文章编号：1003-0530(2021)12-2338-18

人脸视频深度伪造与防御技术综述

周文柏 张卫明 俞能海 赵汉卿 刘泓谷 韦天一
（中国科学技术大学，网络空间安全学院，安徽合肥 230027）

摘 要：近年来，得益于深度生成模型的发展，人脸的操控技术取得了巨大突破，以 Deepfake 为代表的人脸视频深度伪造技术在互联网快速流行，受到了学术界和工业界的广泛重视。这种深度伪造技术通过交换原始人脸和目标人脸的身份信息或编辑目标人脸的属性信息来合成虚假的人脸视频。人脸深度伪造技术激发了很多相关的娱乐应用，如使用面部替换技术将使用者的人脸替换到某段电影片段中，或使用表情重演技术来驱动某个著名人物的静态肖像等。但当前人脸深度伪造技术仍处于快速发展阶段，其生成的真实感和自然度仍有待进一步提升。另一方面，这类人脸深度伪造技术也很容易被不法分子恶意使用，用来制作色情电影、虚假新闻，甚至被用于政要人物来制造政治谣言等，这对国家安全与社会稳定都带来了极大的潜在威胁，因此伪造人脸视频的防御技术至关重要。为了降低深度伪造人脸视频所带来的负面影响，众多学者对伪造人脸视频的检测鉴别技术进行了深入研究，并从不同视角提出了一系列防御方法。然而由于数据集分布形式单一、评价标准不一致、主动性不足等问题，使得防御技术在走向实用的道路上仍有很长一段距离。事实上，人脸深度伪造与防御技术的研究仍旧处在发展期，其技术的内涵与外延正在快速的更新与迭代。本综述将对迄今为止的主要研究工作进行科学系统的总结与归纳，并对现有技术的局限性做简要分析。最后，本文将探讨人脸深度伪造与检测技术的潜在挑战与发展方向，为领域内未来的研究工作提供借鉴。

关键词：人脸视频深度伪造；人脸伪造视频防御；生成技术；视频取证；检测技术；主动防御
中图分类号：TP311 **文献标识码**：A **DOI**：10.16798/j.issn.1003-0530.2021.12.007

引用格式：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. DOI：10.16798/j.issn.1003-0530.2021.12.007.

Reference format：ZHOU Wenbo，ZHANG Weiming，YU Nenghai，et al. An overview of Deepfake forgery and defense techniques[J]. Journal of Signal Processing，2021，37(12)：2338-2355. DOI：10.16798/j.issn.1003-0530.2021.12.007.

An Overview of Deepfake Forgery and Defense Techniques

ZHOU Wenbo ZHANG Weiming YU Nenghai ZHAO Hanqing LIU Honggu WEI Tianyi
（University of Science and Technology of China，School of Cyberspace Science and Technology，Hefei Anhui 230027，China）

Abstract：Benefiting from the development of the deep generative model，face manipulation technology has made great breakthroughs. The deep face forgery technology，represented by Deepfake，has rapidly become popular on the Internet and received extensive attention from academia and industry. This deep forgery technique synthesizes fake face videos by exchanging the identity information between original and target faces or by editing the attribute information of target faces. The deep face forgery technique has inspired many related entertainment applications，such as using facial substitution to replace the user's face in a movie clip，or using expression reenactment to drive a static portrait of a famous person. However，the current deep face forgery technology is still in the fast-growing stage，and its authenticity and naturality require further improvement. On the other hand，this kind of deep face forgery can easily be used by malicious criminals to produce pornographic movies，fake news，or even political rumors about important dignitaries，which is a great potential threat to

收稿日期：2021-07-01；修回日期：2021-09-04
基金项目：国家自然科学基金(U20B2047，62002334)；安徽省自然科学基金(2008085QF296)；中国科学技术大探索类基金(YD3480002001)以及中国科学技术大学青年基金（WK2100000011）资助

national security and social stability, so the defense technology of face forgery videos is crucial. In order to reduce the negative impact of deep face forgery videos, many researchers have investigated the detection and identification techniques of face forgery videos, proposing a series of defense methods from different perspectives. Unfortunately, due to the single form of dataset distribution, inconsistent evaluation metrics and lack of initiative, etc., there is still a long way to go before the detection technology can be applied practically. In fact, the research on deep face forgery and detection techniques remains in the developmental stage, their connotations and extensions are being rapidly updated and iterated. In this review, we will make a scientific and systematic summary of the main research work so far, and briefly analyze the limitations of existing technologies. Finally, we will discuss the potential challenges and development directions of deep face forgery and detection technology, in order to provide insights for future research work in the field.

Key words: Deepfake forgery; Deepfake defense; generation techniques; video forensics; detection techniques; initiative defense

1　引言

2017 年 12 月，一位名为“Deepfakes”的用户在全球流量排名第四的国际互联网社区“Reddit”上发布了一段好莱坞女星盖尔·加朵的伪造人脸视频，掀起了一阵轰动，这一事件作为开端，标志着人脸深度伪造技术的兴起，而该用户的用户名也被引用成为了这一类技术的代名词“Deepfake[1]”。人脸的深度伪造主要针对成对的人脸进行，它可以将目标视频人物的脸替换成指定的原始视频人脸，或让目标人脸重演、模仿原始人脸的动作、表情等，从而制作出目标人脸的伪造视频。近年来，人脸深度伪造技术在深度学习技术的推动下取得了快速发展，自动编码器[2-4]、生成对抗网络[5-7]等深度学习中的生成模型均被应用到了该项技术中，实现了人脸的高真实度生成与替换。深度学习技术的普适性以及高度的开源性使得 Deepfake 相关技术只需要利用少量的原始人物与目标人物肖像便可以完成换脸过程，这带来了很多换脸技术在娱乐中的应用，如 2019 年非常流行的“ZAO APP[8]”和最近新上线的“去演 APP[9]”，用户可以用自己的脸去替换部分影视剧片段中演员的脸，虚拟的体验一把“当明星”的感觉，此外，“抖音”、“Avatarify”等知名的社交软件也提供了换脸相关的新功能，用户可以用自己的表情去驱动一些著名人物的静态肖像，让他们做出跟自己一致的表情，使静态名人“活化”。

然而，人脸深度伪造技术在为人们带来娱乐性的同时也带来了巨大的安全威胁。由于深度学习的技术门槛相对较低，很多开源技术可以直接利用，因此一些不法分子能够轻易未经许可地伪造特定人物的假视频并恶意使用，如将女明星的脸换到成人电影中，对当事人造成隐私和名誉的侵害，或操纵特定人物肖像发布虚假信息，误导舆论，侵蚀社会信任度，甚至能够将换脸技术用于国家政要，发布涉及国防外交的虚假视频，导致严重的政治危机。因此对伪造人脸视频的检测鉴别至关重要。为了减轻人脸深度伪造技术带来的危害，学术界和工业界都展开了对人脸伪造视频检测技术的深入探索，并提出了涵盖空域、时域、频域等多种维度的检测方法，并在特定数据集上取得了一定成功。此外，Facebook（脸书）公司也联合 MicroSoft（微软）和 MIT（麻省理工）等知名企业与高校，在著名竞赛平台 Kaggle 上发布了迄今为止人脸伪造检测领域规模最大、影响力最高，也是 Kaggle 竞赛平台有史以来奖金最丰厚的人脸伪造检测挑战赛[10]（Deepfake Detection Challenge, DFDC），吸引了全球超过 2200 支队伍参赛，以期推动检测技术的发展。尽管众多学者在人脸深度伪造检测领域做出了不小的努力，但由于高质量数据集缺乏、评价机制不统一、实验环境与真实场景数据失配情况严重等问题，导致相关检测技术的实用性受到极大制约，Facebook 官方也在 DFDC 比赛后宣称：“人脸深度伪造检测的问题仍尚未得到解决”。

