过定期的培训，金融从业人员能够更好地理解和识别 AIGC 欺诈的潜在风险，从而在客户身份核验、交易监控等关键环节实施更严格 的审核。例如，金融从业人员可学会如何通过综合验证手段（如多因素认证、 行为分析等）来识别和防止伪造身份的风险。在处理客户请求时，经过培训的 员工能更清楚地识别风险信号，及时采取措施，如冻结可疑账户、要求二次验 证、启动人工审核流程等，从而有效减少 AIGC 欺诈带来的损失。

**增强客户信任。** 受过培训的从业人员能为客户提供更加专业的风险识别和 防范建议， 增强客户对金融机构的信任。例如， 员工可以通过解答客户对于 AIGC 欺诈的疑虑，提供反欺诈教育， 帮助客户了解如何识别深度伪造的音视频 内容，保护个人账户和信息安全；在接到客户关于可能遭遇欺诈的投诉时，员 工能够做出快速且准确的反应，向客户提供有效的解决方案和帮助，提升客户 的满意度。

**推动合规落地。** 培训有助于金融从业人员了解和遵守国家和地区的法律法 规，确保在面对 AIGC 欺诈时采取合规措施。例如，员工了解在 AIGC 欺诈情境 下如何合法获取证据，如何处理客户信息，能够避免因误操作而引发的法律纠 纷。还可以加强员工对于金融监管政策的理解，确保机构能够执行正确的反欺 诈流程和合规标准，避免因应对不当而遭受监管处罚。

**降低金融机构声誉风险。** 培训可以帮助金融从业人员在面对 AIGC 欺诈事件时，采取适当的危机应对措施，减少损失，并通过有效的客户沟通维护机构的公众形象，确保能够及时识别并防范 AIGC 欺诈行为，增强客户和市场对机构的信任度，从而在竞争激烈的市场中获得优势。

## 4.4 管理体系的提升

建立科学的管理体系，提升金融机构的反欺诈能力，是成为应对 AIGC 音视频欺诈的关键。

**构建统一标准的管理框架。** 科学的管理体系首先需要为金融机构构建一个 全方位的风险管理框架，通过综合考虑技术手段、流程控制、合规性要求和人 员管理，金融机构可以对潜在的 AIGC 风险进行全面评估。从风险的预防、监控 到应急响应，管理体系能够对 AIGC 带来的欺诈风险进行全周期覆盖，并能够建 立统一的风险评估标准和操作流程，确保在面对 AIGC 欺诈时，所有相关部门和 人员能够迅速、有效地应对。

**增强安全体系的合规性。** 应对 AIGC 欺诈风险，金融机构需要在科学的管理体系中融入法律和合规管理。随着 AIGC 技术的不断发展，相关的法律和政策也在逐步完善，金融机构必须紧跟法律合规的步伐，确保反欺诈措施的合法性和有效性。实施 AIGC 音视频反欺诈技术时，金融机构需要确保所采取的措施符合数据保护和隐私保护法律的要求，在发生欺诈事件时，能够及时通过法律手段追踪和追责，保障金融机构的合法权益。

**提升反欺诈效率。**通过实现跨部门、跨系统的数据融合，金融机构能够在 更广泛的范围内获取和分析客户数据，提升对 AIGC 欺诈的检测和响应能力。通 过加强数据的实时监控和共享，金融机构能够实时追踪客户交易行为和操作，
及时发现潜在的欺诈行为。例如，结合 AIGC 技术，系统能够识别不正常的登录行为、大额交易频繁等异常情况，实时触发反欺诈措施。通过数据整合和共享，能够跨系统、跨部门实现信息的及时流转和反馈，提升反欺诈工作的效率和精准度。

**实现全行级实时监控。**实时监控系统是科学管理体系的核心组成部分。通过引入先进的监控技术和智能化决策引擎，金融机构可以对每一笔交易、每一次身份验证和每一段音视频内容进行动态监控和评估。通过 AI 技术和多模态分析，系统能够实时监测音视频内容的真实性， 快速检测到换脸视频或伪造音频， 及时采取冻结账户、重置密码、二次验证等应急措施。结合大数据分析和行为 分析技术，金融机构可以实时识别用户行为的异常波动（如不寻常的登录地点、 频繁的大额转账等），并触发自动报警系统，迅速采取防范措施。

**提升应急响应能力。**为了应对 AIGC 带来的潜在风险，金融机构需要定期进行反欺诈演练和模拟测试，检验现有反欺诈措施的有效性。这种持续演练不仅能够帮助金融机构发现系统漏洞，还能提高员工在面对实际欺诈事件时的应急处理能力。通过模拟各种 AIGC 欺诈场景，金融机构可以识别现有防护体系中的漏洞和不足，及时进行技术更新和流程优化。模拟演练能够提高员工的应急反应能力，帮助他们熟悉反欺诈操作流程，确保在面对实际欺诈时能够高效处置。

## 4.5 法律法规护航

### 4.5.1 针对 AI 滥用的法规

中国作为 AIGC 领域的全球领导者，已经采取了多项措施来应对 AIGC 带来的安全挑战， 并逐步建立起了一套 AIGC 安全治理的法律框架。然而， 面对 AIGC 技术日新月异的发展，相关法律体系仍有待进一步完善。

中国是最早针对 AIGC 和深度合成技术制定法规的国家之一 。2023 年 10 月 18 日，中央网信办发布《全球人工智能治理倡议》，这标志着中国在全球 AIGC 治理中占据了重要的主导地位。该倡议提出了明确的 AIGC 安全治理方向，包括建立风险等级测试评估体系、分类分级管理、提高人工智能的可解释性和可预测性等。这些措施不仅有助于保障 AIGC 技术的安全发展，还为未来的 AIGC 立法提供了基础。

此外，中国在 2021 年和 2023 年相继出台了一系列与 AIGC 安全相关的法规，包括《互联网信息服务算法推荐管理规定》、《互联网信息服务深度合成管理规定》和《生成式人工智能服务管理暂行办法》。这些法律对生成式 AIGC 和深度合成技术的研发和应用进行了详细的规范， 尤其是对涉及社会舆论或具有社会动员能力的 AIGC 服务，要求其进行安全评估并履行备案手续。

国内的 AIGC 法律体系主要由几部重要的法律法规构成。除了针对 AIGC 和深度合成技术的具体规定，还涉及网络安全、数据安全、个人信息保护、知识

产权等多个方面的法律。这些法律包括《中华人民共和国网络安全法》、《中华人民共和国数据安全法》、《中华人民共和国个人信息保护法》、《版权

法》、《中华人民共和国科学技术进步法》和《中华人民共和国民法典》。这些法规为 AIGC 的研发、应用及其衍生问题提供了法律依据和保障。

中国在 AIGC 安全治理上采取了一系列实际的措施，以确保 AIGC 技术的安全可控。这些措施包括推动 AIGC 技术的可解释性与可预测性，确保 AIGC 技术处于人类的控制之下。相关企业和研发主体必须确保 AIGC 技术的透明性和可审查性，并在应用中保护个人隐私和数据安全，避免非法数据的收集和使用。

### 4.5.2 针对违法者的惩罚

国内在金融领域防止 AIGC 滥用方面的法律法规已经取得了初步成效，涵盖了数据安全、个人信息保护、AI 生成技术监管、防范金融诈骗等多个领域。随着 AIGC 技术的持续发展，金融领域的监管框架也将不断优化，以确保 AIGC 技术在促进金融创新的同时，能够有效防范潜在风险和滥用行为。

《中华人民共和国数据安全法》、《中华人民共和国个人信息保护法》这两部法律构成了防止 AI 滥用的基础。数据安全法要求金融机构必须确保其处理 和储存的数据安全性。AIGC 系统依赖于大规模的数据集，而这些数据集必须得 到严格的保护，不能被用于非法目的或滥用， 如未经许可的数据分析或生成不 良的金融决策。个人信息保护法进一步规定，金融机构在使用 AI 技术时，必须 确保用户个人数据的隐私和安全。任何涉及个人敏感信息的数据处理活动，如 信用评分或风险评估，都需要用户同意，并遵守隐私保护要求。

《生成式人工智能服务管理暂行办法》主要针对生成式 AI 的应用，要求金 融机构在使用具有生成和社会动员能力的 AI 服务时，必须进行安全评估，并遵 循国家有关规定。该规定旨在防止 AIGC 生成虚假信息、操纵市场或扰乱金融秩 序。该《暂行办法》对具有舆论属性和社会动员能力的 AIGC 服务进行了更加严 格的监管，要求企业在提供此类服务时进行安全评估，并履行算法备案、变更 或注销备案手续。这些法规强化了对 AI 系统的监管，并为未来可能出现的新问 题预留了法律空间。

此外，《中华人民共和国刑法》对通过 AIGC 技术进行非法活动进行了处罚规定。编造虚假信息、通过信息网络或其他平台传播的行为可以被处以有期徒刑或罚金，而《中华人民共和国民法典》则对个人隐私、肖像权和名誉权提供了法律保护。利用深度伪造技术侵犯他人权利的个人可能面临民事赔偿责任。

# 第五章 AIGC 音视频反欺诈技术实现

## 5.1 AIGC 音频伪造检测

AIGC 语音检测技术旨在提高系统对各种类型的伪造语音的检测能力，尤其是在面对不断改进的伪造语音技术时， 仍能保持较高的准确性和可靠性。本小 节将从伪造语音特征线索和鲁棒性建模两方面梳理现有的伪造语音鉴别技术， 揭示伪造语音特征的痕迹或异常，以应对不同生成技术和复杂噪声干扰。

