DeepFake 技术及深度伪造视频检测方法

刘文晨　2018080901006

摘要

人工智能的发展给社会生活带来了巨大的改变。然而，随着这些应用的推广，人工智能的安全问题也日益暴露出来。最近，以 DeepFake 为代表的深度伪造技术，严重威胁着社会安全和公众隐私。本文首先阐述了 DeepFake 技术的发展背景和技术原理，然后分析了近年来 DeepFake 技术在商业、政治、色情和娱乐等方面的影响。其次，综述了基于视频帧的深度伪造视频检测方法，针对深度伪造视频帧的视觉伪影、面部噪声特征的检测问题，介绍了相关机器学习、深度学习等分类算法。最后，总结了深度伪造视频检测技术的未来研究方向。

关键词：人工智能；计算机视觉；DeepFake；深度学习；视频帧

一、 DeepFake 技术介绍

1.1. 背景介绍

DeepFake 技术是一类基于人工智能来进行假数据生成的技术总称。从功能上讲，DeepFake 可以将多份不同的数据混合起来，生成新的数据。和一般的数据伪造方法不同，基于 DeepFake 的伪造方法可以自动生成以假乱真的人脸图片、视频、音频，从而给社会安全和内容安全带来极大的威胁和冲击。

2017 年，一种自称 DeepFake 的开源软件横空出世，在 Reddit 社区引起了巨大反响。作为一种编辑虚假信息的强大工具，DeepFake 的身影开始遍布互联网的各个角落：扰乱政治选举，抹黑公众人物，引发色情片泛滥，一步步侵蚀公众信任，引发社会信任危机。

1.2. 技术原理

我们以假人脸生成为例介绍 DeepFake 技术。传统影视制作中的换脸技术首先通过动作捕捉技术来获取一个人（记为 A，或者称为攻击者）的动作，然后用这些动作去驱动另一个人（记为 B，或者称为受害者）的三维人头模型，并渲染

相应并渲染成相应的二维图像。相比传统的影视制作，DeepFake 技术用计算机视觉方法来替代专业影视设备来做动作捕捉，并且用机器学习中的深度生成式模型来替代三维到二维的渲染过程，从而能显著地降低动作捕捉和三维模型中物理模型的构建过程，降低了假人脸生成的成本。在保持高逼真度的同时降低了造假的成本，是 DeepFake 造假技术泛滥的关键技术诱因。

具体地，DeepFake 技术会用一个称为自动编码器（Auto-Encoder，如图 1 所示）的模型结构来进行假人脸生成。

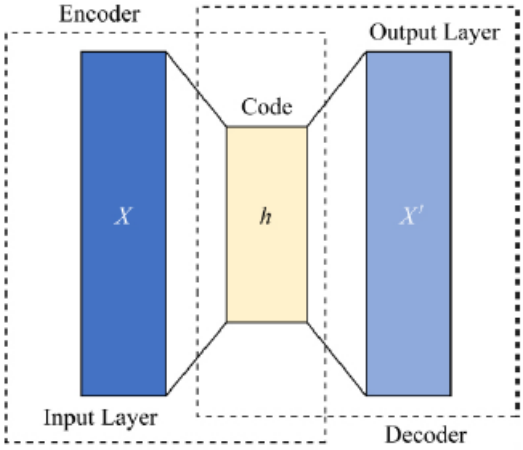


图 1 机器学习中的自动编码器

其中，自动编码器中的编码器（Encoder）会从二维的人脸图像中自动地推理出人脸的语义信息，例如动作、情绪等，起到动作捕捉的作用；自动编码器的解码器（Decoder）会将这些语义信息和另一个人的身份信息结合起来，渲染出假的人脸图片。

基于这套技术，DeepFake 有 2 种潜在的技术功能：人脸替换（face swap）和动作操纵（face reenactment）。人脸替换是指将一个视频中的人脸替换成另外一个人，来伪造受害者在新的场合下做一些实际上不是他做的事情。动作操纵是指在某个场景下，攻击者操纵受害者的表情或者嘴型，来伪造受害者在真实存在的场景下做一些实际上没有发生的事情。