事实上，尽管目前人脸深度伪造所带来的负面影响远大于积极效应，但人脸深度伪造技术本身却依旧值得研究。现有的人脸深度伪造技术在生成质量的真实感和自然度上仍有可以提升的空间，且伪造技术的进步将带来数据集质量的大幅提升，从而进一步推动检测技术的发展。在这样的技术发展背景下，本文将针对人脸深度伪造与检测领域的重要工作展开综述，通过对现有主流技术进行整理与归纳，总结人脸深度伪造与检测领域的发展脉络

与状态,提炼相关技术的内涵与外延,并找出现有工作的局限性和未来潜在的发展方向,促进领域的进步。

本文的贡献主要有以下四点:

(1)整理了三大类不同发展时期的人脸塑造技术手段,并针对每种技术手段的特点和适用场景进行了阐述;

(2)归纳了现有的人脸视频伪造与防御技术,并客观地针对主流方法进行了优劣势分析;

(3)首次对人脸视频深度伪造与检测技术的内涵与外延作出解释与探索,并相应地对现有算法类型进行了归纳;

(4)总结了人脸视频深度伪造与检测领域面临的技术难题与应用挑战,为未来领域发展方向提供指导。

2 "换脸"技术发展史

对于人脸的相关研究已经有数十年的历史,"换脸"技术也是计算机视觉和计算机图形学领域的一项经典任务,但以"Deepfake"为代表的深度换脸技术首次将深度学习技术应用到了换脸任务中,将这一传统任务变得智能化和高度真实化,形成了巨大的轰动效应。在"换脸"的发展历程中,主要有三类技术代表,即以"人脸变换(Face Morphing)[11-12]"为代表的经典换脸技术、以"计算机图形学技术(Computer Graphics, CG)[13-14]"为代表的高真实度换脸技术,以及以"生成模型(Generative model)[15-16]"为核心的智能化换脸技术。这三类技术的发展阶段不同,也具有不同的特点和适用场景。

2.1 人脸变换(Face Morphing)

人脸变换指的是将一张脸逐渐变换成另一张脸的过程,该过程中的任一张图都是由原始人脸S和目标人脸T通过参数控制来混合叠加的,通过改变参数,可以得到过程中的某一张脸M,使其看起来既像原始人脸A又像目标人脸T,如图1所示。

图1 人脸变换过程示意图

Fig. 1 The process of Face Morphing

Face Morphing技术主要有以下几个核心步骤:

1)特征点定位。使用开源的人脸识别工具从图像中检测出人脸,并对人脸关键点进行标定,定位出68个关键点。

2)区域划分。在定位到的特征点基础上,对两个脸分别进行区域划分,最常用的是三角剖分,也可以进行多边形划分。

3)变换关系计算。计算原始人脸和目标人脸同一区域之间的变换关系,通常计算两个人脸上对应区域顶点之间的仿射变换矩阵,该变换关系主要由原始人脸和目标人脸的特征点位置决定。

4)图像合成。根据步骤3中得到的变换关系,合成新的人脸图像。合成过程中,分别对目标图像的每个像素进行处理,根据其所处区域的变换关系插值得到其与原始图像上点的仿射变换关系。最终合成的人脸,其形状将主要由目标人脸决定,而纹理则取决于合成时颜色的权重设置,一般可以设置为$\alpha\times\text{src}+(1-\alpha)\times\text{tar}$,其中src指的是原始人脸图像中像素的颜色值,而tar指的是目标图像中像素的颜色值,通过控制α,可以使合成人脸偏向原始人脸或目标人脸,例如,α取1,则将源图像纹理贴到目标图像,α取0.5,则为二者平均。

人脸变换是在原始人脸和目标人脸的参考下,通过变换合成了新的人脸,并非真正意义上的人脸替换。

2.2 计算机图形学(Computer Graphics, CG)技术

计算机图形学(CG)技术是近二十年来不断完善的一类技术,以CG技术为基础,可以实现高真实度的人脸伪造。CG技术进行换脸首先需要对人脸进行建模,通过"人脸捕捉"过程对原始人脸的表情、纹理甚至是肌肉运动进行捕捉与建模,在得到面部的基础模型后,通过纹理合成与渲染获得逼近真实效果的人脸图像。

为了获得真实的人脸建模结果,通常需要真实扮演者穿戴专业的传感设备来进行人脸捕捉,在面部不同区域设置追踪器,捕捉这些点在空间中的位置,来还原真实演员的面部细节与动作。CG技术能够用来实现人脸的表情重演,如图2(a)所示,在电影《阿凡达》中,由演员穿戴的传感器捕捉得到演员的面部动作,再通过计算机渲染合成出最终的虚拟人物,该虚拟人物能够完全重现演员的动作细

节。此外，CG 技术也能够实现人脸的替换，如图 2(b)所示，左边为演员真实人脸，通过面部的纹理捕捉再结合渲染与合成技术，最终将合成后的人脸贴到替身演员的面部，得到右边的替换结果。

(a) 人脸面部重演
(a) Face reenactment

(b) 人脸合成替换
(b) Face synthesis and swapping

图 2　部分电影片段中的 CG 人脸合成效果

Fig. 2　CG face synthesis in movie clips

聚焦于人脸的 CG 特效，其核心技术包括人脸检测、人脸跟踪、人脸重建以及纹理渲染等，由于 CG 技术本身发展较为成熟，且引入了大量的人工干预手段，所以能够取得非常逼近于真实人脸的效果，是目前最完美的人脸伪造技术之一。但实现高度真实的效果往往需要投入大量的软硬件资源，如传感器、渲染设备以及制作团队等，其技术实现的代价也异常高昂。因此，目前利用 CG 技术进行人脸的合成与替换主要集中在一些商业应用中，如电影工业以及游戏制作行业。

2.3　深度生成模型

深度生成模型是以 Deepfake 为代表的人脸视频深度伪造技术的核心。深度生成模型可以简单描述成一个使用深度神经网络进行数据生成的模型，属于一种概率模型。使用生成模型可以生成出不包含在训练数据集中的新数据。目前深度学习主要包含两大类生成模型：自编码器(AutoEncoder，AE)[2-4]和生成对抗网络(Generative Adversarial Network，GAN)[5-7]。

其中 GAN 模型的训练包含了对抗博弈的过程，这个过程同时训练了两个模型：捕捉数据分布的生成器 G 以及判别器 D，D 用于评估输入的样本来源于真实数据分布还是生成数据分布。GAN 模型可以类比为一个造假团队，其中生成器生产假钞并在未经侦测的情况下使用，而判别器类似于警察，从真实数据中找出并区分这些假钞。整个博弈过程如同一场竞赛，迫使生成器与判别器的博弈对抗中不断提升双方性能，取得良好的视觉质量，直至真假数据无法被分辨。部分由最先进的生成模型 StyleGAN[17-18]。生成的人脸如图 3 所示，其生成的效果非常逼真，已经达到了人眼难以区分真假的程度。