### 5.1.1 语音伪造线索

在 AIGC 语音检测中，伪造线索指的是能够揭示语音是否由人工智能生成或篡改的音频特征。这些伪造线索通常体现在音频质量、自然度、声纹特征和频谱分析等几个方面。

音频质量指的是音频信号本身的清晰度、纯净度和信号质量等方面的综合指标。在伪造语音中，通常会出现一些音频质量的异常或瑕疵。伪造语音的生成过程可能会引入噪声、失真或其他影响音频质量的因素，这些特征可以作为检测伪造语音的一个重要线索。影响音频质量的因素主要有：

**噪声。**伪造语音中可能会包含一些不自然的背景噪声或者音频伪造过程中引入的杂音。

**失真。**伪造语音可能会在音频波形中产生某些突变或不连续的现象，尤其是在快速变化的音节部分，如爆破音、辅音等。

**频谱不一致。** 伪造语音的频谱图可能表现为频率分布不规则， 尤其是高频 部分。

**编码和压缩失真。** 在生成或传输过程中，伪造语音可能会遭遇压缩或编码， 从而丧失一些高质量的音频细节。

自然度是指语音听起来是否像人类自然发出的声音。人类语音的自然度通常具有一些固有的特点，比如语速、语调、音色等方面的自然变化。伪造语音在这些方面可能会表现出不自然的模式，导致它们听起来不像真正的人类语音。影响自然度的因素主要有：

**语音流畅性**。伪造语音可能出现发音不连贯、停顿不自然或语速不一致的现象。

**语调和重音**。人类在说话时会有一定的抑扬顿挫， 而伪造语音可能会缺乏 这种自然的韵律。

**情感表达** 。伪造语音通常无法准确模拟情感变化，听起来更加平淡或呆板。

**语速和停顿**。伪造语音可能会出现语速过快或过慢的现象， 或者在不合适 的地方停顿。

声纹特征是指能够唯一识别和区分每个人的语音中独特的生理和行为特征。由于每个人的声带结构、发声习惯、口腔构造等生理特点不同，因此产生的声音也具有独有的特征，这些特征被称为声纹。声纹特征在语音识别、身份验证和伪造语音检测等领域具有广泛应用。人工智能合成的语音与真实人的语音在

声纹特征上存在明显的差异，AI 合成语音在语音的发音习惯、语速、音高等方面的变化通常比人类语音更加规律化或机械化，缺乏自然人的个性化声纹。

频谱分析是一种将语音信号从时间域转换到频域的技术，通过频谱来分析信号的频率成分及其变化情况。在语音处理、信号处理和语音伪造检测中，频谱分析是一项核心技术，因为它能揭示信号的频率结构，从而提取到音频的独特特征。通过分析语音的频谱图，人工智能生成的语音通常会在高频或低频段表现出与真实人声不同的特征。例如， 生成语音的频率分布可能不自然，偏离 真实人类语音的自然波形特性。

### 5.1.2 线索建模方式

基于上述这些语音伪造线索，AIGC 语音检测的建模方式主要可以分为基于特征工程的传统方法和基于深度神经网络的方法。


图 5-1 伪造语音检测常用结构

基于特征工程的传统方法是伪造语音检测的早期手段，这类方法通过从语音信号中提取人工设计的特征，并利用这些特征进行分类，来判断语音是否为伪造。传统方法主要依赖于信号处理领域的一些经典特征，这些特征能反映出语音信号的物理特性或统计模式。常见的特征包括：

**梅尔频率倒谱系数（MFCC）。** 通常在检测伪造语音时，MFCC 特征中某些频段的异常变化可以作为判断的依据。

**声纹特征**。伪造语音通常难以完全模仿真实人类的独特声纹特性。

**时频特征**。伪造语音的时域特征可能较为机械化，且缺乏真实人类语音的自然变化。通过这些特征训练传统的机器学习模型（如 SVM、随机森林等）进行分类。检测模型通过这些手工设计的特征来判断语音是否为人工智能合成。

基于深度神经网络的方法在 AIGC 语音检测中取得了显著成效，能够识别出合成语音的特征，从而有效区分人工语音和伪造语音。由于 AIGC 语音生成技术越来越逼真，深度学习在 AIGC 语音检测中发挥着至关重要的作用。方法主要包括以下几个方面：

**频谱图分析和卷积神经网络** 。AIGC 语音检测中，频谱图（如梅尔频谱图和 STFT 频谱图）常作为语音数据的主要表示方式。通过将频谱图输入卷积神经网络，可以让模型自动学习到频谱图中的伪造特征。

**时序特征分析与循环神经网络**。循环神经网络 RNN 及其改进模型（如 LSTM 和 GRU）擅长处理时序数据，能够捕捉到语音信号的时间依赖特性。时序信息可以揭示出语音生成过程中的异常，例如发音不自然、节奏不一致等。

**生成对抗网络（GAN）** 。生成对抗网络不仅能生成伪造样本，还可以用于检测伪造内容。

**Transformer 模型与自注意力机制** 。Transformer 模型通过自注意力机制， 能够同时关注语音序列中的不同位置， 非常适合用于分析语音生成过程中的整 体特征。尤其对于长语音数据， Transformer 能够捕捉语音中的长程依赖性， 是 AIGC 语音检测的有效工具。

此外，近年来的研究也探索了将语音信号与其他模态数据（如视频、文本等）结合，进行跨模态检测。这种方法利用多种信号源之间的关系，提高了模型的检测能力和准确性。

AIGC 语音检测技术通过捕捉伪造语音的特征线索并构建强大的检测模型，能够有效应对多样化的生成技术和复杂的噪声干扰。采用多层次特征融合、对抗训练、时序建模等技术手段，确保模型具有更高的检测精度和更强的泛化能力。在未来的研究中，结合多模态信息和更先进的深度学习技术，将进一步提高伪造语音的检测性能。

## 5.2 AIGC 图像伪造检测

AIGC 图像检测任务旨在判断给定的图像是否由人工智能生成或篡改。本小节将从伪造线索和建模方式两方面梳理现有的 AIGC 图像检测技术，前者关注那些能够揭示图像被人工智能生成或篡改的特征或痕迹，后者关注利用伪造线索构建检测模型的具体方法和技术。

### 5.2.1 图像伪造线索

AIGC 图像中的伪造线索主要包括视觉伪影、数字信号异常特征、模型指纹、人脸先验约束、物理成像法则违背等几个方面。

视觉伪影指 AIGC 图像生成过程中产生的不自然视觉效果，其可能是由多种因素造成的，包括但不限于算法局限、训练数据不足、计算资源限制等。 AIGC 图像中的视觉伪影主要表现为以下几类：

**不自然的纹理。** 生成模型可能在训练数据中学习到了某些特定的纹理模式， 并在生成新图像时过度使用这些模式，导致出现重复或不自然的纹理。

**边缘扭曲或模糊。**AI 模型在生成图像时，可能会在边缘处理上不够精细，导致边缘模糊或者过于锐利，与周围环境不协调。

**颜色分布异常。**AI 模型可能会在颜色渲染上存在偏差，导致某些颜色过于饱和或者不自然地分布在图像中。

**几何畸变。**在生成具有复杂几何形状的对象时，AI 模型可能会产生不自然的变形或者比例失调

**过度平滑。** 为了减少计算复杂度，AI 模型可能会在图像中使用平滑技术，但过度平滑可能会使得图像失去细节和纹理的自然变化。

AIGC 图像还可能在数字信号层面留下异常特征，包括频域、噪声域和色彩统计特征等。在频域上，AI 生成的图像在频域中可能表现出特定的模式，这些模式可能与真实图像的频域特征存在显著差异。例如，GAN 生成的图像在频域中可能存在由上采样操作引入的伪影，这些伪影可以作为检测的依据。在噪声域中，真实图像的噪声通常具有特定的统计特性，而 AIGC 图像的噪声模式可能与真实图像不一致。通过分析图像的噪声模式，可以识别出 AIGC 图像。此外，

在色彩统计特征方面，AIGC 图像在色彩分布上可能与真实图像存在差异。例如，生成图像可能在某些颜色通道上表现出不自然的集中或稀疏，这些差异可以通过色彩统计分析来识别。

模型指纹是指 AI 生成模型在其输出结果中留下的独特标记或特征，这些特征可以用来追溯图像是否是由特定的生成模型所创造。这些指纹是由模型的架构、训练数据、参数设置等因素共同决定的。通过训练专门的检测模型来识别这些模型指纹，可以判断一幅图像是否由特定的 AI 模型生成。

人脸先验约束是指人脸图像中固有的结构、比例、纹理等特征，这些特征在真实的、未经修改的人脸图像中通常是稳定且一致的。然而，AI 生成的人脸图像由于生成算法的局限性和训练数据的不足，可能无法完全准确地复现这些先验知识，进而导致生成图像中出现违背先验知识的异常。例如，眼睛的位置可能不准确、瞳孔形状不规则、鼻子的形状可能不自然、皮肤纹理可能过于平滑、头部朝向和脸部朝向不一致等。


