

可视身份深度伪造诈骗犯罪的机理分析及防御策略研究

■ 赖志茂 冯 聪 廖广军

摘 要 深度伪造技术的恶意应用严重威胁社会稳定和政治安全，还衍生出新型的诈骗犯罪。照此态势，电信网络诈骗手法已然进化。不法分子极易通过现有公开技术和素材实施诈骗，而公安机关缺乏有效的应对措施和反制手段。本文从机理分析和防御策略两方面入手，系统梳理了可视身份深度伪造的典型方法，深入探讨可视身份深度伪造机理和技术手段，并从被动防御和主动防御两个方向提出了有效的防御策略，为公安机关打击新型诈骗犯罪提供理论指导和决策支撑。

关键词 深度伪造 新型犯罪 态势分析 防御策略

随着人工智能技术的快速发展，可视身份深度伪造诈骗犯罪（下文简称“深伪犯罪”）已经成为一种新型犯罪形态，给社会稳定和公共安全带来了巨大威胁。深伪犯罪是指利用高科技手段如深度学习、计算机视觉等技术生成高度逼真的虚假身份信息，用于欺骗、伪造证件、冒充他人、非法获利等犯罪行为的一种犯罪形式，主要涉及伪造身份证件、伪造人脸生物特征、制造虚假场景和虚假身份等。其中，深度伪造是深

伪犯罪中的核心技术，最初源于一个名为“DeepFake”的 Reddit 社交网站用户，于 2017 年 12 月在社交网站上发布了将盖尔·加朵等女明星的面孔映射至色情表演者身上的“换脸”视频，此后“DeepFake”和“换脸”也被引用成为了该技术的代名词。它属于一种基于深度学习的图像合成技术，通过伪造合成人脸身份信息或部分属性信息，以使得肉眼无法分辨输出图像或视频中的人脸是否经过篡改。

作者：赖志茂，中国人民警察大学移民管理学院（广州）讲师；冯 聪，广州市公安局网络警察支队警务技术四级主任；廖广军，中国警察协会学术委员，广东警官学院刑事技术系教授
基金项目：本文系中国人民警察大学国家基金培育课题“面向电子物证鉴定的深度伪造人像视频取证技术研究”（项目编号：JJPY202402）；国家重点研发计划项目（项目编号：2023YFC3321600）阶段性研究成果。

早期人们应用深度伪造技术仅仅是出于娱乐的目的，尤其在影视制作中得以运用，在电影制作中创建虚拟角色、视频渲染、声音模拟等，“复活”历史人物或已逝的亲朋好友，实现“面对面”沟通，创造了一种新型的交流方式。但若一些人出于不可告人的目的传播一些经过精心伪造篡改的“换脸”图像或视频，并应用到新闻报道、安全监控、刑事侦查、法律取证等方面，将会给社会带来非常恶劣的影响。不法分子可能利用深度伪造技术对政治人物进行恶意抹黑，甚至危害国家安全。2018年4月，美国网媒媒体 Buzzfeed 发布一段短视频，利用人工智能技术来模仿前总统奥巴马，在视频中批评时任美国总统特朗普，该伪造视频对美国政府形象造成了极大的不良影响。2020年4月，俄罗斯 RT 网站发布了一段使用深度伪造技术制作的讽刺视频，视频显示“美国总统特朗普”在11月大选后去俄罗斯 RT 电视台工作了。这段视频被美媒《每日野兽》网站和“林肯项目”组织的“通俄门”宣扬者利用，借此干扰美国总统大选。2022年3月，俄乌冲突开始不久，网上流传一段使用深度伪造技术，假冒乌克兰总统泽连斯基的投降片段，泽连斯基后来亲身拍片澄清辟谣。2023年6月，克里姆林宫表示俄罗斯境内广播电视网络部分线路遭到黑客攻击，播放伪造俄罗斯总统普京的紧急讲话，并散播乌军已入侵俄罗斯的不实信息，该段视频经过了深度伪造生成，图像与声音跟普京十分相似，目的是给俄罗斯境内居民带来恐慌。

目前，境内外已发生利用深度伪造技术实施的新型电信诈骗案件，2023年5月，国内包头警方发布一起利用人工智能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10分钟内被骗430万元，骗子主要通过智能AI换脸和拟声技术，佯装好友对其实施了诈骗。该事件首次让这种利用深度伪造的新骗局走入了公共的视野，引起社会多方重视。除此之外，包括公安机关宣传民警在内各行各业人物的公开视频，都有可能被不法分子采集面容声纹后制作成深伪视频，进而实施诈骗，欠缺专业知识和鉴伪工具的市民群众，极易被冒充公安民警的深伪视频行骗，也极易产生负面舆情。为此，公安机关和相关领域的研究人员一直致力于对深度伪造进行可靠的检测与标记，以阻止恶意信息的传播和诈骗犯罪的发生。然而，由于深度伪造生成技术的快速演化发展以及防御策略的有效性不足，这仍然是一个有待解决的问题。