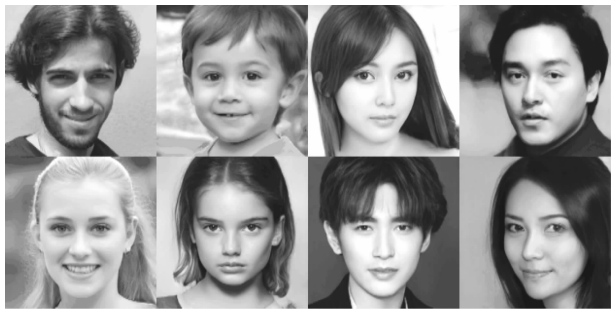


图 3　StyleGAN 模型生成的人脸示例

Fig. 3　Samples of human faces generated by StyleGAN

生成模型为人脸深度伪造奠定了技术基础，随着生成技术的不断完善，人脸伪造过程中所产生的假脸质量也在不断提升，这为人脸深度伪造的快速发展奠定了良好的基础。但这类生成技术也存在明显的缺点，由于其概率模型的特性，人脸的生成过程往往存在一定随机性，导致人脸伪造的过程不可控，且这类生成模型在生成连续视频帧时较难保持帧间一致性，从而使得合成的人脸视频存在较大的抖动，最终的伪造效果仍有待进一步提升。

综上所述，“换脸”技术的发展，实际上是学术化与工业化不断碰撞、螺旋上升的过程，经典的 CG 技术尽管能够取得高度的真实性，但其代价巨大且技术通常用于盈利，并不开源。而以 Deepfake 为代表的人脸深度伪造技术的实现门槛与代价都较低，但实际的效果仍有待提升。现阶段，深度学习技术已经被部分应用于电影工业中的 CG 特效，在未来，人脸的合成也将继续沿着这个方向发展，通过深度学习技术与部分 CG 技术的有机结合，在实现人脸伪造效果提升的同时也保持较低的技术门槛与高程度的开放性。

3　人脸视频深度伪造技术概述

3.1　基于图像域特征编码的方法

现阶段，全智能化的人脸深度伪造技术发展并

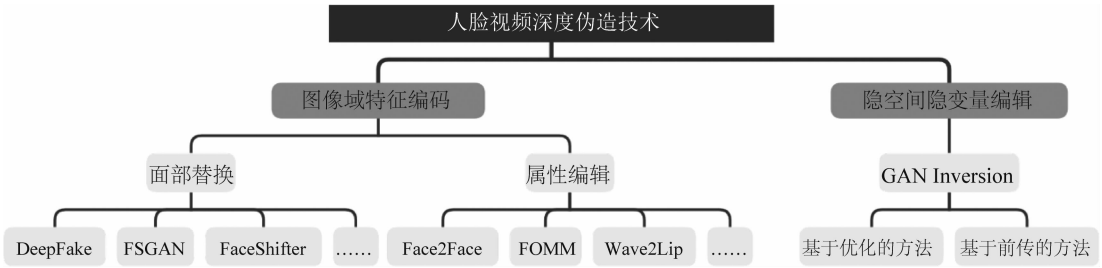


图 4 人脸视频深度伪造技术分类导图

Fig. 4 Classification map of Deepfake forgery technique

不完备,其中主流的伪造技术主要从人脸图像域的角度出发,通过对人脸图像进行特征编码、重构的操作方式实现篡改,根据篡改区域和篡改目标的不同,人脸视频深度伪造的篡改类型可以概括为面部替换和属性编辑两大类。其中面部替换旨在用原始人脸面部替换目标人脸的面部区域,涉及目标图像身份属性的变化。而属性编辑主要针对目标人脸身份信息外的各类属性进行编辑篡改,又包括表情迁移、面部重演、唇形篡改等具体篡改形式。图 4 给出了现有主要人脸视频深度伪造技术的分类引导示意图。

3.1.1 面部替换

面部替换是人脸视频深度伪造技术里最典型的一类算法,由于涉及身份属性的更改,因此面部替换可以实现让原始人物出现在他/她从来没有出现过的视频场景中。面部替换方法需要基于成对数据进行,即需要提供一组原始人脸和目标人脸作为训练数据进行伪造。

早期的伪造技术如 FaceSwap[19],采用的是经典图形学技术,首先获取人脸的若干个关键点,再利用传统的通用人脸三维模型 CANDIDE-3[20] 对人脸进行建模,并对获取到的人脸关键点位置进行渲染,通过缩小目标三维模型形状和关键点定位之间的差异,将渲染图像分解混合到整个人脸模型上,最后利用色彩校正技术进行后处理,得到替换后的人脸。延续类似的思路,许多基于 3D 建模的人脸替换方法被提出并获得发展[21-22]。但由于人脸通用三维模型顶点数目少,因此对人脸的轮廓和细节刻画不够细腻,导致最终的换脸结果不够理想,尽管后续方法部分采用了新式的人脸三维形变模型(3D morphable model,3DMM)[23-24],但这种直接渲染融合的换脸方式过于依赖原始人脸与目标人脸的形状相似度,当二者差距过大时,往往难以取得良好的融合效果。

深度学习的引入为成对人脸的面部替换带来了新的发展空间。如前所述,“Deepfake”算法[1] 是人脸深度伪造技术中的代表性算法,相关项目已在互联网开源。如图 5 所示,Deepfake 方法的主体结构基于自动编码器实现。在训练阶段,对于原始人脸 A 和目标人脸 B,训练一个权值共享的编码器,用于编码人脸特征,而在解码阶段,A 和 B 各自训练一个独立的解码器用于重构人脸。在测试阶段,为了实现 A 和 B 之间的人脸替换,先用训好的编码器对目标人脸 B 进行编码,此时,用训好的 A 的解码器来解码由 B 编码得到的特征,由此可以实现将 A 的面部解码到 B 的人脸肖像上。这种自动编码器的结构大大提升了面部替换的可操作性,同时也大大降低了换脸的技术门槛。但由于自动编码器结构上的限制,其对人脸特征的表征能力有限,因此在生成质量和生成的可控性方面都有待进一步提升。为了达到更好的替换效果和更佳的可操控性,对抗损失和人脸解耦重构等技术也被用于深伪算法进行约束与监督,并产生了很多变体方法,如 FSGAN[25]、FaceShifter[26] 等,使得生成的伪造人脸质量大幅提高。