**图 5-2 姿态一致[6]性/瞳孔形状不规则[7]/左右眼角膜高光一致性[8]**

物理成像法则是指光在传播、反射、折射等过程中遵循的物理定律，如光的直线传播、反射定律、折射定律等。这些定律在自然界中是普遍存在的，也是真实图像形成的基础。而 AIGC 图像由于是由算法生成的，其生成过程中可能并未严格遵循这些物理定律，因此可能在图像中暴露出不符合物理规律的线索。例如，出现光源不一致、阴影方向和长度不符合物理规律、透视关系不符合几何学原理等。

### 5.2.2 线索建模方式

针对 AIGC 图像中伪造线索的建模方法，主要分为手工构造与表征学习两大类。手工构造方法基于对伪造图像生成机制及特性的深入理解，通过人工设计特定特征来识别伪造痕迹。此类方法对数据量的需求不高，通常专注于某一特

定伪造特征，尽管建模过程较为复杂， 但具备较高的透明度与可解释性。这种方式的核心在于“特征工程”，即根据伪造图像中的伪造线索，人工设计并提取出能够区分真实与伪造图像的特征。典型方法包括纹理分析（如灰度共生矩阵和局部二值模式）来提取纹理信息； 分析颜色分布、饱和度及亮度，识别可能存在的颜色失真；以及通过边缘检测和形状分析提取异常轮廓。此外，图像的统计量计算（如均值、方差、直方图）也能揭示真实与伪造图像在统计分布上的差异。提出特征后，可以通过传统的机器学习模型（如支持向量机、决策树等）完成伪造图像的检测。然而，其局限性在于难以全面涵盖所有伪造类型，且随着伪造技术的持续演进，需频繁更新特征集以保持有效性。

表征学习采用深度学习模型，能够从大规模、多样化数据集中自动提取多层次、有效的特征表示。其优势在于适用于类型多样的伪造痕迹检测，涵盖常见与复杂的伪造痕迹，同时适应光照变化、遮挡及图像降质等复杂场景。该方法通常结合特定任务特点，从网络架构、损失函数、数据增广及训练策略等方面优化。在网络架构方面，设计了专门针对伪造特征提取的模块，例如结合提取层次化特征、引入注意力机制关注局部关键特征等；在损失函数的选择上，则注重于强化模型对于细微伪造差异的敏感度；数据增广技术的应用旨在增加模型面对未知情况时的鲁棒性；在训练策略方面，通过调整学习率、正则化和归一化等方式提高模型的稳定性和性能。此方法简化了建模流程，提高了模型的泛化能力，但同时也牺牲了一定程度的透明度与可解释性。

总体而言，表征学习因其自动化、高效能及强大的适应性，在处理日益复杂多变的 AIGC 图像伪造挑战中展现出显著优势，成为当前研究的主流趋势。尽管如此，手工构造方法因其独特的透明度和可解释性，在特定应用场景中仍具有不可替代的价值，两者相辅相成，共同推动着图像伪造检测技术的不断发展。

## 5.3 AIGC 视频伪造检测

针对视频伪造过程中容易破坏语义特征一致性这一特点，研究基于视频语义一致性的鉴伪技术，研究对视频中的目标进行识别与分割，对目标和目标进行纹理、光照、分辨率等特征进行提取，结合这些混合特征进行分类检测通过检测目标网格图像分割帧与帧之间移动的不一致性进行伪造识别。这种方法能够捕捉到伪造过程中容易被忽略的动态语义异常，为视频伪造的精确检测提供了技术支撑。

AIGC 视频检测技术， 其核心目标在于精准判断给定的视频内容是否由人工智能生成或经过篡改。本小节同样从伪造线索和建模方式两方面梳理现有的 AIGC 视频检测技术。在伪造线索方面，AIGC 视频检测不仅需要识别静态图像中的各种伪造线索，还需考虑视频特有的时空连续性、视听不一致性等伪造线索。在建模方式方面，AIGC 视频检测还需进一步考虑视频特有的复杂性。

### 5.3.1 视频伪造线索

AIGC 视频中的伪造线索除了静态图像中的伪造线索，还包括时序视觉伪影、视听不一致性、人脸先验约束、运动轨迹自然性等几个方面。

其中，AIGC 视频中的时序视觉伪影主要表现为以下几类：

**时空连续性不一致** 。AIGC 视频可能在相邻帧之间出现不自然的过渡，如物体位置突然跳跃、背景信息不一致等。

**光照和阴影一致性。**AIGC 视频在生成时，可能无法准确模拟复杂的光照和阴影变化，特别是在动态场景中，光源方向或强度的变化可能导致阴影不一致。

**边界和边缘异常。**AIGC 视频生成时，可能无法完美地处理物体或人物的边界，导致边缘出现锯齿状或模糊现象。

**纹理和颜色突变。**AIGC 视频中的纹理和颜色可能在帧间或帧内出现突变，表现出局部区域的颜色或纹理与周围环境不一致。

真实视频中的视觉和听觉信息应保持高度的同步和协调，而 AIGC 视频可能出现视听不一致性。视听不一致性主要表现为以下几类：

**音画协同性较差。** 视频中的音频和视频内容应在时间上保持一致， 不应有 明显的延迟或错位。导致音频和视频内容的不匹配，如人物动作与背景音效的 不协调。

**音视频情感不一致。** 视频中人物的情感表达应与语音的情感相一致， 如高 兴的表情与欢快的语气，不能出现人物的表情与语音情感不匹配。

**音唇不一致。**表现为人物说话时嘴唇动作与声音不同步。

**身份不一致。**表现为视频中人物的声音与形象不匹配。

**语义内容不一致。**AIGC 视频中的对话内容应与人物的肢体动作、表情或唇部动作所传达的语义内容相吻合。语义内容不一致时，表现为这些元素之间的矛盾。

**音视频环境一致性。**视频中的音频环境应与视觉环境相一致， 如室内环境 的回声与视频中的室内场景相符。环境不一致表现为音频环境与视觉环境不匹 配。

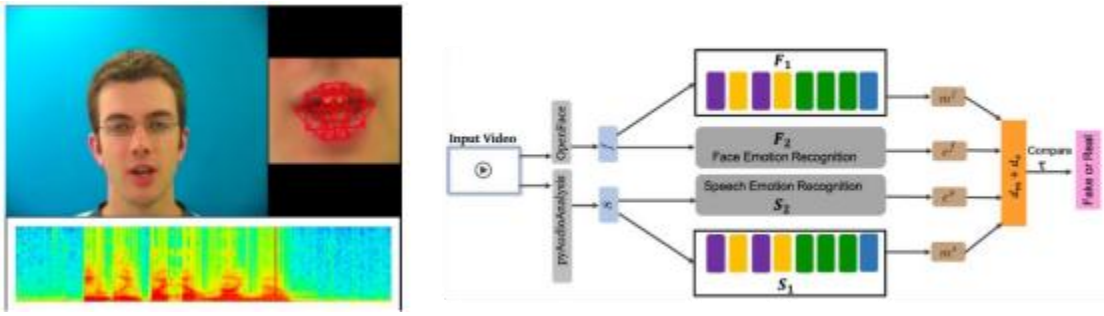


**图 5-3 音视频语义内容一致性[9]/音频视频情感一致性[10]**

AIGC 视频中也会违背一些先验人脸约束，主要包括：

**表情不自然。**AI 在模拟人类面部表情时，可能会因为缺乏对人类情感表达的深刻理解，而导致生成的表情显得生硬、夸张或不符合常理。

**五官比例失调。** 人脸的五官比例是长期进化过程中形成的自然规律，但在 AIGC 视频中，由于算法或数据的问题， 生成的人脸五官比例可能出现失调现象。

**身份特征不一致。** 在连续的视频帧中，AI 生成的人脸应保持一致的身份特 征， 包括脸型、五官形态等。然而， 由于算法的不稳定性或数据的局限性， AIGC 视频中的人脸可能在某些帧中发生微妙的变化，导致身份特征不一致的现 象。

**眨眼与头部运动不自然。**真实人类在交流过程中会自然地眨眼和进行头部运动，这些细微的动作对于维持视觉的真实感至关重要。

此外，AIGC 视频还会存在违背运动轨迹自然性的伪造线索，AI 生成的物体运动可能显得僵硬、不流畅，或者不符合物理运动规律，如速度变化不连续、加速度异常等。

### 5.3.2 线索建模方式

针对视频伪造线索的建模方式，可以借鉴并扩展 AIGC 图像伪造线索建模的方法论，将其应用于更为复杂和动态的视频数据中。视频伪造线索建模同样可以划分为手工构造与表征学习两大类， 但考虑到视频的时间维度、运动信息及 连续性等特点，这些方法需要相应调整和优化。

在视频伪造线索的手工构造建模中，核心仍然是“特征工程”，但特征的设计需更加关注视频特有的属性。例如，通过光流法、轨迹跟踪等技术分析视频中物体的运动轨迹，检测是否存在不自然的运动变化或速度突变；分析视频帧间的时空连贯性，包括颜色、亮度、纹理等特征在连续帧中的一致性，以及物体形态和位置的连续性变化，可以识别可能的伪造区域；验证音频与视频内容的同步性，可以检验视频是否被编辑过；在重力、光影变化、透视原理等方面的检查，可以识别违背自然规律的场景。

在视频伪造线索的表征学习中，核心在于利用深度学习模型自动从大规模、多样化的视频数据中学习到有效的特征表示。与图像伪造线索的表征学习相比，视频伪造线索的表征学习需要特别关注视频的时间维度、运动信息及连续性等特点。例如，通过 3D CNNs 和 RNNs 的结合， 同时捕捉视频的空间和时间特征；
通过多模态学习和跨模态一致性验证，确保视频中的视觉和听觉内容保持一致。