二、 DeepFake 技术的机遇与挑战

随着实现成本的逐渐降低，DeepFake 开始走向技术普及，在社会生活各方面的渗透也越来越深。目前，公众对于 DeepFake 的态度正走向 2 个极端：一是陷入 DeepFake 娱乐化的漩涡；二是一味抵制 DeepFake。但正如“技术本身没有

善恶”这一科技价值观，DeepFake 也是一项新颖的科技创新。除了伦理道德层面的负面影响，DeepFake 也将产生全新的产业价值，比如在影视工业、数字媒体、广告公关行业中的应用。

目前，DeepFake 的衍生内容越发繁多而且真实，“硬币”的两面到底暗藏着哪些挑战和曙光，还需我们深入探究。

2.1. 商业

2019 年美国政府问责办公室发布的 DeepFake 简要报告指出，DeepFake 合成技术在影视娱乐、数字媒体、广告植入等领域有着良性应用，暗藏商业机会。

2.1.1. 影视特效

试想一下，电影拍摄完成后某特定镜头出现状况但无法重新拍摄，影视剧组该如何解决？传统的视觉特效实验室可能会采用真人镜头和计算机生成图像（CGI）的合成，而基于 DeepFake 的影像合成提供了一条更快捷、更低成本的“换脸”方案。通过提供演员的大量图片和视频，DeepFake 采用的神经网络能够自动学习到演员的面部数据，自动分析目标人脸的运动、光照条件，然后就可以直接完成替换，毫无违和感。整个过程只需要一段代码，没有复杂耗时的渲染过程，极大地缩短了制作周期与实现成本。

2.1.2. 广告营销

DeepFake 技术的出现将有可能实现千人千面的智能广告投放，大幅提高广告收益。日本京都某数据网络公司开发了一套“自动生成全身模型 AI”，基于 DeepFake 技术，以顾客自身形象生成虚拟模特，再通过在线“一键换衣”呈现出真人的试穿效果。不局限于线下，DeepFake 也为未来电商开展“线上试衣”，进行网络销售提供了新思路。

2.2. 政治

眼看许多西方国家的总统大选即将到来，很多国家开始担心 DeepFake 虚假视频会被参选人员用于操纵选举、煽动毫不知情的民众和控制舆论。为了保证大选的顺利进行，部分国家正式立法禁止在选举期间进行换脸行为。

2.2.1. 操纵选举

在西方国家，假新闻（fake news）一直是政治选举中长期面临的重要问题。但与以往不同的是，DeepFake 伪造视频以假乱真的能力更强，谎言也因此变得

更加强大。北美法学界不少学者认为 DeepFake 会在更大程度上影响选举进程，并破坏选举的合法性。2018 年 4 月，DeepFake 首次被用在政客身上。当时美国喜剧演员约旦皮尔制作了一段调侃视频，但在另一个视频中，他被换脸成前总统奥巴马，然后在演讲中发表了侮辱现总统特朗普的讲话。自此，自由创作者们利用 DeepFake 针对许多政客开始了“大规模报复”，英国首相热门候选人伯尼桑德斯、英国工党领袖杰里米科宾等人无一幸免，沦为公众笑柄，形象受损在所难免。