一、可视身份深度伪造典型方法

（一）深度伪造典型方法的生成原理

现有的深度伪造技术大致可分为以下四种类型：人脸生成、人脸交换、局部属性编辑和人脸重现。人脸生成是使用生成对抗网络生成现实中不存在的人脸，可以生成面部五官区域、头发、姿态等。人脸交换是将图像或视频中一个人的面部换成另一个人的面部。局部属性编辑主要是修改人脸的某些局部属性，如头发颜色、皮肤颜色、性别特征、年龄特征、是否佩戴眼镜等。人脸重现主要是修改面部表情，即把给定原始人脸表情进行复制，在目标人脸上进行表情重现。其中，人脸交换和人脸重现是目前深度伪造方向上最流行的方法，也是带来风险隐患较高的技术手段。

1. 人脸交换

人脸交换对两个人进行身份交换，利用

原始人物图像取代目标人物图像的整个头部信息，保持其他属性信息不变，可以实现让原始人物出现在特定场景中。基于深度生成模型的人脸交换主要流程如图 1 所示，图中原始人物视频和目标人物视频来自公开数据集 Faceforensics++。在模型训练阶段：首先，收集原始人物 A 和目标人物 B 的图像或视频信息，提取人脸并进行裁剪和对齐等预处理操作；随后，训练权值共享的编码器 A 和编码器 B，用于提取人脸 A 和人脸

B 的共有面部属性；最后，各自训练一个独立的解码器 A 和解码器 B 分别学习特有的面部信息，完成对应人脸重构。在身份交换阶段：首先，利用编码器 B 对人脸 B 进行面部属性编码；然后通过解码器 A 重构人脸 B 的图像，生成具有人物 A 外貌特征但保留人物 B 面部表情的伪造图像；最后，完成后期图像融合与视频合成操作。