## 5.4 AIGC 欺诈鉴定技术

现有的 AIGC 图像和视频检测工作大部分聚焦于真假判别任务，即判断给定的图像或视频是否被伪造。AIGC 生成工具鉴定则是一种更精细的任务，需要分析伪造图像或视频背后的生成细节，包括合成方法、网络结构等。这类技术可以提供伪造数据的更多历史信息，增强真假判别结果的可信度。根据能否预先接触待溯源的目标数据，AIGC 生成工具鉴定技术可以被划分为被动式溯源和主动式溯源。

### 5.4.1 被动式溯源

被动式溯源按照溯源粒度可以进一步划分为方法级溯源、结构级溯源、模型级溯源和超参数级溯源，任务难度由易到难。方法级溯源主要集中在识别伪造内容所采用的具体技术或算法。Jia 等人[11]提出的 DMA-STA 方法通过结合空间注意力机制和时序注意力机制来提取和聚合视频帧的特征，最终实现对伪造视频的多分类。Ciftci 等人[12]则通过放大视频中的人脸肌肉运动，利用 3D 卷积神经网络（3DCNN）来识别伪造内容。 Girish 等人[13]关注到了快速迭代的 GANs 方法带来的挑战，提出了一个开展深度伪造溯源任务及解决方案，旨在动态地识别新出现的未知 GANs 方法。而 Narayan 等人[14]引入了“深伪种系”

的概念，强调了一张人脸可能经过多次不同的伪造处理，因此将深伪溯源问题定义为多标签分类任务， 即给定一张图像， 输出所有使用的伪造方法类别。

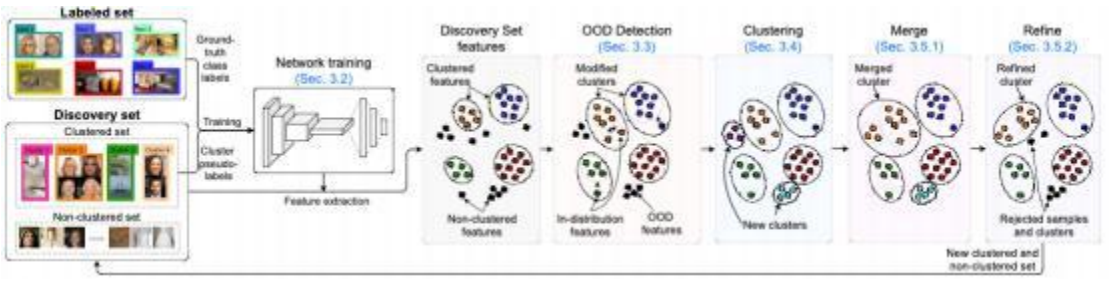


**图 5-4 开集深度伪造溯源方法流程图**

结构级溯源旨在确定伪造内容背后的特定神经网络架构。Yang 等人[15]通过实验证明了神经网络结构指纹的存在，并开发了 DNA-Det 网络来学习和识别这些结构指纹，这有助于更深层次地理解伪造内容的生成过程。

模型级溯源进一步细化到识别具体的模型实例，即识别出用于生成伪造内容的具体模型（包括其权重和参数）。Yu 等人[16]通过实验证明了 GAN 模型指纹的存在性及其独特性，并提出了一种基于自编码器的方法来提取和利用这些指纹进行溯源。Guarnera 等人[17]构建了一个专门的数据集， 用于研究不同模型实例之间的细微差异，并提出了一种有效的模型识别方法，该方法在特定数据集上展现了良好的识别性能。

超参数级溯源是最细粒度的溯源层次，目标是从伪造内容中反推出生成模型时使用的具体超参数设置。Asnani 等人提出了一种框架，该框架包括两个主要组件：指纹估计网络（FEN）和解析网络（PN），前者用于从生成的图像中估计出生成指纹，后者则负责从这些指纹中解析出模型结构和损失函数等超参数信息。

### 5.4.2 主动式溯源

主动式溯源要求在图像或视频生成过程中预先嵌入特定的指纹信息，推断时通过提取这些指纹信息来获取伪造数据的生成细节。这种方法依赖于制作深度伪造时嵌入的隐藏签名。这种隐藏的签名信息伴随着图像或视频的整个生命周期，因而推断时只需将签名信息提取出来就能得到相关信息。

主动式溯源通常通过数字水印（Digital Watermarking）技术和神经网络 水印（Neural Network Watermarking）技术来实现。数字水印技术将信息直接 嵌入图像的像素中，而神经网络水印技术则将信息嵌入到神经网络的参数里。 Yu 等人[8]提出了一种基于神经网络水印的方法，该方法首先通过训练指纹编 码器将指纹信息编码到训练数据中，随后训练 GAN 模型。在生成过程中， GAN 的生成器会生成包含指纹信息的图像， 最终通过指纹解码器提取出嵌入的指纹 信息，从而完成伪造内容的溯源。

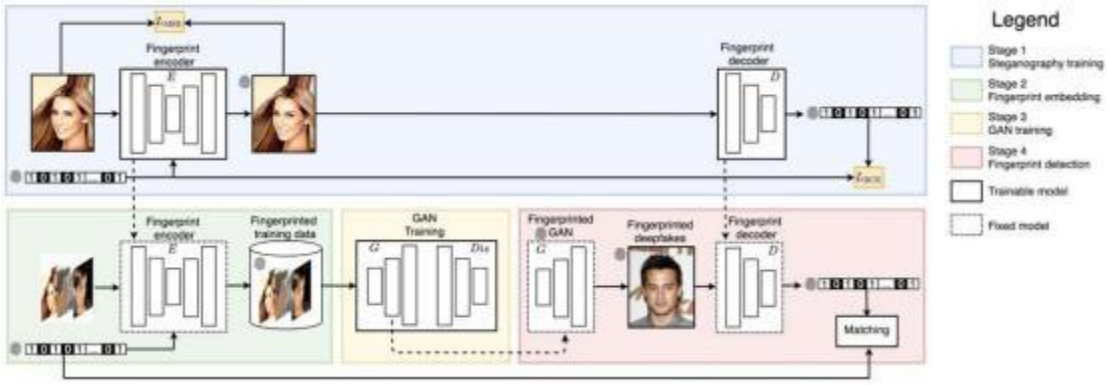


图 5—5 基于神经网络水印的主动式溯源方法

## 5.5 基于知识图谱的特征关联分析

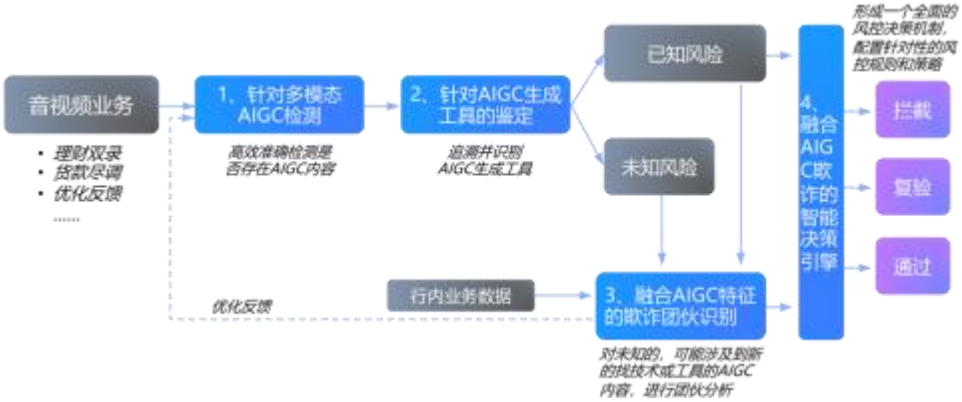


图 5—6 业务流程图

基于知识图谱的 AIGC 特征关联性分析是一种强大的图结构数据模型，通过构建与深度学习、特征提取和指纹识别技术相结合的知识图谱，用图形的形式将实体（如人物、事件、地点等）之间的关系进行结构化描述，揭示不同数据对象间潜在的复杂联系。这种方法不仅能有效识别和追踪 AIGC 生成内容中的潜在欺诈团伙，还能够揭示多个欺诈个体之间的联系和行为模式。

**揭示隐藏的关系。**通过将各类 AIGC 生成的内容（如虚假视频、伪造音频等）作为图谱中的节点，并通过特征匹配、生成工具搭建节点间的关系，知识图谱可以揭示不同欺诈行为之间的潜在关联。例如，多个虚假账户通过相似的生成工具、相近的生成时间或一致的伪造特征连接成一个欺诈团伙的潜在网络。

**帮助识别团伙组织模式。** 通过社群发现算法（如 Louvain 算法、Label Propagation 算法等），可以在图谱中识别出高密度的节点社群，这些社群可能代表着有组织的 AIGC 特征欺诈团伙。例如，某些欺诈团伙可能通过类似的生成工具进行内容制作，或在特定的时间窗口内频繁进行虚假内容生成，形成具有明显联系的社群。