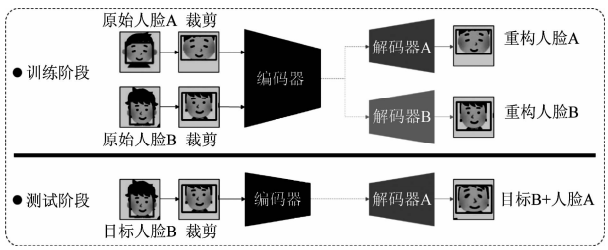


图 5 DeepFakes 基本原理框图

Fig. 5 Framework of DeepFakes

随着对人脸深度伪造理解的不断深入,更多提升面部替换逼真度的方法被提出。近期,微软亚洲

研究院的学者提出了 FaceShifter[26]，该方法将面部替换的效果提升到了一个新的高度。该方法共分为两阶段，第一阶段主要处理面部交换问题，第二阶段用于解决用于交换的人像中可能存在的遮挡问题。如图 6 所示，第一阶段中使用了两组独立的编码器，分别将人脸的身份信息和属性信息进行独立表征，这种表征方式有利于对人脸的身份和属性进行解耦并重组。在解码端，采用了基于自适应注意力机制的融合方法，将身份特征和属性特征在人脸不同的位置进行加权融合，在面部区域保留更多的原始人物身份信息，而在其他区域保留更多的目标人物属性信息，由此可以大大提升生成的真实感和生成过程的可控性。在第二阶段中，同样采用了类似 U-Net[27] 的网络结构，将目标图像与重建图像的差值作为遮挡部分区域的线索，联合第一阶段得到的生成图像，一同输入网络中，得到最终的恢复遮挡物体的生成图像。

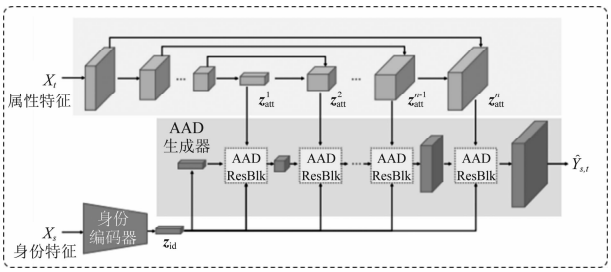


图 6　FaceShifter 面部替换方法基本原理框图

Fig. 6　Framework of FaceShifter

然而，之前的大多数面部替换算法主要只考虑了身份的改变，对于面部细节和光照分布以及整体的优化关注有限，这样当原始人物与目标人物的肤色、光照等属性差异较大时，生成的替换结果视觉质量上将较差。为了解决这一问题，中科院自动化所的研究者从最优化传输的角度提出了 AOT 算法[28]，该方法认为面部替换问题类似于一种最优传输(OT)问题，首先需要对原始图像和目标图像中的人脸肤色以及光照条件等进行建模，对于替换后但未进行优化前的人脸以及目标人脸，采用一个感知编码器来获取多尺度的特征，并对二者进行特征上的融合。这里的感知编码器使用了双分支的结构，一个分支用来提取高维特征，另一个分支用来计算坐标以及向量这些三维数据，以提供人脸的几何和光照信息。最终，在编码器的训练过程中最小化从原始分布到目标分布的代价，通过这一针对性的方法，能够有效改善现有面部替换方法光照、肤色不协调的情况，并且取得了良好的视觉质量。这一方法也凸显了在人脸视频深度伪造的过程中，后处理环节的重要性。

3.1.2　属性编辑

属性编辑是人脸深度伪造技术中另一类重要算法。该类算法以人脸属性为对象进行篡改，不涉及到目标人物身份信息的改变。通常，属性编辑可以改变视频人物的外观或动作表情特征，这类方法的输入可以是成对人脸视频，来实现目标人脸对原始人脸表情的模仿，也可以是单一的目标人脸加上某一指定的条件，将目标人脸的某种属性改变为指定的条件，如给定指定的语音内容，对目标人物的唇形进行篡改使得其口型配合上给定的语音内容等。

在深度学习技术成熟以前，可以通过消费级的 RGB-D 相机，追踪并重建原始人物和目标人物的人脸 3D 模型，再通过图像渲染技术进行最后融合，实现高逼真度的表情迁移。或基于 photoshop 等图像视频编辑软件通过手工的方式对图像或视频中目标人脸属性实现篡改。但这类方法所需要的专业化技术和成本较高，一定程度上制约了相关技术的应用。近期，深度生成模型被广泛应用到了人脸的属性编辑中。如表情迁移技术中的代表性算法 Face2Face[29]，该方法在 GitHub 上的开源版本利用了图像翻译模型实现原始人物到目标人物的表情迁移，其核心原理如图 7 所示。该算法使用成对的原始人脸和目标人脸作为输入，以五官的关键点作为表征，刻画并驱动不同表情的生成。在训练阶段，首先对目标人物的五官关键点进行提取，并以此作为图像翻译网络的输入，驱动网络训练学习并重构出具有相应表情的目标人脸。为了实现表情迁移，在模型的测试阶段，同样对待迁移的原始人脸提取五官关键点，以该关键点图形作为驱动输入到训练好的图像翻译模型中，重构出具有驱动表情的目标人脸，再根据需求合成视频。除了使用五官关键点作为驱动，部分后续工作如 NeuralTexture[30] 以及 Head2Head 等[31-32]也采用了人脸 3D 模型的渲染图像作为驱动，通过交换原始人脸和目标人脸的 3D 表情参数实现表情的迁移，并结合了时序一致性技术保障生成视频的连贯性。这类方法主要采用 AE、GAN 等生成结构，需要大量的原始人物和目标

人物的人脸素材,训练好的模型只能使用在特定人物上。

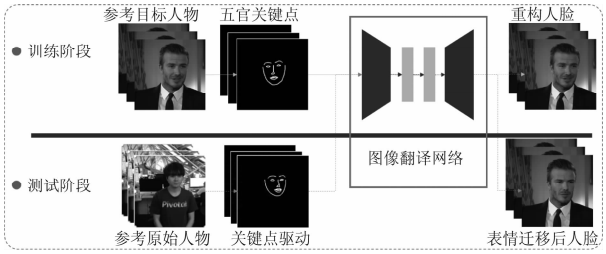


图 7 基于深度生成模型的 Face2Face 基本原理框图

Fig. 7 Framework of Deep generative model based Face2Face

除了这类特定人物之间的表情迁移,一项由视频驱动图片进行表情迁移方法 First Order Motion[33] 也获得了学术界的大量关注,在其基础上也发展出了一系列娱乐应用。该方法大体上来说分成两个模块,一个是运动估计模块,另一个是图像生成模块。其中运动估计模块的输出有两个,一个是密集运动场,它表征了驱动视频人物图像中的每个关键点到目标图像的映射关系;一个是混淆遮罩,该参数约束了驱动图像的不同区域,表明了哪部分可以通过扭曲得到,哪部分需要通过图像填充得到。而在图像生成模块中,输入目标人物图像,可以在编码器得到的特征层中进行形变,在解码回去,得到最终的输出。以该方法作为基础,已经发展出了 Avatarify 等互联网上非常流行的换脸软件。

近期,一些跨模态的唇形篡改方法如 Wave2Lip[34] 及类似方法[35-36] 等也逐渐成为研究热点。这类方法通常以驱动音频以及目标人物视频作为输入,首先对音频和目标人脸分别进行特征提取,再通过不同类型的联合共享特征空间的表征,配合不同视角的损失函数进行约束,最后进行音频和篡改后视频的同步,得到能够配合驱动音频的篡改唇形视频。