2. 人脸重现

人脸重现传递原始人脸动作表情与姿

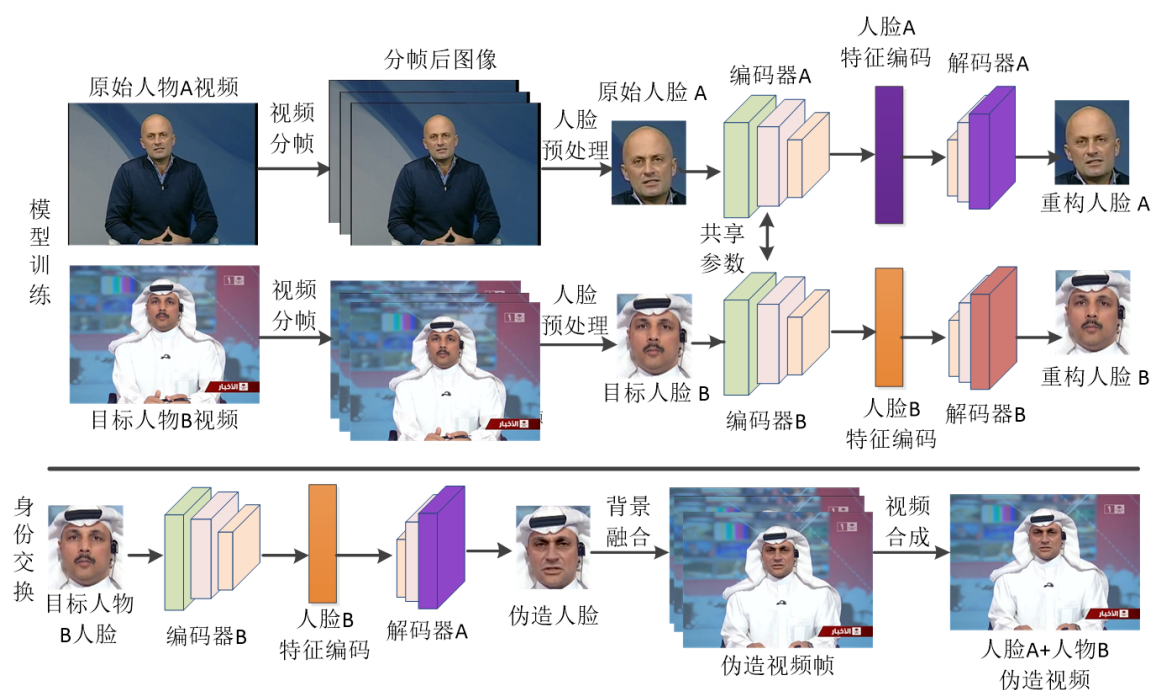


图 1 基于深度生成模型的人脸交换基本原理图

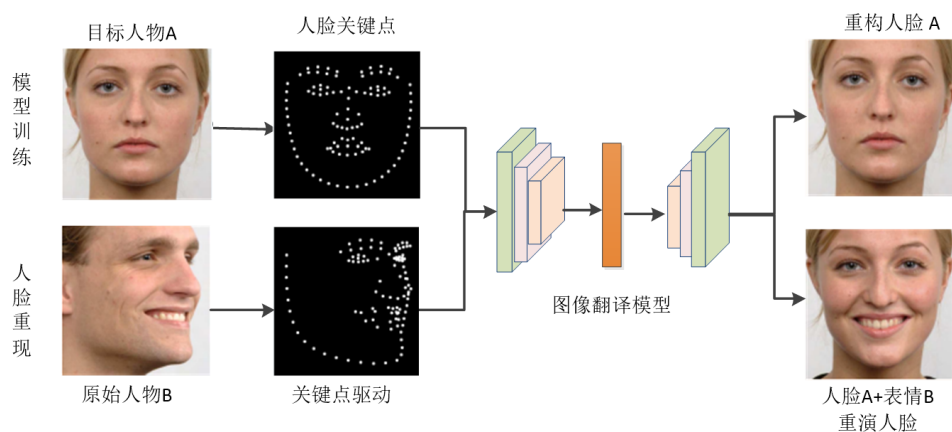


图 2 基于深度生成模型的人脸重现基本原理图

态，不改变目标人物的身份信息，使修改后的人脸保留目标人脸的长相和原始人脸的表情与姿态，再配合指定的语音内容，可以实现让目标人物做出指定的表情动作，发表目标人物非本人真实意图的言论和行为。基于深度生成模型的人脸重现主要流程如图 2 所示。首先，提取目标人物 A 的人脸关键点，并以此作为图像翻译模型的输入，对网络进行训练学习重构目标人物 A 的表情动作；其次，提取待传递的原始人物 B 的人脸关键点，作为表情驱动输入图像翻译模型，从而重构出具有人脸 A 和表情 B 的重演人脸，达到表情迁移的目的。最后，根据需求完成后期图像融合与视频合成等操作。

(二) 深度伪造制作工具

目前，深度伪造开源库中最具代表性的两种核心技术分别为 Deep-FaceSwap 和 FaceSwap-GAN，基于这两种技术开发的深度伪造视频常用制作工具如表 1 所示。下面将对这几种制作工具进行简要描述。

1. FaceSwap 是一款功能全面而强大的深度伪造应用程序和软件，可用于以最小的投入创造出惊人的效果。作为一种在线工具能在几秒钟内交换面孔，一张图片加上一个视频，就能把视频里面的人物进行深度伪造。功能包括在任何视频或图像中轻松交换

人脸、动画和卡通人物，通过移除和替换背景快速重现图像场景，并提供快速编辑和配音功能。

2. Deepfake Web 是一项可让用户在网络上创建深度伪造视频的服务，可能是最好的在线人工视频制作工具。它采用机器学习算法，将视频中的人物与另一个人进行脸部互换，用户界面友好，适合没有丰富技术知识的用户使用。在 Deepfake Web 的帮助下，用户无需具备高超的技术就能创建自己的深度伪造作品，快速简单地将深度伪造技术应用到自己的照片和视频。它利用深度学习来吸收各种复杂的人脸数据，可从任何有网络连接的设备上方便地访问。

3. Reface 是一个强大的基于人工智能的平台，可以让用户用自己选择的视频、GIF 和图片创建逼真的深度伪造效果。Reface AI 是生成式对抗网络的幕后功臣，要使用这款应用程序，只需捕捉用户照片，然后选择用户想使用的 GIF 即可。Reface 还提供各种滤镜和特效，帮助创建独一无二、令人惊叹的深度假脸，也可用于快速制作各种假冒名人的视频。在社交媒体上分享这些视频也同样方便，只需一个按钮即可完成，不少社交名人都曾使用该应用制作过精彩的合成视频。

表 1 深度伪造视频常用制作工具总览

工具名称	技术特点
FaceSwap	以最小的投入创造出惊人的效果
Deepfake Web	在网络上创建深度伪造视频的服务
Reface	能快速、准确地创建高质量深度伪造视频，并提供一键分享
DeepFaceLab	集成大量开源模型，可自由设置参数
FakeApp	提供基于 Windows 系统的图形化界面
Zao	移动端应用，提供视频模板，上传少量视频在短时间内生成伪造视频

4. DeepFaceLab 是一款 Windows 应用程序，允许用户创建深度伪造的视频，主要面向计算机视觉研究人员和学生，使用时需要一台配备专用高端 GPU 的强大电脑。它利用机器学习和人脸图像合成来替换视频中的人脸，这个开源的深度伪造系统是深度伪造软件领域的佼佼者。由于 DeepFaceLab 是一款主要面向研究人员的高级工具，其界面并不友好，学习时间较长，用户必须从文档中学习其用法。该软件具有灵活的设置、精确的面部映射和尖端的算法，可实现高度逼真的效果，深受希望获得高质量效果的伪造者青睐。