**动态追踪和推理。** 知识图谱不仅能表示当前的关联性，还可以进行动态推 理，发现潜在的隐性关系。通过关联推理，可以追踪 AIGC 生成内容在图谱中的

传播路径，进一步发现潜藏的团伙成员，甚至识别出新的欺诈行为模式。

### 5.5.1 基于 AIGC 特征的关系建立

在 AIGC 特征欺诈团伙的识别中，首先需要从 AIGC 生成内容中提取出独有的特征信息，作为构建知识图谱节点的依据。这些特征可以包括但不限于以下几类：

**音频频谱特征。**音频伪造（例如语音克隆或 AI 合成语音）生成的音频频谱具有某些特征，可以通过频谱分析算法（如 MFCC、Chroma 等）提取音频的频率、 音高、语速、韵律等信息。这些特征是识别伪造音频的关键。

**人脸微表情特征。**在 AI“换脸”生成的视频中， 人脸微表情（如眨眼、口 唇动作等）往往呈现出人工智能生成的失真特征。通过计算机视觉算法提取面 部表情和微表情的特征，可以有效识别 AIGC 生成的人脸伪造内容。

**设备的指纹特征。**每种生成工具（如 DeepFaceLab、GAN 等）都有独特的生 成痕迹和指纹。通过指纹识别技术，可以从 AIGC 生成的内容中提取出工具的使用痕迹，从而识别出相同工具生成的虚假内容。

**时间和行为特征** 。欺诈团伙往往会在特定的时间段内频繁进行欺诈活动 。 通过对生成内容的时间戳、频率等信息进行分析，可以识别出是否存在团伙行 为的规律。

将上述特征映射到知识图谱中， 可以形成多个节点， 每个节点代表不同的 虚假账户、伪造合约、篡改证据等行为，每个节点通过特定的关系（如相似的 生成特征、相近的时间生成等）进行连接，形成知识图谱的边。这些关系不仅 反映了个体之间的相似性，还揭示了可能的组织结构和行为模式。

### 5.5.2 发现与识别团伙欺诈

构建基于 AIGC 特征的知识图谱后，可以通过社群发现算法分析图谱中潜在的高密度节点群体，通过关联推理进行可疑行为模式标记，以及通过追踪节点关系进行团伙扩展与路径追踪，从中识别出可能的欺诈团伙。

**高密度节点社群识别。**社群发现算法的基本思想是，图中节点间的边越多， 越可能代表一个紧密的群体。使用社群发现算法（如 Louvain 算法、Girvan–Newman 算法等），可以识别出高密度的节点社群。社群中的节点通常具有相似的特征或行为，如使用相同的 AIGC 生成工具、相似的伪造特征等，这意味着它们很可能属于同一个欺诈团伙。

**可疑行为模式标记。**通过图谱中的关联推理，能够发现同一社群内多个节 点之间存在重复或一致的行为模式。例如，同一社群中的多个节点可能频繁使 用相同的音频伪造特征或人脸“换脸”特征， 或者在相似的时间内进行高频次 的虚假内容生成。通过标记这些可疑的行为模式，能够有效识别潜在的欺诈团 伙。

**团伙扩展与路径追踪。**通过追踪节点之间的关系， 可以发现特定特征在图 谱中的传播路径，进一步挖掘潜在的团伙成员。例如，某些节点可能与多个虚 假账户存在关联，而这些账户又与其他疑似欺诈行为密切相关，通过追踪这些 节点间的传播路径，可以揭示出完整的诈骗网络。

### 5.5.3 提升反欺诈的能力

通过基于知识图谱的 AIGC 特征关联性分析，能够在以下几个方面提升反欺诈能力：

**高效识别欺诈团伙。知识图谱**通过节点和边的关系，可以快速识别出潜在的欺诈团伙，避免传统方法中对个体行为的逐一排查。同时，依托社群发现算法，能够自动化地发现隐形的犯罪网络，提升识别效率。

**动态监控与实时预警。**知识图谱具有实时更新的特性，能够随着新数据的加入不断更新图谱结构，实时反映 AIGC 生成内容的变化，及时发现新的团伙成员和新的欺诈行为模式。一旦识别到潜在的欺诈团伙，系统可以立即发出预警，为监管机构提供阻止欺诈行为扩散的机会。

**精准防范与风险管理。**通过对知识图谱的深入分析，企业和监管机构能够更准确地预测和防范 AIGC 特征欺诈团伙的活动，从而优化风险管理策略，减少经济损失。

基于知识图谱的 AIGC 特征关联性分析，通过构建与深度学习、特征提取和指纹识别技术相结合的知识图谱，能够有效揭示多个欺诈个体之间的潜在关联性，识别有组织的欺诈团伙及其行为模式。

## 5.6 融合反 AIGC 欺诈计算引擎的处理系统

融合反 AIGC 欺诈技术引擎的高性能实时流处理系统，通过集成多种数据源（如设备、图像、音频、文本等），运用实时流处理技术、跨模态特征分析以及智能决策引擎，能够实时检测、评估和识别 AIGC 生成内容中的欺诈行为，从而在复杂的欺诈环境中提供更为精准的防护。

### 5.6.1 数据采集与预处理

融合反 AIGC 欺诈技术引擎的实时流处理系统由多个功能模块构成，涵盖数据采集、特征提取、规则匹配、风险评估、决策引擎和风险响应等多个环节。系统的设计目标是实现对不同模态数据进行实时分析，并根据检测结果进行精准的欺诈判定和风险决策。

系统首先从多个数据源获取输入信息，包括但不限于以下几种类型：

**设备数据。包括**设备指纹、IP 地址、设备型号、操作系统等信息，通过设备行为模式判断是否与以往行为存在差异。

**图像数据。**通过计算机视觉技术对视频帧、图像进行分析，检测人脸“换脸”、图像篡改等伪造特征。

**音频数据。**通过语音识别和声纹分析技术检测伪造音频或语音生成内容，识别生成的语音与实际语音的差异。

**文本数据。**对生成的文本进行语义分析，检查是否存在由 AI 生成的钓鱼邮件或伪造合同等。

这些数据会经过预处理步骤，进行清洗、去噪和格式化，以便后续处理。数据预处理阶段还包括特征提取，主要是从视频、音频、设备和文本中提取出

关键特征信息（如音频频谱特征、人脸微表情特征、设备指纹等）。

### 5.6.2 特征与规则

不同模态的数据（设备、图像、音频、文本）具有不同的特征，系统需要对每种特征分别进行单独和综合的分析。

**设备分析。**通过对设备行为的模式分析，检测是否存在不合常规的行为，如频繁变更 IP 地址或使用不同的设备指纹。

**图像分析。**使用计算机视觉技术（如人脸识别、图像一致性检测）分析图像中的伪造痕迹，如“换脸”、虚假修图、图像变形等。

**音频分析。**通过声纹识别和语音生成模型分析音频中的特征，识别是否为 AI 合成语音或深度伪造的音频内容。

**文本分析。**通过自然语言处理（NLP）技术对文本内容进行检测，识别是否为 AI 自动生成的钓鱼信息或虚假合约。

在特征提取后，系统将这些特征与规则库中的预定义规则进行比对。这些规则可以包括基础规则（如音频频谱特征是否正常）以及复杂规则（如多个特征之间的关联性、跨模态特征的匹配等）。

### 5.6.3 智能决策引擎与风险评估

当系统通过规则匹配识别出潜在的欺诈特征后，将对相关行为进行风险评估。评估过程中，系统将根据多个维度的分析结果计算出一个综合风险评分。这包含生成的音频或图像与真实内容的差异程度、行为模式的异常性、模态特征的一致性、历史行为的对比等。最后，系统通过多层次的计算模型（如决策树、随机森林、神经网络等）进行风险评估，区分潜在的欺诈行为与正常行为区分开。

### 5.6.4 实时响应与行为拦截

基于智能决策引擎的分析结果，系统会根据设定的规则进行实时响应及处置，以减少可能的损失。

**自动拦截。**当某个请求的风险评分超过预设阈值时，系统可自动阻止该请求进入业务流程，例如自动撤销伪造的交易请求或视频生成请求。

**人工审核。**对于中等风险行为，系统会将风险事件标记为待审核状态，生成通知并提交给人工审核人员进行深入分析。

**反馈与监控。**系统会将识别出的可疑行为反馈给业务流程，同时通过实时监控系统持续跟踪潜在欺诈的活动。

### 5.6.5 业务价值及优势

**高效实时识别。**融合反 AIGC 欺诈技术引擎的实时流处理系统，能够在欺诈

行为发生时实时捕获并处理多模态数据。通过实时流处理技术，系统能够实现毫秒级的响应时间，大幅提升对 AIGC 欺诈行为的识别能力。无论是伪造音频、视频，还是虚假文本，系统都能够在生成内容发布的瞬间进行风险评估，并及时采取拦截措施。

**多模态数据综合分析。系统**通过将设备数据、图像数据、音频数据和文本数据结合，系统能够进行跨模态分析， 揭示潜在的欺诈行为。例如，在一段视 频中，如果图像内容与音频特征不一致，系统能够根据图像和音频之间的关联 性判断该视频是否为 AI 伪造。系统综合分析各类数据源的特征，通过能够有效 识别 AI 生成的欺诈内容。

**智能决策与实时反馈。** 基于决策引擎的智能决策能力， 系统能够根据实时 数据自动调整响应策略，确保对 AIGC 欺诈行为的快速反应。通过结合多种算法 模型，系统可以不断优化风险评估和决策能力，逐步提升防御的精度。

**灵活应对新型欺诈手段。** 由于 AIGC 技术日新月异，欺诈手段不断变化，系统能够根据不断更新的规则库和算法模型，灵活适应新型的 AIGC 欺诈行为。例如，随着生成工具的更新，系统能够动态调整规则，确保始终处于最新的防护状态。

融合反 AIGC 欺诈计算引擎的高性能实时流处理系统，通过集成多种数据源及其特征，结合先进的跨模态分析、智能决策引擎和实时流处理技术，为识别和应对 AIGC 欺诈行为提供了强有力的支持。通过高效的实时识别、综合分析和智能决策，系统能够有效防范和应对 AI 生成内容带来的各种欺诈风险，确保在 复杂和隐蔽的网络环境中提供精确的防护，为业务流程的安全性提供保障。