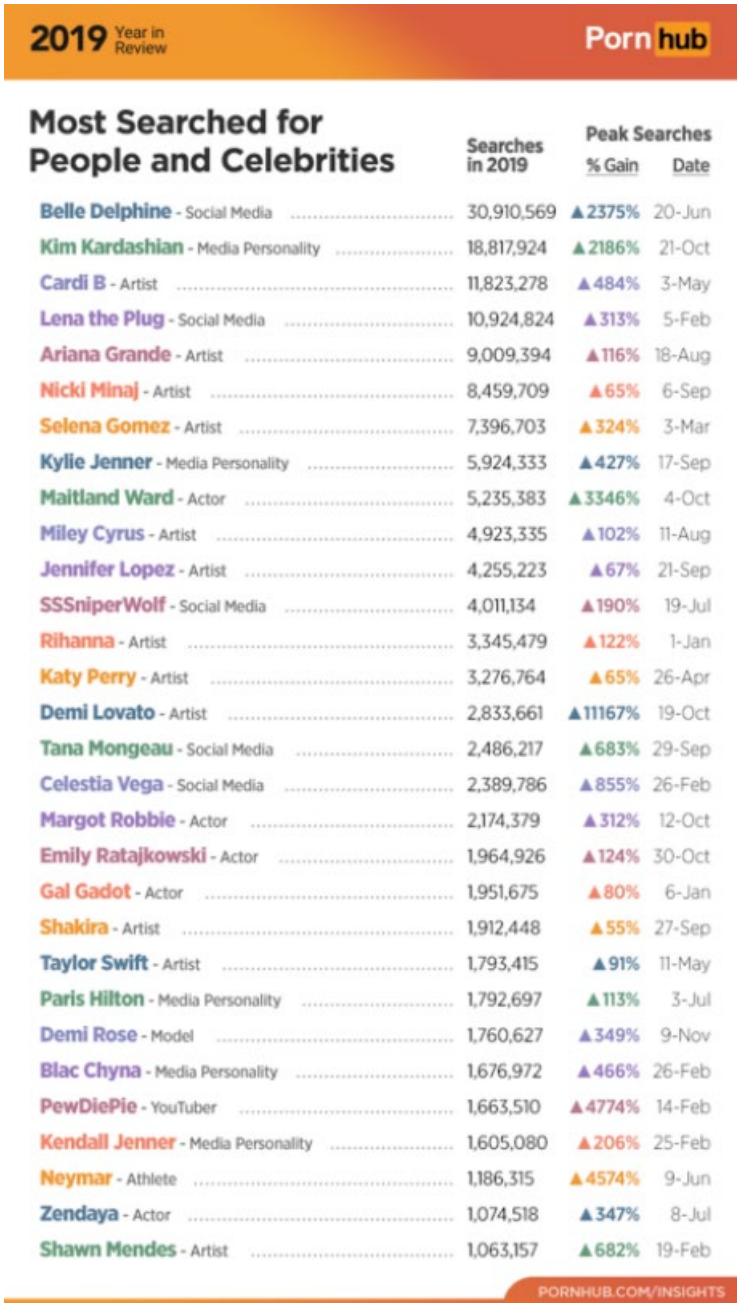
2.2.2. 信任危机

美国独立性民调机构皮尤研究中心发布的一项民意调查显示，近 70%的受访者称虚假新闻和信息极大地影响了他们对政府机构的信心。该调查同时显示，近一半的受访者将虚假信息视为比恐怖主义、非法移民等更严重、更值得国家着手去解决的首要任务。随着 DeepFake 虚假内容的激增，社会被错误的信息激化，民众对于执政者的不信任感增加，继而对国家的信赖感降低。

2.3. 色情

2019 年底，全球最大成人网站 Pornhub 发布了“年度回顾”，其中统计了 2019 年度该平台上搜索量最多的 30 位明星，但 Pornhub 没有说明，这些搜索背后显示的都是这些明星的“假色情片”。据统计，“名人色情”已经成为 Pornhub 平台上最受欢迎的搜索类型之一。

色情是互联网的第一生产力。过去 1 年，DeepFake 爆炸式增长，其中色情相关的 DeepFake 成为最大的利基市场，呈急速增长态势。那为什么平台会纵容这些虚假视频内容的出现呢？深究之后可以发现，这类“名人色情”由于曝光度高，往往会被投放广告，上传者上传这类视频，累计足够观看量后能够争取不少费用，并与 Pornhub 平台分成，违反规定成本却极低，只有被换脸的主人公成为整个环节中唯一的受害者。