5. FakeApp 的 AI 照片深度伪造技术可让用户用人工智能改变自己的照片，提供各种特效、背景和滤镜，可以用它们来改变自己的外貌。这款应用的编辑功能非常广泛，可以调整面部表情、玩转时空旅行，甚至在性别之间进行切换，让用户彻底改变自己的照片。它快速、简单、高效，只需点击几下，用户就可以应用多种滤镜和特效，创建一个逼真的深度伪造版自己，还可以修饰任何斑点或皱纹，使用胡须特效来尝试不同的面部发型。

6. 移动深度伪造应用程序 ZAO 彻底改变了深度伪造技术，不仅能让两个人交换面孔，还能在任何视频中添加自己的面孔。通过 ZAO 先进的深度学习算法，人们可以像自己喜欢的表演者和虚构人物一样出现，用户友好的界面和令人印象深刻的效果使其在深度伪造爱好者中大受欢迎。用户通过 ZAO，只需一张自拍，即可将本人脸部无痕换到明星头上，生成同款视频或者表情包。

二、深度伪造诈骗犯罪手段及态势分析

（一）深度伪造诈骗犯罪的主要类型

近年来，公安机关聚焦人民群众深恶痛绝的电信网络诈骗，全面加强“四专两合力”建设，深入推进打防管控各项工作，有效遏制了案件快速上升势头，有力维护了人民群众合法权益。但是，电信网络诈骗犯罪形势依然严峻，尤其随着技术的发展，出现了以深度伪造类新型诈骗犯罪，其中冒充公检法类诈骗和冒充领导、熟人类诈骗已经成为最为突出的十大类电诈犯罪。靳高风团队在中国犯罪形势分析总结与预测中提到：“网络犯罪已经成为主要的犯罪类型，寄生网络黑灰产的信息网络犯罪大幅增加，应当重点防范网络新技术变革引发的新型犯罪。”

1. 冒充公检法类诈骗

诈骗分子首先通过非法渠道精准获得受害人的个人信息，随后冒充公检法机关工作人员与受害人取得联系，再以受害人涉嫌违法犯罪为理由进行威胁、恐吓，要求其配合调查并严格保密。为了增加可信程度，诈骗分子利用深度伪造技术，生成深伪视频以执法人员的名义角色实施诈骗。2023 年以来，利用苹果手机 FaceTime 视频通话功能实施的电信网络诈骗案件频发。由于 FaceTime 有着易拨通、可匿名、难溯源等特点，便衍生出多种诈骗作案手法。其中，利用 FaceTime 致电实施的“猜猜我是谁”“冒充公检法”就是诈骗分子的常用伎俩。

此外，还新增了假扮绑架的诈骗。诈骗分子通过话术诱骗年轻人，令其以为自己身涉罪案危险，必须远赴境外失联藏匿，然后再联络年轻人的家人，声称已绑架此人并索要赎金。此类绑架诈骗案实质是冒充公检法

手法的变种，同样是境外远程用话术遥控受害人实施的。其中，轻信“被绑架”的受害人就是关键环节。在近期的某宗类案里，诈骗分子就利用了深伪技术对受害人进行话术控制。诈骗分子在与受害人视频通话时，播放了一段某地反诈民警公布在网上的公开视频，再加以话术配音，就令受害人相信自己正在跟公安民警进行通话，进而确信自己真的涉及罪案而要失联藏匿。

2. 冒充领导、熟人类诈骗

诈骗分子首先通过非法渠道获取受害人的个人信息及领导同事关系、亲友关系等，并根据收集的信息，选择一个合适的诈骗角色，如冒充受害人的上级领导、亲戚、朋友或同学等，以便更容易获得受害人的信任。随后，诈骗分子通过电话、短信、社交软件等方式与受害人建立联系，并尽可能模仿目标人物的言行举止、使用相似的头像和昵称，利用深度伪造技术合成模仿目标人物的言行举止。最后，就是设定好骗局剧本，进而引导转账汇款。诈骗分子通常会编造紧急情况或诱人情境，例如领导因公急需周转资金、亲友遭遇突发事件急需用钱，或者推荐一个看似有利可图的投资机会等，让受害人产生同情心、责任感或贪婪心理。在受害人对虚构情境信以为真的情况下，诈骗分子会催促受害人尽快转账汇款，并强调保密性，不让受害人有足够的时间去核实信息或咨询他人意见。

2024年2月，香港警方公布案例，某英国跨国公司的香港分公司员工，在假冒的“高管视频会议”中，被深度伪造的“英方高管”骗取2亿港币。据称，这位香港分公司的员工先收到了自称英国总部CFO的邮件，称总部正在计划“秘密交易”，需要分公司将资金转到几个香港本地的账户中

待用。而这个员工一开始并没有相信邮件内容，认为是钓鱼诈骗邮件。但是骗子不断给他发邮件讨论这个秘密交易的必要性，并向该员工拨通了视频通话。在通话中，这位员工看到了他见过的公司CFO和其他几位“同事”，会议中的“英方高管”要求他立即转账。该员工信以为真，就分15次向5个香港本地账户陆续汇款了2亿港币。