3.2 基于隐变量编辑的方法

在人脸伪造相关技术中,有一类方法通过编辑人脸图像的隐空间变量实现篡改,这类方法大多基于生成对抗网络(GAN)来实现。与基于图像域特征编码的方法不同,基于 GAN 隐空间实现人脸语义篡改的方法依赖于已训练好的 GAN 网络,探索人脸图像在隐空间中对应的隐变量,找到待篡改的语义方向,再利用预训练好的 GAN 生成器来生成编辑后的人脸。这类方法需要在图像隐空间进行操作,对于已知分布的生成人脸图像可以实现高自由度的编辑和高真实感的伪造,但对于真实人脸图像,首先需要将其映射到 GAN 的隐空间,此时非常依赖 GAN 反演(GAN Inversion)[37-38] 技术的效果。

GAN 反演的目的是将真实图像反转到预先训练好的 GAN 模型的隐空间中,如 StyleGAN[39-40],以便生成器准确地从被反演的隐编码中重建图像。它不仅提供了一个灵活的图像编辑框架,而且还有助于揭示深度生成模型的内在机制。作为连接预训练的 GAN 模型和实际图像篡改任务的新兴技术,GAN 反演对质量和效率有很高的要求。它主要可分为基于优化[17-18,41] 和基于前传的方法[42-44]。基于优化的方法通过反向传播直接对给定的单一图像进行隐编码优化。它能够产生高质量的反演,但优化过程过于耗时,因此大大限制了其实时应用。基于前传的方法使用编码器网络来学习从图像空间到隐空间的映射,在推理中只需要一次前传,为实时应用提供了更高的效率。

通过 GAN 反演方法获得了真实图像在隐空间中对应的隐编码后,便可以利用有监督[9] 或无监督[10] 的方法找到某个具体面部语义在隐空间中的方向,沿该方向调整隐编码,即可实现面部相应语义的篡改。其中,典型的有监督的代表方法是由香港中文大学汤晓鸥团队提出的 InterfaceGAN[45]。其基本原理如图 8 所示。针对某个待篡改属性,例如:眼镜、年龄、性别等,首先通过预训练的相应属性的预测模型对于从隐空间中随机采样的隐编码所对应的合成图像进行判决,经过判决后便可以对该隐编码给予确定的标签,即是否包含目标属性。接着,利用这些被给予标签的隐编码即可训练 SVM 分类器。对于给定的二分类属性,在隐空间中存在一个线性超平面可以很好地将隐编码分为两类。因此,我们便获得了该属性对应的语义方向。

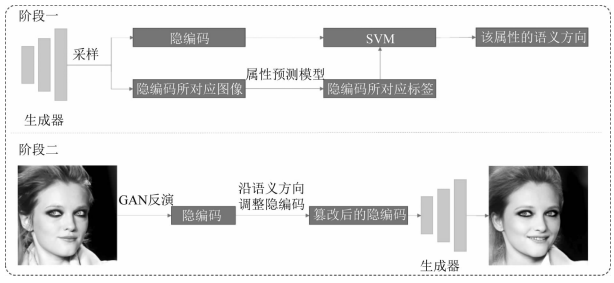


图 8 InterfaceGAN 基本原理图

Fig. 8 Framework of InterfaceGAN

有了相应属性的语义方向和通过 GAN 反演方法获得的真实图像对应的隐编码，我们可以通过令隐编码在该语义方向进行线性插值的方式篡改相应语义，之后将篡改后的隐编码送入生成对抗网络生成篡改后的图像。这类方法可探索性强，且思路多样化，近期在学术界得到了快速发展。

4　人脸视频伪造防御技术概述

在当前互联网流行的音视频媒体中，已经充斥了大量深度造假的内容，尽管这些伪造的媒体素材本身不具有明显的破坏能力，但其所涉及的虚假信息的快速传播，将可能引发严重的“蝴蝶效应”，可能会损害他人名誉、涉嫌伪造证据、传播谣言，影响政客形象甚至干涉政治选举等，这种潜在威胁激发了一系列对于 Deepfake 人脸视频防御技术[46-56]的研究。根据防御策略的不同，现有的防御技术可以大体划分为被动式检测和主动式防御两大类，其中被动式检测技术侧重于事后取证，即针对已经制作并传播的视频进行检测，判别其是否属于伪造人脸视频；另一类主动式防御技术侧重于事前防御，即在人脸数据发布传播前添加隐藏信息，如水印[57]、对抗噪声[58-59]等，进行主动溯源或使得恶意使用者无法利用添加了噪声的人脸视频进行伪造，从而达到保护人脸，实现主动防御的目的。图 9 给出了现有人脸伪造视频防御技术的分类导图。

4.1　被动式检测方法

现阶段大多数 Deepfake 人脸伪造视频的防御方法均从被动式检测的视角出发，这主要是由于视频来源的多样性决定。对于开放互联网环境下的人脸视频，其发布的来源以及发布方式具备多样性，难以做到在每次发布时都主动添加上干扰信息预防伪造。因此，当这些人脸素材被用于恶意伪造时，如何通过检测技术鉴别视频是否为伪造成为了防御技术的研究主流。

被动式检测技术指仅从人脸视频自身获取信息或提取特征，对伪造人脸视频进行鉴别的技术，这个任务本质上是一个二分类任务。二分类任务在多媒体取证[60-61]、计算机视觉等[62-63]领域都已有深入研究，但深度伪造的检测具有其特殊性，既不同于一般取证场景下像素级别修改条件下的检测，也不同于计算机视觉中语义级别的理解分类，而是与二者都有关联但又有实质差别的一种技术。现阶段 Deepfake 的被动式检测方法也大多从这两个领域中借鉴思路，根据检测方法的特点，可以概括为三类。

4.1.1　有伪造样本学习方法

这类方法的核心特点是利用真假成对数据作为训练的数据驱动，分类模型的学习过程需要有伪造人脸样本的参与。通过人工设计或神经网络挖掘的形式提取到真假人脸的可区分特征，从而训练分类器并进行分类。根据提取特征方式的不同，有伪造样本学习的方法可分为基于人工特征的方法和基于深度学习方法两种。其中基于人工特征的检测方法多见于早期阶段，针对伪造技术不完善导致的篡改痕迹进行特征提取，如利用传统图像取证中的隐写分析特征[64]，或捕捉眨眼异常、头面姿势不一致、异瞳和牙齿细节异常[65]等。图 10 给出了部分早期人脸伪造方法的伪像示意图。

而随着深度学习生成技术的发展，伪造人脸视频的质量越来越高，早期利用伪像鉴别真伪的方法难度逐渐提升，基于深度学习的方法成为主流。基于深度学习的方法更多地建立在深度神经网络强大的图像视频理解能力之上，从人脸媒体的空域、时域、频域等角度挖掘可鉴别的细节特征。

在基于深度学习方法中，根据信息提取视角的不同，有伪造样本学习方法又可细分为基于空域信息的方法、基于时域信息的方法、基于频域信息的方法、基于通用伪造痕迹的方法、基于注意力机制的方法、跨模态检测方法等等。