# 第六章 典型业务场景

## 6.1 远程音视频反欺诈

### 6.1.1 背景

AIGC 技术的迅速发展，也为诈骗分子提供了可乘之机。利用 AIGC 的音视频生成能力，欺诈活动变得愈发复杂和隐蔽， 为金融机构的安全运营带来了新 的挑战。

某银行基于企业级音视频能力推出的在线业务办理平台， 致力于为客户提供便捷、全面的在线金融服务。黑灰产势力利用 AIGC 进行面部替换、表情驱动、全脸合成、背景生成等伪造技术，使银行远程视频服务面临的欺诈风险挑战，亟需构建全新的防范体系，保障客户资金财产安全。

### 6.1.2 风险分析

通过 AIGC 技术，黑灰产能够合成出完全虚拟的人脸图像，或者从网络上获取目标客户的面部图像，通过全脸合成技术生成几乎无法辨别的虚假面部，使得伪造的客户面部特征和表情与真实客户高度相似。不仅能够模拟客户的外貌，还能通过表情驱动技术，精准地再现客户的面部动作和情感变化，此外，伪造的背景环境也是重要的一环，攻击者可以生成与客户真实环境相符的背景，以消除人工座席对环境不匹配的怀疑。这种“背景生成”技术能够极大提高欺诈的隐蔽性，增加人工客服判定欺诈的难度。 黑灰产还通过 AIGC 语音模拟技术，能够精准复制客户的语音特征，包括音调、语速、语气和语音模式等。

传统的人工客服依赖客户提供的身份信息进行服务，但对于那些没有明显漏洞的 AIGC 伪造内容，人工客服很难从细节中分辨真伪。这使得黑灰产能够通过电话或视频通话冒充客户，进行账户查询、资金转账、贷款申请等高风险操作。

### 6.1.3 解决方案

为应对不断升级的远程欺诈风险，该银行以关键的伪造音视频检测技术为核心进行了全面的防御升级。

**积累伪造合成数据集。**针对伪造攻击的多样性和复杂性， 建立伪造合成数据集，涵盖 AIGC 生成的面部替换、嘴型驱动、声音合成等多种类型的虚假内容。通过数据集的积累和细分，构建精确的识别标准，有效区分正常与伪造内容。

**研发音视频伪造检测算法。**开发并优化基于深度学习的检测算法，采用技术手段分析音视频内容中不易察觉的伪造特征，例如视频帧间的细微错位、音频波形异常等，针对深度伪造内容的技术原理、统计特点进行识别，实现动态检测，提高伪造内容检测的精准度和效率。

**建立高效检测与风险控制机制。** 针对伪造内容高质量和快速变化的特点，及时生成欺诈风险评分，快速识别并标记高风险内容，形成预警信号。对于高于预设阈值的高风险内容，自动触发二次验证流程或转接人工审核，并通过对内部员工进行反欺诈培训，进一步控制风险。

**加大客户信息安全意识普及。** 通过反诈知识宣传和客户提醒，向客户传达欺诈风险信息，提醒客户在远程视频交互时保持警惕，不点击来历不明的链接，不轻易分享敏感信息。

### 6.1.4 实施效果

一方面，通过欺诈数据的积累，持续进行算法优化，不断提升检测算法的判伪精度，极大地降低了客户隐私泄露和财产受损的风险。进而基于高效检测算法和风控机制，以技防+人防的手段实现了对深度伪造内容的高效识别。伪造音频整体平均检出率 99%，伪造视频整体平均检出率 100%，大幅降低伪造音视频通过的概率，有效保障了业务流程的安全性。

与此同时，通过持续的内外部欺诈知识培训及宣导，提升内部员工及客户的反诈意识，更加主动和从容地应对欺诈风险，保障人民群众资金财产安全。

## 6.2 人脸识别身份认证反欺诈

### 6.2.1 背景

人脸识别和声纹识别等多模态生物识别技术为金融机构扩大服务半径、提升服务效率提供了必要手段，但随着人工智能技术的发展，AI 深度伪造技术（如深度学习驱动的面部合成）被犯罪分子应用在欺诈行为中，为金融服务的安全性带来挑战，对用户资金安全和银行声誉构成了严重威胁。

目前大多数金融机构所使用的人脸识别算法，主要基于深度学习算法，通过大量人脸数据进行特征提取、训练和识别。这一过程虽然高效，但也存在过度依赖数据的问题，透明度和可解释性不足，容易成为黑客攻击的目标。常见的攻击方式包括对抗样本攻击和深度伪造攻击，前者可通过在人脸图像上添加细微扰动来误导人脸识别系统，后者则能将一个人的面部特征转移到另一个人身上，制造逼真的动态视频，用以冒充他人完成身份验证。

### 6.2.2 风险分析

目前，对于人脸识别安全防御主要重心主要是通过摄像头或其他传感器直接捕捉到用户的面部图像或视频实时分析检测用户是否为“活体”的前端活体检测，以及通过对获取到的面部图像或视频数据进行更深层次的分析和比对，识别潜在的伪造内容的后端活体检测。

新型攻击方式针对上述防御手段进行针对性攻击，主要有以下几种：

针对活体检测增强模块的攻击。活体检测对屏幕重放攻击有一定的防御效

果，但若对 APP 或操作系统进行攻击，绕过摄像头采集，可直接将准备好的伪造图像传输给后台人脸比对算法，造成威胁。

针对人脸特征比对增强算法的攻击。人脸特征比对增强本质上是提升了比对算法的阈值，在提升自身安全性的同时，也降低了对于用户的友好体验，往往要进行反复拍照、核验才能通过比对，容易造成客户厌倦并采用其他手段进行校验，从而绕过人脸检测方案。

针对脸部异常结构识别的攻击。脸部异常结构识别旨在应对对抗眼镜样本等攻击方式，但对抗样本的攻击方式千变万化，目前已出现眼镜形式之外的形式，防御手段难以快速跟进。

针对眩光活体增强检测的攻击。眩光活体检测对常规黑灰产攻击有一定的防御效果，但对于当下新型的 AI 伪造+4K 屏翻拍模式基本无效，因其可在不对银行 APP 做任何更改的情况下，实现传统活体需要的用户动作+炫彩的无延迟反馈。

综上所述，传统安全加固措施，难以解决现有面临的新型攻击形式，需要从 AI 伪造本身对攻击的样本进行特征分析。

### 6.2.3 解决方案

针对上述挑战，某银行研发了 AIGC 伪造检测系统，该系统具有显著的技术价值和独特的特点，能够精准地识别并拦截 AIGC 伪造内容，有效提升现有生物特征识别系统的安全性，并专门针对对抗样本和深度伪造等新型攻击方式进行防御。

**识别深度伪造。**通过在现有活体检测和人脸识别的基础上加装 AIGC 伪造检测模块，系统能够为生物识别系统提供额外的防护层。系统通过深入到图像生成的细节层面，识别那些微小的伪造痕迹，深入分析图像特征、动作一致性、细节与纹理等多重维度，能够有效识别伪造图像或视频中的细微差异，例如不自然的面部表情、纹理的失真、光线变化异常等，从而提高了对深度伪造和对抗样本的防御能力。

**实时快速检测。** 通过对人脸图像的即时监控和分析， 系统可以在几秒钟内 完成伪造检测，并将结果即时反馈给业务系统。如果检测到图像为伪造或存在 可疑迹象，系统会自动阻止该图像进入后续识别流程，确保伪造内容无法突破 生物识别层级并影响后续的业务操作。

**AI 与人工结合判断。** 系统会在检测到可疑图像时，标记为“待人工复审”状态，并将其交由经过专业培训的人工审核员进行进一步验证。这种人工智能与人工审核相结合的方式，能够提高检测的准确性，避免系统误判或漏判，确保客户身份验证的可靠性。

**系统自我更新。** 该系统通过机器学习算法，能够在接收到新的欺诈样本时， 实时更新检测模型。随着新型攻击手段的出现，系统可以基于历史数据的反馈 机制，自主调整算法，及时识别新类型的伪造图像。这种持续学习的能力，使 得该系统在面对未来不断变化的欺诈方式时，始终能够保持有效性和精准度。

这套 AIGC 伪造检测系统通过针对深度伪造技术的专门防护，提升了银行对于 AIGC 相关欺诈行为的识别能力 ，有效避免由于身份盗用和虚假认证引发的金

融风险。系统的高效检测与自动化响应机制能够在第一时间识别出欺诈行为，及时止损，从而保障了金融交易和客户资金的安全。

### 6.2.4 实施效果

该系统已成功上线并应用于 20 多个在线业务场景，涵盖账户注册、登录、资金转移、设备更换和信用额度调整等关键环节，累计处理人脸识别请求约 3.8 亿次，有效抵御针对人脸识别的身份攻击事件 1.8 万起，保护了超过 2000 名潜在受害者的财产安全，精准拦截数亿元经济损失，显著增强了银行的风险控制能力，为维护客户信任和市场稳定做出了重要贡献。