（二）深度伪造诈骗犯罪的关键环节及态势分析

随着社交网络的发展应用，个人照片或视频分享成为了公众的爱好，这也给犯罪分子带来了数据信息收集的渠道便利。此外，部分网民利用换脸软件制作大量虚假视频哗众取宠、提升关注热度；还有网民热衷寻求新鲜刺激，随意将自己和他人的人脸信息上传到换脸软件，通过制作心仪的虚幻场景来满足虚荣心，进一步扩大了公民隐私安全隐患。

深度伪造类诈骗案作案手法中的关键环节，便是生成能足以轻信受害人的深伪视频及语音，而这些深伪素材用料都是源自于被假扮对象公开发布的视频或语音，比如短视频平台上发布的某公司高管的视频、某名人明星的视频、各国国家领导人的视频等。然后用以上的明星名人角色身份，运用精准骗术在特定的场景向特定的对象实施诈骗。与此同时，随着计算机图形学技术和深度学习的发展，任何没有特殊技能的个人均可利用网上免费提供的软件包对图像或视频中的人脸进行篡改，生成高仿真深度伪造人脸图像或视频。不法分子可以通过互联网开源社区直接获取源代码并进行修改使用，大大降低了使用成本。其中，公安机关宣传工作中，无论是反诈、禁毒、交管、网安业务口的民警所作的宣传视频，还有高层警官在各

类新闻发布会的演讲，已经涌现了不少知名的“网红警官”。这些“网红警官”的面容、声纹数据，都有可能被诈骗分子精准采集，用于生成行骗所用的深度伪造视频。

未来的“终极 MAX 版”诈骗手法中，有可能受害人看到的人物视频、听到的熟人声音、朋友说话的内容，都是用生成式人工智能深度伪造的。目前，冒用公司高管、明星名人的角色身份行骗的作案手法，已越来越多地被诈骗分子用于向公司职员、追星粉丝实施诈骗。今后，诈骗分子还会更多地使用网络知名的公安民警形象身份行骗，研发出“冒充公检法”的“终极版”诈骗手法，以此达到范围更广、难度更低地骗取我国境内受害人财产的目的。以往已流传在网上的全国各地公安民警拍摄的正面宣传视频，如被不法分子更换嘴型唇语后制作成深伪诈骗视频，无论是制成广撒网诈捐的视频，还是用于点对点诈骗中的取信受害人，都可能会危及人民群众财产安全，不但伤害人民群众的感情和公安机关的形象，更有可能引起连串的负面舆情。细思极恐之下，公安机关必须重视，在各类宣传中都要对深伪诈骗进行主动防御。

三、可视身份深度伪造诈骗犯罪的防御策略

深伪犯罪作为一种新型犯罪形态，存在隐秘性强、取证难度大、犯罪形式多样、涉及领域广、危害性大等特性，预防和打击深度伪造犯罪已经成为当前公安机关的一个迫切需求。一些不法分子利用深度伪造技术恶意制作违法有害信息，并通过社交网络发布、传播，诋毁他人名誉、侵犯个人隐私，或利用深度合成技术仿冒他人身份实施诈

骗，严重扰乱网络传播秩序和社会秩序，危害公众利益、国家安全和社会稳定。这类违法有害信息，难以追溯制作源头，进一步增加了网络内容管理的难度。针对深度伪造内容生成传播平台非可控与可控两个不同场景，可从深度伪造的可靠被动检测防御和主动防御两个方向形成治理策略。

（一）被动检测防御

深度伪造检测对我国当前的公安技术而言，也是一块崭新的领域。从深度伪造技术诞生至今，有不少的研究工作展开对伪造内容进行检测，但是依然没有完美的解决方案。同时一些新的生成技术的发展成熟会让此类深度伪造的鉴别工作越来越困难，在被动检测领域依然存在着诸多研究难点问题。

早期的深度伪造检测主要是基于手工特征的机器学习方法。Zhang Y 等最先提出利用关键点描述子和经典机器学习方法检测人脸伪造图像。Yang X 等人提出了一种基于头部姿态位置的 3D 姿态差异的检测方法，该方法使用脸部中心区域的坐标位置和使用 DLib 提取的 68 个人脸关键点的差异作为区分真假人脸的特征。随着深度学习技术在图像识别与检测领域的广泛应用，研究人员利用能够对特征进行自适应提取的卷积神经网络进行深度伪造检测，主要分为基于图像级检测技术和基于视频级检测技术两类，如图 3 所示。

1. 基于图像级的检测技术

基于图像级的检测技术以单个帧为输入，主要关注深度伪造过程产生的空间伪影（如边缘模糊、纹理不一致、颜色异常等）。由于 DeepFake 中的人脸通常是以有限的分辨率创建的，需要进一步扭曲以匹配源视频中的原始人脸，导致伪造人脸区域和源背景内容的分辨率不一致，这种转换在生成的