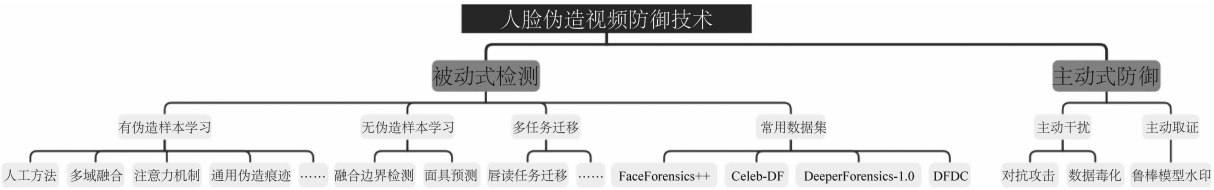
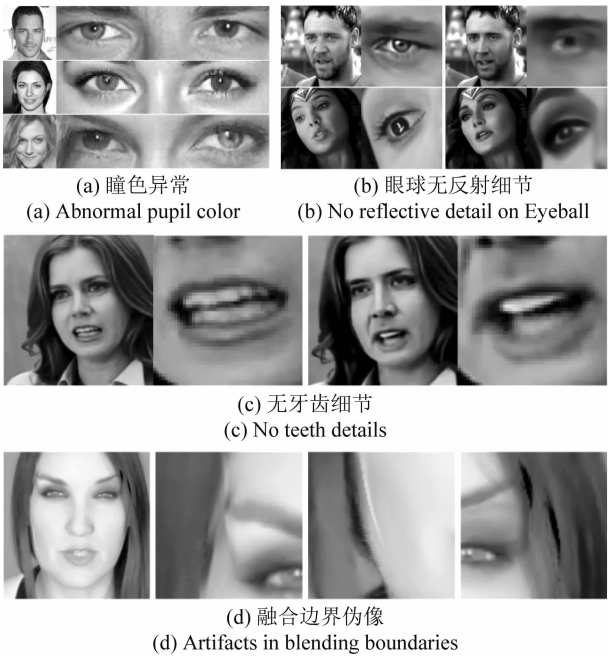


图 9　人脸伪造视频防御技术分类导图

Fig. 9　Classification map of Deepfake forgery defense technique



(a) 瞳色异常
(a) Abnormal pupil color

(b) 眼球无反射细节
(b) No reflective detail on Eyeball

(c) 无牙齿细节
(c) No teeth details

(d) 融合边界伪像
(d) Artifacts in blending boundaries

图 10 早期人脸伪造方法常见伪像示意图

Fig. 10 Common artifacts in early face forgery methods

1)基于空域的典型方法。一些经典的网络结构被直接应用于 Deepfake 检测,如 XceptionNet[47]、EfficientNet[66]等,这些网络结构直接从视频帧图像的空域提取特征信息,在 Deepfake 检测任务中发挥了良好的效果。但由于网络结构并非针对 Deepfake 检测任务进行特殊设计,且空域信息的提取能力有限,这类方法在不同数据集上所取得的性能上限受到了一定制约。

2)基于时域的典型方法。随着研究的深入,更多检测视角和机制被引入 Deepfake 检测,通过多个信息域特征的融合进行综合判决。典型的,Masi 等人提出的了 Two-branch[67]的检测方法,将时域信息进行了很好的利用,并与其他信息域的特征进行了融合。该方法构建了一个两分支的网络结构,分别采用了多尺度拉普拉斯算子和高斯运算符来提取视频帧图像 RGB 空间域和频域的信息,并通过一个双向的 LSTM[68]网络对输入的连续帧图像提取时域信息,最终实现空域、时域、频域信息的融合,进行综合判决检测。该方法由于充分利用了频域和时域信息,因此在人脸伪造视频检测的跨数据集迁移能力方面表现出了一定的优势,但其库内检测能力有待进一步提升。

3)基于频域信息的典型方法。在频域信息的利用上,Qian 等人做出了很大的突破,针对现有方法对高压缩率伪造视频的检测方面性能不佳的问题,Qian 等人提出了 F3-Net[69]方法。该方法精心设计了频域信息提取模块,使用频率感知分解和局部频率统计信息来放大视频帧图像的频域中由 Deepfake 伪造过程带来的伪像,该方法更多依赖于频域上的信息保留,因此对于高压缩率的伪造视频检测取得了良好的性能。但该方法也存在一定的局限性,即只能针对某一特定压缩率的视频取得良好的检测结果,当测试数据包含多种压缩率时,该方法需对每一种压缩率都进行重新训练,通用性不足。Chen 等人在 F3-Net 频域特征提取的形式上做了改进,提出了 MPSM[70]方法。该方法在网络骨干结构的不同层分别对空域特征和频域特征进行融合,并在最终融合的特征图上进行多尺度的块相似度计算,当某个块中存在修改痕迹时,该块中不同通道的融合特征相似度会出现明显异常,由此可以对比检测出待测人脸视频是否经过伪造。

4)基于通用伪造痕迹的典型方法。随着对 Deepfake 伪造技术理解的不断深入,研究者们不再将目光局限于从不同的信息域获取特征,而是希望通过新的视角来取得突破。Liu 等人近期发表了基于相位谱信息提取的空域浅学习方法 SPSL[71]。该方法不同于现有方法仅关注人脸的伪造过程,而是将视角转换,关注了整个伪造流程中,生成模型所必须经历若干次的上采样过程。该过程会在频域中留下可检测的伪像,这种伪像会在相位谱中更为显著,而多次上采样的累积过程会放大伪像。基于此,SPSL 从相位谱中捕捉频域伪造痕迹,由于上采样过程在不同生成模型中是通用的,因此 SPSL 实现了更为通用的检测和更好的跨库迁移性能。

5)基于注意力机制的典型方法。Zhao 等人最近提出了一种多注意力机制的方法 Multi-Attention[72],该方法首次将 Deepfake 检测定义为细粒度分类任务,属于领域内的一次创新。该方法有三大主要模块,即纹理增强模块、注意力生成模块和双线性注意力池化模块,其算法原理如图 11 所示。其中纹理增强模块用于增强骨干网络浅层提取到的纹理特征,作者认为浅层的纹理信息在 Deepfake 检测任务中能起到更为关键的作用。注意力生成模块用于生成多个注意力图,用于关注输入视频帧的不同位置,发掘并区分局部且细微的纹理伪像。而双线性注意力池化模块则用于对注意力图进行处理,