2019 Year in Review

Pornhub

Most Searched for People and Celebrities

	Searches in 2019	Peak Searches % Gain	Peak Searches Date
Belle Delphine - Social Media	30,910,569	▲2375%	20-Jun
Kim Kardashian - Media Personality	18,817,924	▲2186%	21-Oct
Cardi B - Artist	11,823,278	▲484%	3-May
Lena the Plug - Social Media	10,924,824	▲313%	5-Feb
Ariana Grande - Artist	9,009,394	▲116%	18-Aug
Nicki Minaj - Artist	8,459,709	▲65%	6-Sep
Selena Gomez - Artist	7,396,703	▲324%	3-Mar
Kylie Jenner - Media Personality	5,924,333	▲427%	17-Sep
Maitland Ward - Actor	5,235,383	▲3346%	4-Oct
Miley Cyrus - Artist	4,923,335	▲102%	11-Aug
Jennifer Lopez - Artist	4,255,223	▲67%	21-Sep
SSSniperWolf - Social Media	4,011,134	▲190%	19-Jul
Rihanna - Artist	3,345,479	▲122%	1-Jan
Katy Perry - Artist	3,276,764	▲65%	26-Apr
Demi Lovato - Artist	2,833,661	▲11167%	19-Oct
Tana Mongeau - Social Media	2,486,217	▲683%	29-Sep
Celestia Vega - Social Media	2,389,786	▲855%	26-Feb
Margot Robbie - Actor	2,174,379	▲312%	12-Oct
Emily Ratajkowski - Actor	1,964,926	▲124%	30-Oct
Gal Gadot - Actor	1,951,675	▲80%	6-Jan
Shakira - Artist	1,912,448	▲55%	27-Sep
Taylor Swift - Artist	1,793,415	▲91%	11-May
Paris Hilton - Media Personality	1,792,697	▲113%	3-Jul
Demi Rose - Model	1,760,627	▲349%	9-Nov
Blac Chyna - Media Personality	1,676,972	▲466%	26-Feb
PewDiePie - YouTuber	1,663,510	▲4774%	14-Feb
Kendall Jenner - Media Personality	1,605,080	▲206%	25-Feb
Neymar - Athlete	1,186,315	▲4574%	9-Jun
Zendaya - Actor	1,074,518	▲347%	8-Jul
Shawn Mendes - Artist	1,063,157	▲682%	19-Feb

PORNHUB.COM/INSIGHTS

图 2 Pornhub 上包含“假色情片”搜索量排名

2.4. 语音

除了假视频，深度伪造的假音频已经被用于诈骗。2019 年 9 月，据《华尔街日报》报道，英国某能源公司发生 1 起电话诈骗案件，黑客利用 DeepFake 技术生成的语音冒充该公司的 CEO，指示英国子公司对 1 笔急单汇款，成功骗取 24.3 万美元。

就像互联网可以被武器化一样，“语音克隆”技术也呈现出“武器化”趋势。而且不同于最开始的机器学习方法，需要使用大量音频素材提高“克隆语音”的

质量，现在的“语音克隆”技术已经发展到可以使用非常小的音频样本大规模克隆声音的地步。比如 5 分钟的音频素材就可以训练出相当逼真的 DeepFake 版数字语音，而 5 小时或更长时间的音频素材则可能完全“复制”出你的声音，这一“复制”的声音甚至能够完全欺骗声纹识别系统。

事实上，除了电话诈骗，利用 DeepFake 生成的音频还可以与电子邮件欺诈、网络钓鱼攻击等形式相结合，形成更大规模的诈骗。

2.5. 娱乐

DeepFake 在一定程度上可以满足目前日益多元化的娱乐需求，视频可以直接带给人们强烈的视觉冲击。目前，DeepFake 在娱乐领域已经有诸多成功的案例。

最近，1 家名为 Boogie 的初创公司声称，可以使用单一照片生成舞蹈视频（如图 3 所示）。Boogie 希望“采用 AI 实现社交媒体幻想民主化”，人们可以通过一张自拍来生成自己在社交媒体平台上的“化身”，这些“数字化身”可以随心随意地更换服装、环游世界，并尝试最新的 TikTok 热门舞蹈。