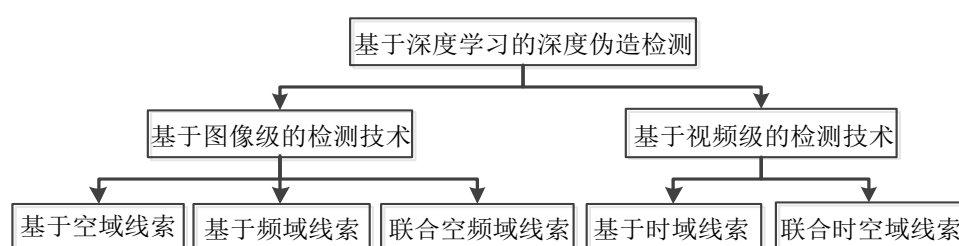


图3 基于深度学习的深度伪造被动检测技术分类示意图

DeepFake 视频中留下了独特的伪影。基于此，Li Y 等利用卷积神经网络获取仿射面部扭曲伪影特征，来区分真假人脸图像，通过模拟伪造操作过程（比如对面部区域进行高斯模糊处理）动态生成假样本的数据增强方式，来提高网络检测性能。每幅图像都有自己独特的标记或统计特征，来自硬件（CFA 阵列、传感器噪声等）和软件（有损压缩等），留下特定的特征往往是周期性的或均匀的，这可能在改变的图像中受到干扰。同时，人脸深度伪造算法都需要一个共同的操作步骤，即将改变后的人脸混合到现有背景图像中，在混合过程中存在内在差异。Li L 等通过使用混合边界的图像统计数据的不一致性来检测伪造的人脸图像，并对伪造图像的混合边界位置进行准确定位。伪造的图像将在不同的位置包含不同的源特征，而原始图像必须在所有位置上是一致的。Zhao T 等提出了一种利用源特征不一致的人脸深度伪造图像检测方法，通过提取局部源特征并测量其自一致性，可以检测到伪造的人脸图像。基于真实人脸和伪造人脸类内分布的差异性，增强低质量下检测效果，Li J 等提出了一种完全由数据驱动自适应频域特征提取方法，在一个基础网络中增加自适应频域特征生成模块和特征融合模块。输入图像首先经过两个分支分别提取空间域特征和频域特征，在融合模块中将两种特征进行融合，并进行端到端训练。Zhao H 等引入了

细粒度分类技术，提出了一种新的多注意网络结构来捕获人脸不同区域的局部鉴别特征，提升检测性能。

2. 基于视频级的检测技术

基于视频级的检测技术以一个帧系列为输入，并试图提取时间伪影（如帧间的不一致信息）。Guera D 等使用卷积神经网络对视频帧图片进行特征提取，然后将特定数量的连续视频帧图片的特征输入到循环神经网络中，用来区分视频是否为换脸视频，该方法结合了视频的空域和时域特征。Li Y 等认为换脸视频无法合成真实视频中人眼自然流畅的眨眼过程，提出了一个检测合成人脸眨眼的算法，根据是否出现眨眼异常来区分真实视频和换脸视频。一个正常的成人每次眨眼间距时间通常为 2-10 秒，每次眨眼时间大概是 0.1-0.4 秒。而在合成的换脸视频中，假人眨眼的频率则比真人较低。文献提出了一种融合卷积神经网络和递归神经网络的长时循环卷积神经网络，用于检测视频人物眨眼。它包含三个模块：用 CNN 进行特征提取；将提取的特征输入到 LSTM 中学习时域特征；最后将序列学习的结果输入到全连接层进行眨眼状态的预测。充分利用帧内和帧间信息是提高检测器有效性的途径之一，Chen B 等提出了时空注意机制，用于捕获和增强时空相关性，引入了卷积长短期记忆来考虑帧结构信息，同时对时间信息进行建模。为解决高视觉质量的深度伪造

检测问题，Hu J 等提出了一种基于帧推理的检测框架，首先学习当前和未来帧的人脸参考表示。然后，利用当前帧的面部表征和自回归模型预测未来帧的面部表征。最后，设计了预测损失算法，使真假视频的可判别性最大化。

总体来说，针对深度伪造的现有检测方案中普遍存在检测准确性不高、泛化性能差、计算耗时过长和抗干扰能力不强等缺陷，使得检测方案存在很大的局限性。因此，深入研究深度伪造检测技术，并将其应用在打击深度伪造诈骗犯罪和电子物证取证的公安业务工作中，研发实现服务公众的鉴伪平台是当前公安机关的迫切需求。公安机关可会同网信等有关部门，牵头研发面向人民群众的深伪鉴别平台工具，开设一个公众易用的鉴伪平台，或是在“国家反诈中心 APP”等手机软件中加入相应功能，当人民群众收遇到疑似深伪诈骗的视频时，可以将存疑视频上传到平台进行初步的鉴伪，从而及时阻断深伪诈骗作案，而群众提交的这些鉴伪需求也应当同步推送给警方，用以适时介入预警和精准宣防。