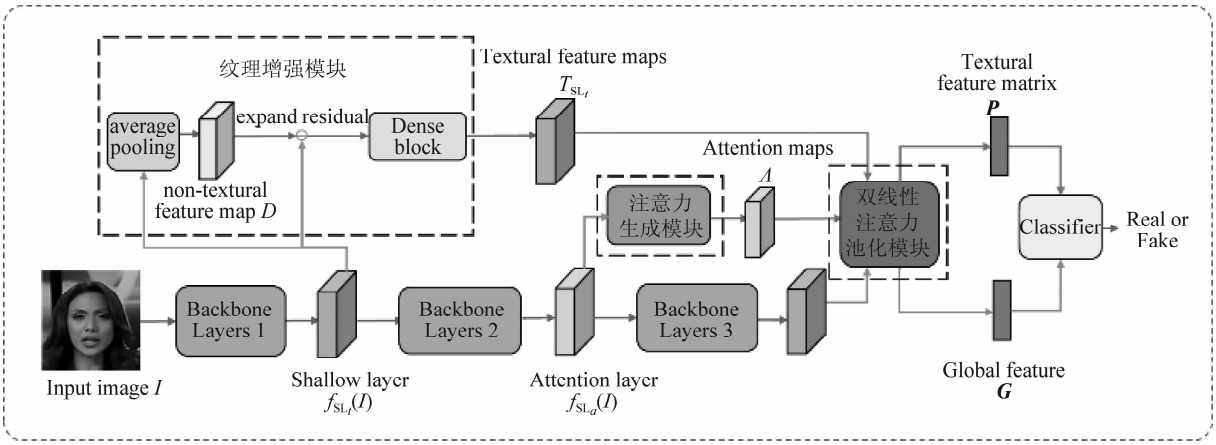


图 11　Multi-Attention 算法原理框图

Fig. 11　Framework of Multi-Attention Method

使每个注意力图的关注区域和响应位置不会重合，更加有效的利用多个注意力图的优势。该方法也是 DFDC 挑战赛中取得第二名的方法[73] 的扩展版本。

6）跨模态典型检测方法。除了从视频的空域、时域、频域等提取信息，部分学者也尝试从结合音频信息，从跨模态的视角来进行 Deepfake 检测。Lin 等人抓住了伪造人脸与真实人脸在发不同的声音时，对应的唇形上会产生不同这一现象，提出了基于唇形音素匹配的检测方法[74]，该方法受唇读任务的启发，建立了从人脸唇形到语音音素的 12 类鲁棒映射，根据划分好的音素将对应音素的真假唇形进行分类，得到 12 个唇形-音素匹配的子数据集。在每个子数据集中，利用分类好的真假唇形进行有伪造样本的配对训练，得到子分类器，并进行集成，最终获得性能优异的 Deepfake 检测模型。此外，Chugh 等人[75] 通过真伪视频中音频信号与视觉内容同步性上的差异进行伪造检测，也取得了一定效果。但这类跨模态的方法需要伪造视频中包含音频信息，而当前的主流数据集往往只包含视觉内容，只有少部分数据集包含音频内容，因此这类方法的发展也受到了一定制约。

除了上述总结的各类方法，也有学者尝试从生物信号的角度来解决 Deepfake 检测问题。Ciftci 等人[76] 参考了深度学习医疗信号处理中的心率分析方法，通过从伪造人脸视频图像的残差中提取生物特征，建立人脸纹理与心率之间的映射关系，在利用伪造人脸反应出的心率异常来实现 Deepfake 伪造视频的鉴别。该方法从生物信号的角度来对待 Deepfake 检测任务，为该领域提供了不一样的思路。最近，中国科学技术大学的 Dong 等人和微软亚洲研究院联合提出了一种通过引入额外的身份辅助信息进行 Deepfake 伪造视频检测的方法[77]。区别于以往的“伪像驱动”的思路，该方法基于“身份驱动”的思路，通过与找到待检测视频中的人物同身份的人脸作为身份属性参考，比对除脸部之外周围属性的差异，以此来鉴别待检测视频是否经过了换脸操作。类似的还有 Cozzolino 等人提出的 ID-Reveal 方法[78]。

4.1.2　无伪造样本学习方法

与“有伪造样本学习”的方法相对应，“无伪造样本学习”方法的模型训练过程不需要使用伪造人脸的负样本作为数据驱动，而是抓住了人脸这一特殊信息载体的某些特性，或抓住了深度伪造过程中某一固有的流程漏洞实现检测与鉴别。该类方法中最具代表性的算法是由微软亚洲研究院的 Li 等人提出的 Face X-ray[79]。该方法抓住了伪造方法大多需要进行融合操作这一固有流程，进行针对融合操作痕迹的检测，其原理示意图如图 12 所示。由于只关注融合边界的痕迹，因此检测模型的训练只需要使用真脸数据和部分人工制作的融合图像，这里训练所需的负样本并不是伪造的人脸图像，而只需要将不同图像进行拼接并融合进行构造，所以训练过程不需成对的真假脸，因此称为“无伪造样本学习”的方法。

另一个具有代表性的算法是 Zhao 等人提出逐块一致性学习的方法 PCL[80]，该方法通过在待检测人脸视频帧图像上测量其与源人脸图像特征的一致

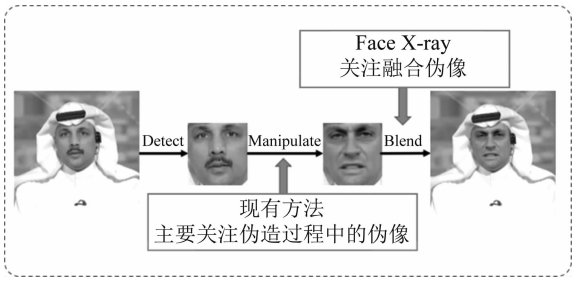


图 12 Face X-ray 算法原理示意图

Fig. 12 Framework of Face X-ray

性来进行学习,通过一致性程度来判断当前块内是否存在融合的痕迹,从而预测出类似融合操作的遮罩 mask,当待检测人脸中存在大面积 mask 区域,则可认为是高概率的伪造人脸。这类无伪造样本学习的方法由于不依赖于成对真假人脸数据,因此对不同的伪造算法具有较强的迁移检测能力,其跨库检测的性能也基本处于领域内领先的地位。

4.1.3 基于多任务迁移的方法

多任务迁移的方法本质上是利用其他取证或视觉任务中已有的方法进行迁移改造,应用到 Deepfake 人脸伪造视频的检测任务中。这是最近新兴起的一类方法,其核心分类器依旧是依赖于成对真假数据进行训练,与前述方法最大的不同在于,该类方法会利用其他任务中的预训练模型或预筛选特征,在 Deepfake 的真假数据对上进行微调。这类方法中,Haliassos 等人提出的 Lip Forensics[81] 是当前最具代表性也是最先进的算法,其方法原理如图 13 所示。该方法将唇读任务中的预训练模型迁移到了 Deepfake 伪造人脸视频的检测任务中。利用了针对嘴部动作中的高级语义不规则性,以及现有伪造技术在唇形生成方面的弱点,区分真假唇形,进而鉴别伪造人脸视频。其特征提取过程与唇读任务一致,首先对在大规模唇读数据集上对时空融合的骨干网络进行预训练,学习与自然嘴部运动有关的丰富表达形式,再固定网络中提取唇形特征的这部分,在 Deepfake 伪造数据集上对最后的分类模块进行微调,最终形成可区分真伪人脸视频的检测模型。该方法由于在大规模的唇读任务数据集上进行了预训练,因此在 Deepfake 检测任务中,网络不会过拟合到少量的伪造人脸数据唇形上,所以在库内和跨库迁移性方面都表现出了非常优秀的性能。

这类基于多任务迁移的方法,其原始的预训练模型经过了其他任务中的特有大规模数据集预训练,再在 Deepfake 检测专用数据集上微调,其模型相比于直接训练与 Deepfake 检测数据集的方法要具有更好的泛化能力与鲁棒性,这也预示着这类方法在 Deepfake 伪造人脸视频检测任务中具有广阔的发展前景。类似的,如表情识别[82]、身份识别[83]等具有良好研究基础的任务中的方法,也可遵循这类方法的基本设计原则,迁移到 Deepfake 检测任务中,理论上也能够获得良好的检测性能。