## 6.3 伪造人脸考勤反欺诈

### 6.3.1 背景

保险公司的人力成本中，薪资占比达 85%以上，福利成本占比高于 12%。截至 2022 年 6 月 30 日，全国保险公司在保险中介监管信息系统执业登记的销售人员 570.7 万人。

人员管理已成为保险公司的核心竞争力。然而，虚假考勤、代替打卡等现象屡见不鲜，造成人力资源浪费。2023—2024 年，黑灰产通过破解多家公司考勤系统，制作出打卡作弊工具，并向保险公司员工兜售“代打卡服务”，可以让保险公司员工不出门不到岗也可以实现“上班打卡”，轻松领取全勤奖。

### 6.3.2 风险分析

黑灰产主要通过伪造人脸视频和考勤 App 等手段，提供考勤欺诈服务。

**伪造人脸视频代打卡。**个人提供真实的人脸视频，黑灰产将视频上传至考勤系统，绕过考勤系统的人脸识别，帮助购买者完成每日的考勤打卡。最新的攻击已经可以只通过照片制作出具有高真实性的人脸视频，模拟目标员工的面部表情和动作，使考勤系统将错误地认为是员工本人，从而完成考勤打卡。

**伪造考勤 App 。**黑灰产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 并伪造新的考勤 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。个人可在任何地点使用该 App，该 App 会模拟出公司场景、位置等信息并发送至后台，后台系统将错误地识别为正常的考勤打卡。

### 6.3.3 解决方案

某保险公司将人脸考勤系统同设备指纹、行为识别等技术融合，并通过强化考勤 App 的安全性以提高防范能力。

**检测识别考勤打卡机上传的 AIGC 伪造的视频。**通过对考勤打卡机设备环境、人脸信息、图像鉴伪、用户点击动作等多维度信息进行智能核验，结合考勤打卡机所在环境中声音与视频帧，检测出音画不同步、考勤背景声不一致等问题，提高整体伪造检测的鲁棒性。

**检测识别终端 App 上传的 AIGC 伪造的视频。**通过对终端 App 上传的音视频进行鉴伪分析，综合判断办公背景、光照、背景音、地理位置等情况，进行音视频鉴伪分析。

**提升考勤 App 安全性。**一是通过定期进行渗透测试，对 App 进行运行环境安全检测，观察是否有存在代码注入等行为，同时增加 App 安装包 (SDK) 的合法性检测，检测包的签名、大小、进程信息、 App 版本号等，并持续进行不同版本 App 的信息检验；二是通过获取终端设备指纹、 IP 地址等信息，识别同一设备频繁登录多个账户等情况，防范虚拟终端风险。

### 6.3.4 实施效果

分析发现，某些地区考勤作弊的保险员工数量占比高达 25%以上，该方案 在保险公司部署后，拦截阻止虚假考 15 万次，节省人力成本近百万元。

## 6.4 虚假视频聊天反欺诈

### 6.4.1 背景

AIGC 为威胁行为者提供了新工具， 黑灰产正在使用 AI 来瞄准员工、创建网络钓鱼电子邮件、冒充供应链合作伙伴。部分案例表明，诈骗分子在视频会议中创建深度伪造的企业负责人形象， 要求受害人转账或提供账号密码，导致 个人或企业的资金被盗或涉及重大合同或交易信息泄露，影响客户利益、商业 利益。

此外，诈骗分子通过 AIGC 生成更加逼真的电子邮件内容，骗取受害者的信任，要求受害者点击恶意链接或下载附件，从而泄露敏感信息或者控制受害者的电脑，进而通过视频软件、聊天工具等多种手段，实施进一步的诈骗活动。

根据德勤的最新报告[18]，与深度伪造相关的网络攻击损失预计将从 2023 年的 123 亿美元飙升至 2027 年的 400 亿美元，复合年增长率达 32%，银行和金 融服务业将成为主要目标。

### 6.4.2 风险分析

犯罪分子主要通过音视频伪造技术伪造成被骗人员的关系人，进而通过音视频、交友 App 等手段逐渐诱导实现诈骗。下面以一名企业财务人员遭遇的电信网络诈骗为例进行说明。

**关系人信息收集与伪造。** 黑灰产首先了解该财务人员所在公司， 并收集该 公司领导的公开照片、视频素材，甚至通过社交网络或社交工程手段获取私人

化信息。

**AIGC 技术合成。**利用 AIGC 技术，基于收集的相关素材，诈骗分子合成该领导的声音，并将自己的面部表情、语言和动作准确地转换到该公司领导的脸上，创造一个虚拟且可信的领导形象。

**建立联系与信任。**诈骗分子通过交友 App 与该财务人员建立联系，基于仿真的音视频合成技术，关心员工工作、生活状态，允诺在工作中的相关待遇，逐步建立信任，使受害人放松警惕。

**紧急资金需求诱导。**案发当时，诈骗分子同该财务人员紧急视频，伪装成 领导并通过语音和视觉传达紧急资金需求，要求该财务人员迅速转账以应对公 司资金周转问题。

### 6.4.3 解决方案

通过构建一个多维度、全方位的欺诈识别系统，结合音频、图像和视频等不同模态的信息，利用深度学习和专家知识，实现欺诈检测。

**检测识别 AIGC 伪造的视频。**在视频伪造检测中，结合声音的语义分析技术与视频图像分析技术，对音频的语调、音频频谱异常、视频中的光影不一致等特征进行综合分析，检测音画不同步、情绪不一致等问题，提高伪造检测的精度。

**设备维度关联分析。**通过对伪造音视频设备进行记录，建立欺诈视频终端 数据库，并建立名单动态运营维护机制，沉淀并维护相应的黑白名单数据。防 范团伙作案。基于设备指纹、IP 地址等信息，同溯源数据库进行比对，识别出 是否存在欺诈行为。

**基于记录数据持续优化模型。**基于沉淀及积累的欺诈案例及数据，利用智 能模型平台构建专属欺诈风险模型，并根据防范情况及时更新风控策略，实现 风控策略实时迭代更新，提升防范虚假人脸风险的水平。

### 6.4.4 实施效果

某银行部署该系统后已经成功拦截数十起虚假视频的诈骗案件，及时识别和阻止虚假视频的侵入，避免了近百万元的资金损失。

# 第七章 展望与倡议

AIGC 技术以其优异的表现、普适的场景而受到社会的广泛认可，其持续发展是数字经济时代的必然。然而，技术作为双刃剑，其快速发展也会带来更复杂的欺诈风险。随着反欺诈技术的持续演变，AIGC 音视频欺诈将呈现智能化、多维化、自动化、个性化等特点，需要持续关注、及早布局、从容应对。

## 7.1 未来技术挑战

**超逼真生成技术的发展提升了鉴伪难度** 。一是随着算法的优化、算力的扩 容、音视频数据的不断采集，更加逼真、更加精细的音视频生成技术将出现，
高分辨率人脸生成技术可细致到毛孔、瞳孔纹理及微表情，并结合动态模糊处理增强真实性，达到近乎真实的效果，甚至在人眼和传统算法检测下无法辨别；二是实时换脸与拟声技术将得到极大改进，可以在视频通话中即时伪装为他人，实时互动而无卡顿或破绽；三是声音和影像的配合将更加紧密，情感语调与面部表情同步，整体的欺骗性更强。

**个性化 AIGC 伪造内容生成导致攻击更加精准** 。AIGC 技术结合用户数据，可生成高度定制化的 AIGC 伪造内容，对目标对象实施精准攻击。基于社交媒体上公开的音视频内容，AIGC 可以学习目标的外貌、声音、行为习惯，模拟复杂的生活、工作场景，利用语气模仿、信任场景构造等手段诱导目标作出决策。

**生成与篡改技术融合形成组合型风险**。音视频生成技术结合篡改技术可大幅提升欺骗性和攻击力，通过拼接真实音视频片段并补充伪造内容，可实现更高可信度的 AIGC 伪造材料。

**自动化与规模化攻击使“技防”成为必须路径** 。AIGC 模型的训练和使用门槛不断降低，攻击者可编写脚本，批量调用成熟工具生成伪造音视频内容，形成规模化攻击。犯罪团体可能开发基于 AIGC 的欺诈服务平台，向不具备技术能力的犯罪分子提供自动化工具，生成音视频逐步渗透目标系统，在多个环节实施欺骗行为。

**自学习、自适应等技术的提升导致攻防对抗更激烈**。学习是不断提供标签反馈再进化的过程， 随着反欺诈样本的增多，AIGC 也将不断优化生成能力，以逃避检测。例如，利用生成对抗网络（GAN）提升伪造内容的检测逃逸能力，甚至针对特定检测算法进行优化；利用自适应学习进行试探性攻击，学习反欺诈技术的特征，动态调整生成策略，持续提高欺骗成功率；采用多模态技术及对抗思想，同时生成文字、语音、图像、视频等多模态伪造内容，形成复合型欺诈，增加欺诈内容的可信度等。

## 7.2 相关倡议

技术发展带来的挑战不可避免，需要同时兼顾发展和安全。需要 AIGC 技术提供者、AIGC 使用者、行业机构等产学研各方的通力协作，构建安全、可信的 AIGC 音视频生态体系，才能在技术创新与安全防护之间找到平衡。需要在法律、