图 3 Boogie 宣传网站

三、 基于视频帧的深度伪造视频检测方法

帧是组成视频的基本单位，视频通过逐帧播放向观众传递信息。DeepFake 往往通过逐帧的方式对面部的特定区域进行篡改，其在各帧内部会出现视觉伪影和视觉噪声，在帧间会出现人物时空状态的连续性不一致的情形，为检测 DeepFake

视频提供了依据。

3.1. 基于帧内差异的检测方法

DeepFake 视频由于通常选择在人的面部中心区域交换人脸，而不是对整个面部进行篡改，因此会出现视频中人脸中心的伪造区域与人脸边缘真实区域无法很好拟合的视觉差异，如亮度、颜色、像素不同。这些差异能够通过机器学习算法、深度学习模型（如卷积神经网络）或者其他分类算法进行区分。

3.1.1. 基于机器学习等算法的面部关键部位伪造特征检测

基于机器学习等算法对面部关键特征进行检测时，首先需要提取人脸关键部位。图 4 为人脸面部关键点（face landmark）提取的示意图，其通常通过 ASM 等人脸对齐算法实现，以便于后续的人脸识别和分析。

图 4 面部关键点提取

在使用面部识别算法提取人脸关键部位之后，可针对眼睛颜色等面部特征、3D 头部姿态和人脸关键点位置的概率分布等方面的差异检测出伪造人脸。

1） 眼睛颜色等面部特征差异

如图 5、图 6 所示，在提取眼部特征点后，可通过常用的计算机视觉方法（如颜色直方图、颜色聚合向量）提取眼睛的颜色特征，然后将其作为 KNN 分类器选取的特征，从而对人脸图像做出真伪鉴别。

此外，伪造人脸的面部还会出现鼻边阴影、眼睛缺少反射细节、牙齿没有规则的几何结构等情况，这些特征可以通过逻辑回归算法或者简单的神经网络检测到。

图 5 伪造人脸眼睛颜色差异

图 6 人眼颜色特征的提取

2） 3D 头部姿态差异

用旋转矩阵R和平移向量t评估伪造视频和真实视频人物头部的方向、姿态差异。式（1）通过最小化表征视频帧头部方向和姿态的三维和二维坐标之差求解R和t，其中$[U,V,W]^T$和$[x,y]^T$为从人脸面部关键点获取的标准坐标和图像坐标，(c_x,c_y)和(f_x,f_y) 为相机的光学中心和焦距。如图 7 所示，由于伪造过程通常选择在中心人脸区域进行篡改，因此真实视频的整个人脸面部和中心人脸区域所评估的平移向量基本吻合（见图 7（k）），但深度伪造视频中两个向量会在方向、大小上显示出较大差异（见图 7（n））。选取真假人脸的中心区域（见图 7（i）、图 7（l））和整个面部区域（见图 7（j）、图 7（m））的旋转矩阵差$R_a - R_c$和平移向量差$t_a - t_c$放入 SVM 分类器进行分类训练，能够检测出伪造视频。

$$\min_{R,t,s} \sum_{i=1}^{n} \left\| s\begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix} - \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \left(R\begin{pmatrix} U_i \\ V_i \\ W_i \end{pmatrix} + t \right) \right\|^2 \tag{1}$$