（二）主动防御

深度伪造的检测仅仅依靠技术手段不能完美地解决问题，因为生成与检测是一个永恒博弈的过程，仅依靠一门检测技术来杜绝深度伪造现象不太现实，需要在立法监管、新闻宣传、知识科普等多个维度协同攻坚。同时，需要对伪造图像与视频传播的源头进行控制，要明确在图像或视频上打上是否伪造的标签，以减少新闻媒体的误导能力。对此，公安机关除了加强反诈宣传和案件打击等常规工作外，还可以从网络安全及电子数据取证警务技术方向开展主动防御工作。

1. 基于对抗扰动或数字水印的主动防御

由于被动检测的方式难以做到“事前防御”，因此主动防御的思想得到公安机关的高度关注。这类技术的核心思想就是在图像或视频发布之前注入特定的对抗扰动或数字水印，当恶意用户需要进行深度伪造时，这些保护信息一方面对伪造结果造成视觉上的破坏，另一方面可以开展溯源工作，并对图像进行真伪鉴定。以中山大学网络空间安全学院的研究成果为例，针对深度伪造的潜在危害，可在公安机关今后的外宣视频中，加入主动防御与数字水印技术结合的对抗深伪技术方案。对于视频图像，一方面可通过在待保护图像中添加不宜觉察的对抗噪声，使得含噪图像输入到深伪软件之后会生成严重失真的图像，难以生成不法分子预期的输出图像，实现对特定人像的主动保护；另一方面，可进一步在含噪图像中嵌入鲁棒水印，这样即使待保护人像被用于深度伪造，依然可以通过所嵌入的水印信息，对伪造图像的来源及真实性进行溯源认证，为图像提供双重保护。对于声纹语音，同样可以加入主动防御噪声，防止不法分子的恶意采集。

2. 基于终端厂商的硬件防御

公安机关可通过警企合作等渠道，推动手机、电脑等网络通讯厂家，对深伪诈骗进行主动防御。比如，手机厂商可从设备底层硬件和操作系统中加入主动防御技术，当网民进行网络视频通话时，手机硬件一旦检测到对方可能是在用虚拟摄像头或播放视频或图音不一致时，能主动提示用户防范深伪合成的诈骗伎俩。而这种硬件层的主动防御技术终端厂商的制造成本和商业信誉，可能需要公安机关推动更高层面进行立法强制，

方能实现相关的技术主动防御。《反电信网络诈骗法》立足于前端治理的立法理念，直指电信治理、金融治理、互联网治理三个环节，贯彻了习近平总书记强调的“源头治理、综合治理，坚持齐抓共管、群防群治，全面落实打防管控各项措施和金融、通信、互联网等行业监管主体责任”重要指示精神。

3. 基于社交网络分析和知识图谱的深伪犯罪预测与风险评估模型构建

深伪犯罪作为一种新型网络犯罪形态，主要利用社交网络采集的数据作为工具来达到非法获利的目的，如利用 AI 换脸技术破解线上支付系统的人脸识别系统、利用深度伪造照片或视频进行诈骗等。因此，可以

对深伪犯罪样本数据进行分析 and 特征提取，采用社交网络分析和知识图谱引导，实现对社交网络中深伪犯罪的预测与风险评估。具体方法如图 4 所示。

4. 基于人技一体的深伪生成、传播、用户全链条管控防范机制

深伪犯罪链主要涉及深伪内容的生成端、网络平台传播端和网络用户端，在不同链条上，需要采取相应的有效监管措施，形成一体化管控防范机制，守住整个链条、每个节点的安全，建立“全链条”屏障。采用基于人技一体、分级管理的方法，对深伪犯罪产生的不同节点进行有针对性的管控。政府相关部门、网络平台方、深伪技术开发方、