技术、生态等层面同步发力， 推动全球化治理，形成多层次、全方位的防护网，

才能更有效防范 AIGC 音视频风险带来的威胁。

### 7.2.1 健全合规体系

#### 一、 法律法规

**强化法律法规的向善引导作用。**压实 AIGC 技术及使用各方的主体责任，促进技术的合规、合理使用，避免误用、滥用， 严厉禁止将先进技术应用于违背 社会伦理道德、违反社会价值、违反法律要求的场景中。

**与时俱进更新法律法规。**随着 AIGC 技术的创新发展，与时俱进更新法律法规，适应人民群众的新生活方式、有效抑制新风险、从容应对新挑战，既需避免过度抑制技术创新，又应在保障社会安全与道德底线的前提下推动行业健康发展。

**进一步强化数据隐私与生物特征保护。**在第六章应用场景的风险分析中可 以看到，犯罪分子如能获取公民的生物特征、隐私信息，其攻击能力将极大增 强。宜强化对源头数据的保护与治理， 细化在对隐私信息采集、处理、使用等 过程中应明确的使用范围、安全保护要求，避免信息泄露。

#### 二、标准建设

**明确 AIGC 生成的音视频内容附加标识的规范。**让使用方可以清晰辨别原始影像与合成数据，实现充分告知、及时验证、不影响用户体验；在此基础上，进一步推动跨国共识，防止跨境规避的情况发生。

**细化 AIGC 应用透明度和可解释性的标准。**形成技术规范或应用指引，并进一步建立相关内容评测标准，通过第三方机构测试等方式提升技术应用的可信度。

**建设 AIGC 鉴伪能力指引。**明确对 AIGC 生成的音视频进行鉴伪的必要性，并指导企业在技术应用、业务流程、管理机制等方面进行音视频伪造风险防范。

### 7.2.2 创新发展技术

#### 一、研制性能更优的欺诈检测模型

随着 AIGC 技术的进步，AIGC 与反欺诈技术之间将进行持续的竞赛。AI 生成的虚假内容将不断优化其隐蔽性和仿真度，而反欺诈系统则需要持续更新算法及机制来应对愈发高级的攻击。

**持续优化检测模型性能。**综合利用积累的数据及业务经验，利用小样本学 习、持续学习等方式，实现模型的快速迭代，适应新型欺诈手段。

**融合多模态分析技术进行综合分析。**通过对图像、音频、文本等多种数据 的分析，实现多模态交叉验证，进行更全面的威胁检测。

**针对典型的 AIGC 音视频生成算法进行逆向检测**。针对 AIGC 算法的固定模式、技术特征，同真实影像的统计学差别等特点，进行针对性研究；也可通过生成对抗网络（GAN）识别模型生成的伪造内容，进一步提升伪造检测的精度。

### 二、构建自动化反制体系

**强化技术创新，研制持续进化的反 AIGC 技术。**开发具有自适应能力的反欺诈 AI 模型，通过深度学习和对抗训练实时学习新的欺诈模式。

**建立防范机器人攻击的防御体系。**针对 Bot 攻击的特点，利用历史流量分 析、设备指纹等技术，检测机器人攻击行为并进行溯源分析，从而进一步屏蔽 该团伙大量生成的伪造内容。

### 三 、提供定制化的身份验证及风险预警服务

**研制多维度验证手段。**在客户授权的前提下， 基于客户授权的相关信息及 存量信息，结合行为分析、动态生物识别技术，以及多因素验证等手段，对音 视频进行综合化分析，可有效提升检测精度，提升服务安全性。

**提供个性化风险预警模型。**针对不同用户群体的特点建立个性化风险模型， 并生成客户定制化的风险预警内容，在保障客户体验的前提下实现风险充分预 警，使用户更容易理解潜在威胁。

## 7.2.3 构建健康生态

### 一、产学研用多方协同，构建健康应用生态

**提升先进技术应用转化效率。**强化技术应用方同科研机构等技术提供方的 合作，提升技术转化效率，加快先进技术普及进程，以最快的速度防御日新月 异的挑战。

**构建协同防御机制。**新型欺诈往往跨机构作案，单一机构难以看到完整的 资金链路及欺诈过程，金融机构、电信运营商、视频通讯服务提供商、聊天工 具服务提供商等宜建立联合防御机制，应对复杂、隐蔽的作案手法。

**提供便捷的虚假视频检测手段。**宜提供开放便捷、轻量的反伪造工具，通 过数据完整性、元数据特征和数字水印等方法提升伪造内容的识别能力，供社 会公众使用，使社会公众及时获知音视频伪造情况，降低操作使用的风险。

### 二、促进行业共享，建立反诈案例数据库

**建立威胁情报共享合作机制。**在合规的前提下，通过名单共享、通报交流、 案例研讨等方式及时共享黑名单数据、新型攻击手段和典型欺诈案例。

**在情报共享基础上形成典型案例数据库。**收录最新伪造技术及特征，供企 业和机构进行运营策略、算法调整和风险评估。

### 三、强化宣传引导，提升反诈意识

**强化内部员工培训。**及时学习掌握新技术， 了解新技术可能带来的攻击手 段，在业务办理时强化风险意识，在运营策略、风险防范、操作流程等方面考 虑新技术手段可能带来的业务风险。

**面向公众进一步普及最新的反诈知识，提升反诈意识。**防范换脸诈骗、拟声电话、远程视频伪造等形式的诈骗；强化公众隐私保护意识，谨慎上传人脸、声音等个人数据，避免授权给不可信的平台或应用；提升公众在资金交易或敏感信息操作过程中的风险意识，面对可疑或不信任场景时，及时求证或报告可疑情况，迅速向相关平台或机构举报。

**普及 AIGC 知识，减少不必要的恐慌。**先进技术虽然可能被犯罪分子利用，但更多的是用于促进社会发展、提升人民群众幸福感。强化 AIGC 技术知识的普及，可以有效增强企业、消费者、公众对新技术的信任，为行业信任和秩序奠定基础，真正实现科技为民、科技向善。

# 后记

《金融 AIGC 音视频反欺诈白皮书》旨在引起全社会对 AIGC 技术带来的日益复杂的安全威胁的关注，金融企业、行业从业者以及社会各界应保持高度警惕，形成合力，共同应对 AIGC 滥用带来的安全风险。

我们希望通过本白皮书，提升金融行业对 AIGC 风险的认知，强化防范措施，并推动反欺诈技术的创新。推动政府、行业和企业共同携手，形成一个全方位、多层次的防控体系，从而减少 AIGC 技术滥用带来的潜在风险，确保金融体系的安全和可靠，保障用户资金安全和业务稳定的同时，应对 AIGC 带来的挑战。

在此，我们特别感谢北京市科委、中关村管委会 、上海金融科技产业联盟的大力支持，本白皮书也是“远程银行音视频系统反 AIGC 欺诈和智能决策系统关键技术及应用研究”项目研究成果之一。

我们诚挚邀请更多的合作伙伴加入 AIGC 音视频反欺诈的行列，与我们一道，共同推动金融行业在应对 AIGC 带来的安全威胁方面不断创新，确保金融安全与用户信任的长期可持续发展。

# 参考文献

[1]央视网，《热解读｜从这八个字理解人工智能治理中国方案》

[2]新华社，《中共中央关于进一步全面深化改革 推进中国式现代化的决定》

[3] GB/T 38671—2020《信息安全技术 远程人脸识别系统技术要求》

[4] T/IIFAA3001.1—2021《远程人脸识别应用技术规范第 1 部分：金融账户管理》

[5]JR/T0164—2018《移动金融基于声纹识别的安全应用技术规范》

[6]Yang X, Li Y, Lyu S. Exposing deep fakes using inconsistent head poses [C]//ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 8261—8265.

[7]Guo H, Hu S, Wang X, et al. Eyes tell all: Irregular pupil shapes reveal gan—generated faces [J]. arXiv preprint arXiv:2109.00162, 2021.

[8]Hu S, Li Y, Lyu S. Exposing gan—generated faces using inconsistent corneal specular highlights [C]// ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 2500—2504.

[9]Korshunov P, Marcel S. Speaker inconsistency detection in tampered video [C]//2018 26th European Signal Processing Conference (EUSIPCO). IEEE, 2018b: 2375—2379.

[10]Mittal T, Bhattacharya U, Chandra R, et al. Emotions don't lie: A deepfake detection method using audio—visual affective cues [J]. arXiv preprint arXiv:2003.06711, 2020.

[11]Jia S, Li X, Lyu S. Model Attribution of Face—swap Deepfake Videos[J]. arXiv preprint arXiv:2202.12951, 2022.

[12] Ciftci U A, Demir I. How Do Deepfakes Move? Motion Magnification for Deepfake Source Detection[J]. arXiv preprint arXiv:2212.14033, 2022.

[13]Girish S, Suri S, Rambhatla S S, et al. Towards discovery and attribution of open—world gan generated images[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 14094— 14103.

[14] Narayan K, Agarwal H, Thakral K, et al. DeePhy: On Deepfake Phylogeny[J]. arXiv preprint arXiv:2209.09111, 2022.

[15]Yang T, Huang Z, Cao J, et al. Deepfake Network Architecture Attribution[J]. arXiv preprint arXiv:2202.13843, 2022.

[16] Yu N, Davis L S, Fritz M. Attributing fake images to gans: Learning and analyzing gan fingerprints[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 7556—7566.

[17] Guarnera L, Giudice O, Nießner M, et al. On the Exploitation of Deepfake Model Recognition[C]//Proceedings of the IEEE/CVF Conference

on Computer Vision and Pattern Recognition. 2022: 61-70.
[18] Deloitte , 《Generative AI is expected to magnify the risk of deepfakes and other fraud in banking》