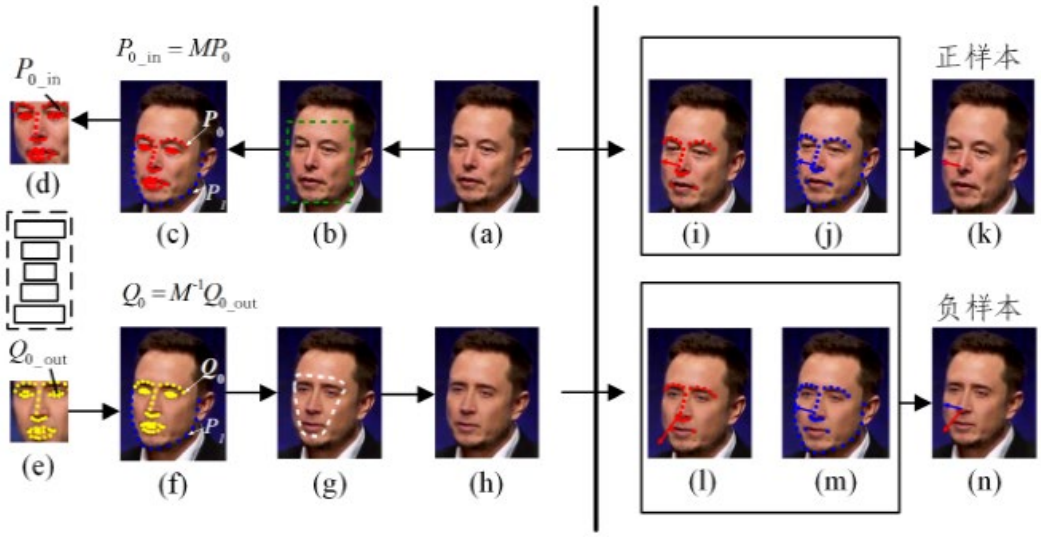


图 7 基于 SVM 的 3D 头部姿态差异检测

3) 人脸关键点的概率分布差异

如图 8 所示，通过上述人脸面部关键点的提取，得到面部的 68 个标记点；接着使用人脸对齐算法将所有的标记点通过仿射变换归一化到标准坐标系下，并去除面部边界上的点；再将这些面部区域标记点的位置向量化后作为特征向量来训练 SVM 分类器，从而检测出人脸的真伪。

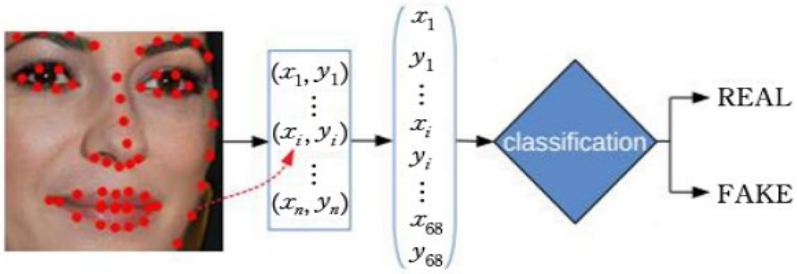


图 8 基于人脸关键点的概率分布差异检测

此外，由于通过血管的血液量越大，人皮肤表面反射的光线就越少，因此可通过对视频帧进行时频分析来估算人的心率和面部反射细节。欧拉影像放大算法，通过空间滤波减少视频图像噪声，利用时域滤波提取所研究的若干频带，用泰勒级数来差分逼近频带信号并线性放大所得结果，通过对比真伪视频中人脸面部变化信号的频率、振幅来鉴别深度 DeepFake 视频。

3.2. 基于帧间差异的检测方法

由于深度伪造视频在生成的过程中是逐帧进行的，因此对每一帧进行深度伪

造操作时难以兼顾之前已经伪造过的帧序列，从而导致深度伪造视频的连续帧会在时空分布上显示出差异，即伪造视频中的人物随着视频的逐帧播放会显示出眨眼频率明显较低、面部动作变化不协调、人脸亮度逐帧发生变化的情况，因此深度伪造视频能够被循环神经网络 RNN 或其他与序列数据有关的算法捕捉到。