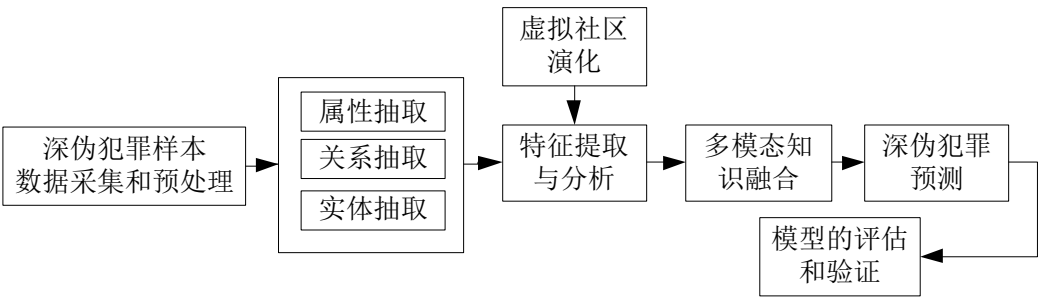


图 4 基于社交网络分析和知识图谱的深伪犯罪预测与风险评估模型

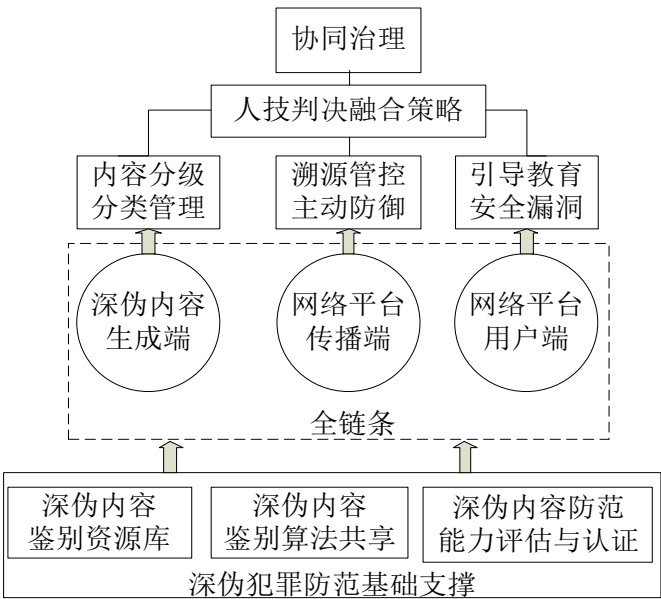


图 5 基于社交网络分析和知识图谱的深伪犯罪预测与风险评估模型

深伪制作方、网络用户等五方之间应建立类型多样的协同治理机制。对违法的深伪主体、深伪内容和相关的传播链等，进行协同治理，从而提升应对深伪犯罪问题的整体治理效果。具体方法如图 5 所示。

生成式深度学习算法的发展，驱动了以深度伪造为代表的人脸图像和视频伪造技术在互联网快速流行，催生了深度伪造诈骗新型犯罪模式。本文介绍了人工智能时代深伪犯罪的机理，并对其引发的深度伪造诈骗犯罪手段进行了分析，最后从被动检测和主动防御两方面提出了相应的防范策略，旨在为公安机关提供一套可行的深伪犯罪预防和打击体系，以应对当前和未来可能出现的深伪犯罪。在社会公共安全和队伍公安建设方面，将为管控深度伪造内容、深度伪造诈骗犯罪取证等提供理论和决策支撑。

参考文献：

- [1]Janko Roettgers. Porn Producers Offer to Help Hollywood Take Down Deepfake Videos [EB/OL]. 2018. 2. 21. <https://variety.com/2018/digital/news/deepfakes-porn-adult-industry-1202705749/>
- [2]梁瑞刚、吕培卓、赵月等. 视听觉深度伪造检测技术研究综述 [J]. 信息安全学报. 2020. 5
- [3]张俊杰、张纯琍. 基于深度伪造技术的AI诈骗风险与预防 [J]. 广西警察学院学报. 2023. 5
- [4]周文柏、赵汉卿、刘泓谷等. 深度伪造防御技术新进展 [J]. 人工智能. 2023. 3
- [5]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. 12
- [6]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. 4

- [7]李旭嵘、纪守领、吴春明等. 深度伪造与检测技术综述 [J]. 软件学报. 2021. 2
- [8]Rossler A, Cozzolino D, Verdoliva L, et al. Faceforensics++: Learning to detect manipulated facial images[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2019
- [9]Zhang H, Goodfellow I, Metaxas D, et al. Self-attention generative adversarial networks[C]. International conference on machine learning. PMLR. 2019
- [10]董凡超. 刷单返利类诈骗占发案总量近三分之一 [N]. 法治日报. 2023. 6
- [11]靳高风、张雍锐、郭兆轩. 2022—2023年中国犯罪形势分析与预测 [J]. 中国人民公安大学学报(社会科学版). 2023. 2
- [12]吴雪冬、刘彬、王子杰. 深度伪造技术新型犯罪的探析与治理 [J]. 政法学刊. 2023. 6
- [13]王琦. 生成式人工智能 (AIGC) 犯罪风险及应对研究 [J]. 公安研究. 2023. 9
- [14]杨骏. 人工智能领域电信网络诈骗犯罪防控策略研究 [J]. 公安研究. 2020. 6
- [15]Zhang Y, Zheng L, Thing V L L. Automated face swapping and its detection[C]//2017 IEEE 2nd international conference on signal and image processing (ICSIP). IEEE. 2017
- [16]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
- [17]Li Y, Lyu S. Exposing DeepFake Videos By Detecting Face Warping Artifacts[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition Workshops. 2019
- [18]Li L, Bao J, Zhang T, et al. Face x-ray for more general face forgery detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020
- [19]瞿左珉、殷琪林、盛紫琦等. 人脸深度伪造主动防御技术综述 [J]. 中国图象图形学报. 2024. 2
- [20]庄华、廖广军. “供给-需求”视角下电信网络诈骗犯罪治理框架 [J]. 公安研究. 2023. 2
- [21]张璐、季昊龙、江宁等. 涉深度伪造技术案件的打击与治理策略 [J]. 辽宁警察学院学报. 2024. 1
- [22]邱雅娴. 深度伪造型网络犯罪及其治理研究 [J]. 浙江警察学院学报. 2023. 3

责任编辑 张树彦