3.2.1. CNN 和 RNN 结合的检测办法

如图 9 所示，一种 CN 和长短期记忆网络（Long Short-Term Memory, LSTM）相结合的方法来检测深度伪造视频。

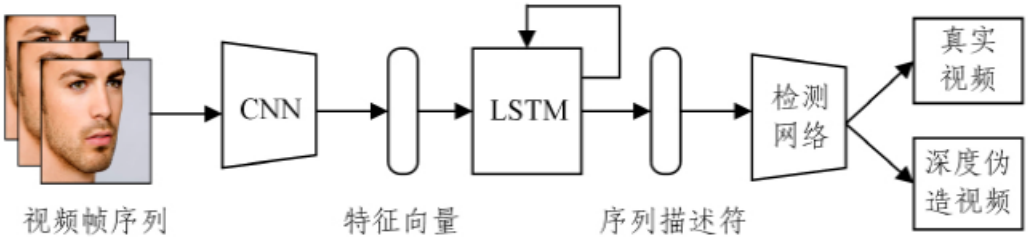


图 9 基于 CNN 和 RNN 结合的检测办法

LSTM 通过门控控制传输状态，选择性遗忘与当前帧时序特征无关的上一节点信息，仅保留对相关人脸特征进行逐帧分析的相关信息。实验在数据处理阶段提取关键帧，数目分别为 20,40 和 80，在将数据样本送入 CNN 之前通过减去样本通道均值实现快速收敛。实验采用 InceptionV3 模型提取视频帧的特征，该模型通过非对称的卷积结构拆分增强了其非线性表达能力，节约了大量参数，同时减轻了过拟合。经池化层降维，将提取到的特征向量送入 LSTM 对帧序列的时序状态进行学习，引入反向随机失活概率减轻验证阶段的负担。最后，由全连接层对帧序列特征做特征加权，使用逻辑回归与 softmax 函数求得深度伪造视频的概率。实验使用 HO-HA 数据集，按照 70:15:15 的比例划分训练集、验证集和测试集，最终得出各视频提取 80 帧进行检测时准确率最高（为 97%）的结论。

因为生成深度伪造视频时采用的数据集图像人物很少有闭眼的状态，所以伪造视频中人物的眨眼频率要低于真实视频中的 17 次/min，甚至会出现不眨眼的状态；同时，眨眼动作是一个与时间有关的序列，因此可以将 RNN 应用于眨眼检测。将视频中帧序列的人脸对齐到同一坐标系下后单独提取与眼部有关的区域，采用 VGG16 模型通过连续 5 次的卷积操作提取出眼部区域的可区分特征，然后将其输入 LSTM，并采用基于时间的反向传播算法（Back Propagation Trough Time, BPTT）沿需要优化的参数的负梯度方向不断寻找更优的点直至收敛，最后由全

连接层做出分类。在两次眨眼之间通常会维持一段睁眼状态，该模型可以检测眨眼的持续时间和两次眨眼之间的时间间隔在真伪视频中的差异。实验验证，CNN与 LSTM 结合的模型因引入对时间序列的学习，因此较仅使用 CNN 学习睁闭眼图像来预测眨眼状态更具优势。

四、 总结与展望

本节对上述深度伪造视频检测技术各自的优缺点和未来面对的挑战进行以下总结。

1) 针对基于帧内差异的检测方法，对比分析机器学习算法和深度学习算法。

利用机器学习算法依据人脸面部关键点特征做出决策，较深度神经网络对人脸所有特征进行全面学习和训练，能够从较低的数据维度对真伪视频人脸做出分类，且模型训练用时较短；但不能应对高质量的深度伪造视频。

利用深度学习模型能有效解决机器学习依赖于提前对人脸面部关键部位进行定位的弊端，并且能够使用端到端的训练方式对视频帧的高维数据特征进行充分学习；但是该方式依赖于 GPU 的运算能力，且训练时间很长。因此，可以结合机器学习和深度学习模型，先使用机器学习算法对伪造视频做初步分类，再使用深度神经网络对伪造人脸细节做进一步鉴别。

2) 对比分析帧内差异检测方法和帧间差异检测方法。

基于帧内差异的检测方法通常只提取深度伪造视频的个别帧，将伪造视频检测转化为伪造图像检测。其优点是能够更多地关注伪造视频帧内人脸面部主要器官（眼睛、鼻子、嘴巴等相关区域）的细节特征，从而针对不同特征提出具体的检测方案。但是该方法缺少对人脸面部表情和动作随时间变化的时序特征的理解，同时随着深度伪造技术的不断成熟，伪造人脸会更加逼真，因此会给帧内差异的检测方法带来更大的挑战。

基于帧间差异的检测方法能够充分挖掘伪造视频逐帧播放过程中的上下文信息，从而更充分地提取其时序特征。由于现有的主流伪造方法通常对视频人脸逐帧进行训练替换，因此该方法能够起到很好的检测效果。但是其缺乏对帧内伪造特征细节的充分理解，同时如果视频长度过短，会因提取关键帧数量不足而无法充分挖掘其时序特征，导致检测效果下降。

因此，可以分别提取 DeepFake 视频帧内面部特征细节信息和帧间全局时序

信息，并进行特征融合，用更全面的方式对真伪视频进行分类。

3) 由于当前主要检测方法还是依赖于机器学习算法和深度学习模型，因此人工智能本身存在的瓶颈成为阻碍深度伪造视频检测技术进一步发展的难题。

首先，人工智能计算机视觉领域的对抗样本生成，使得攻击者能够通过在源数据集上增加人类很难利用视觉分辨出的细微变化而导致相关模型做出错误的分类决策。由于深度伪造视频数据维度过高，因此在训练样本没有覆盖的区域可利用模型对分类边界的不确定性生成相关对抗样本，通过增加对抗扰动来干预人脸面部检测的结果，使得分类效果大幅降低。针对模型易受对抗样本攻击的问题，可在伪造人脸检测模型的训练阶段加入适当的对抗样本，以增强模型抵御对抗样本攻击的鲁棒性。

其次，深度学习模型对特定数据分布具有依赖性，通过训练能够对给定的深度伪造视频数据集做出准确的判断，但对跨数据库的检测准确率会下降。针对此问题，可以引入元学习和小样本学习，增强模型对数据集的泛化能力，增加模型的可解释性，使模型拥有真正类似人类的快速学习能力，提升模型应对不同场景的鲁棒性。

此外，视频的压缩和分辨率的差异也会影响模型的检测效果，使用低压缩率的伪造视频数据集进行训练后，在对高压缩率的视频进行检测时准确率会大幅下降。因此，可以在 CNN 的基础上，结合视频帧的噪声流或者对视频做傅里叶变换将帧转化到新域，将离散信号分解成不同频率的正弦分量，通过分析其谱相图解决此问题。

最后，仅通过人工智能模型无法应对所有篡改算法，应该结合区块链智能合约等新兴技术构建数字媒体信任机制，以实现对视频真实性的溯源。同时，可以借鉴数字水印等传统信息安全手段或预先提取视频的关键信息作为视频指纹，来作为视频真实性检测的补充手段，并制定相应的法律法规实现对深度伪造视频的全面打击。

五、 参考文献

[1]李永强,白天.基于卷积 LSTM 的视频中 Deepfake 检测方法[J].信息技术与网络安全,2021,40(04):28-32.

[2]耿鹏志,樊红兴,张翌阳,唐云祁.基于篡改伪影的深度伪造检测方法[J/OL].计算

机工程:1-13[2021-06-18].https://doi.org/10.19678/j.issn.1000-3428.0060733.
[3]暴雨轩,芦天亮,杜彦辉.深度伪造视频检测技术综述[J].计算机科学,2020,47(09):283-292.
[4]姬德强.深度造假：人工智能时代的视觉政治[J].新闻大学,2020(07):1-16+121.
[5]Murphy Gillian,Flynn Emma. Deepfake false memories.[J]. Memory (Hove, England),2021.
[6]Nguyen Xuan Hau,Tran Thai Son,Le Van Thinh,Nguyen Kim Duy,Truong Dinh-Tu. Learning Spatio-temporal features to detect manipulated facial videos created by the Deepfake techniques[J]. Forensic Science International: Digital Investigation,2021,36.
[7]Whyte Christopher. Deepfake news: AI-enabled disinformation as a multi-level public policy challenge[J]. Journal of Cyber Policy,2020,5(2).