4.1.4 常用数据集

Deepfake 数据集主要用来训练以及评估检测模型的性能。Deepfake 发展至今,出现过很多数据集如 UADFV[65]、Deepfake-TIMIT[84] 等等,这些数据集大多发布于研究早期,其规模和伪造效果均有一定局限性。随着 Deepfake 伪造技术的发展,更大规模和更高质量的数据集不断被发布,目前最常用的数据集主要有四个,各数据集的基本情况如表 1 所示。其中衡量算法的库内检测性能通常用 FaceForensics++[85],而衡量算法跨库性能时则多使用 Celeb-DF[86] 测试。商汤科技对伪造效果进行了进一步优化,并发布了

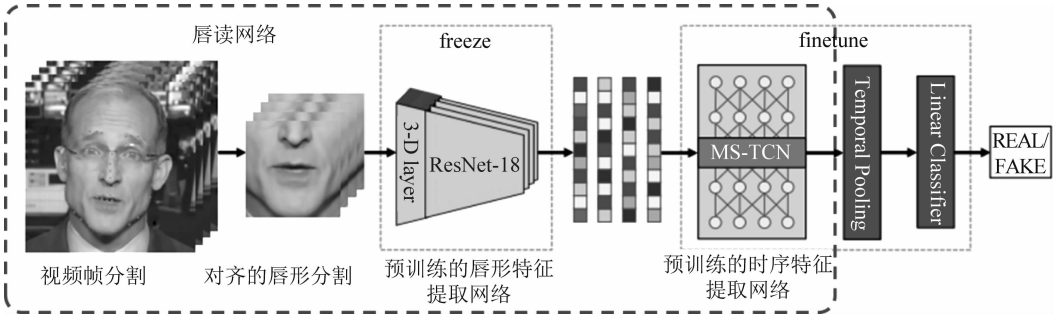


图 13 基于唇读任务的 Deepfake 人脸伪造检测算法框图

Fig. 13 Framework of Lip-Forensics

表 1　Deepfake 检测主要数据集基本情况

Tab. 1　Overview of mainstream Deepfake detection datasets

数据集名称	数据规模	数据来源	特点
FaceForensics++[85]	1000 real,4000 fake	YouTube	早期包含 4 种伪造方法(现新增 FaceShifter),包含 3 种分辨率,整体质量偏低,有明显伪像
Celeb-DF[86]	590 real, 5639 fake	YouTube	基于 Deepfake 方法进行优化,视频视觉质量较高,伪造方法与数据类型单一
DeeperForensics-1.0[87]	1000 real, 11000 fake	YouTube+ Actors	增加了数据规模和大量亚洲人种数据,合成方法更好,但真实人脸源自 FF++,易造成数据泄露问题
DFDC[73]	23654 real, 104500 fake	Actors	全真实场景模拟,全真实演员拍摄,迄今为止规模最大的 Deepfake 数据集,8 种伪造方法,19 种干扰手段,难度极大。但部分数据有异常

具有更多人种和效果的 DeeperForensics-1.0[87]。Facebook 为了配合 DFDC 深度伪造检测挑战赛的开展,融合了 8 种伪造方法、19 种干扰手段,发布了迄今为止最大规模且最接近真实场景的音视频融合数据集 DFDC[73],这一数据集目前仍旧是的 Deepfake 检测领域最具挑战性的数据集。

随着领域内技术的进步和研究视角的转变,不同类型的数据集还将继续产生和发展,如近期中国科学技术大学和微软亚研院联合组建的包含人物身份参考的特定人伪造鉴别数据集“Vox-Deepfake”[77],Zi 等人提出的真实场景下的数据集“WildDeepfake”[88],He 等人采用了更多伪造策略构建的大型人脸图像视频伪造数据集[89]以及 Kwon 等人提出的针对韩国人的大型数据集[90]等。

在上述数据集上,本文针对当前最主流和性能表现处于第一梯队的部分 Deepfake 检测算法进行了检测性能的对比,所有数据均由论文实验整理获得,部分细节设置可能不完全统一,因此实验数据仅可用作性能对比的参考。主要评估指标有准确率(Acc)、ROC 曲线面积(AUC),HQ 和 LQ 分别代表低压缩率(c23)和高压缩率(c40)。评测结果如表 2 所示。由表中结果可知,各类采用不同思路的方法均在某一方面具备一定的优势,但 Deepfake 人脸伪造检测的任务,依旧是一个尚未完善解决的任务。

表 2　主流实验设置下,代表性方法在主要 Deepfake 检测数据集上的库内性能和跨库性能评测结果(%)

Tab. 2　In-dataset and cross-dataset performances of typical Deepfake detection methods under mainstream experimental settings (%)

方法	主要特点	测试集:FF++ HQ(训练集:FF++ HQ)		测试集:FF++ LQ(训练集:FF++ LQ)		测试集:Celeb-DF(训练集:FF++ HQ)
		ACC	AUC	ACC	AUC	跨库测试 AUC
DSP-FWA[55]	将真实人脸以不同分辨率模糊处理后贴回,生成训练集	–	56.89	–	59.15	64.60
Xception[85]	FF++论文中效果最好的 baseline	95.73	96.30	86.86	89.30	65.50
SSTNET[91]	双流网络,Xception+约束卷积的浅层 Xception+RNN	98.57	–	90.11	–	–
Two-Branch[67]	双流网络,独立分支对特征图提取高频信息+双向 LSTM	–	98.7	–	86.59	73.41
F3-Net[69]	基于频域图像分解和局部频域特征统计的双流网络	97.52	98.1	90.43	93.3	65.17
Face X-ray[79]	仅关注假脸的边缘融合区域	–	87.40	–	61.60	80.58
PCL[80]	检测区块特征的不一致性	–	–	–	–	81.80
Multi-Attention[72]	细粒度分类思想,多注意力机制	97.60	99.29	88.69	90.40	67.44
LRNet[92]	利用对 landmark 的预测和校准进行特征增强与检测	–	97.30	–	95.70	56.90
MPSM[70]	对特征图进行多尺度的块相似度计算	97.59	99.46	91.47	95.21	78.26
M2TR[93]	多尺度 transformer 作 backbone	98.23	99.48	92.35	94.22	65.70
SPSL[71]	检测伪造的上采样痕迹	91.50	95.32	81.57	82.82	76.88
LipForensics[81]	利用唇读任务进行迁移检测	98.80	99.70	94.20	98.10	82.40

4.2 主动式防御方法

如前所述,目前最主流的 Deepfake 伪造防御技术大多通过事后取证的方式进行,属于被动式检测方法,但此时往往伪造人脸多媒体的危害已经形成。因此,近期一些主动式防御的工作被提出。相比于被动式检测方法,主动式防御的核心思路是在人脸视频发布前便添加一定程度的信号干扰,使得恶意使用者无法利用添加了干扰后的人脸素材进行伪造,或即便伪造也能够顺利溯源,找到伪造者,以实现“事前防御”的目标。这类技术理论上能够从源头杜绝 Deepfake 视频的产生,但相关技术的执行条件较为苛刻,需要人脸素材的发布方主动添加干扰信息,或要求人脸素材发布的场景可控,这在绝大多数素材自由分享的互联网平台上是难以实现的。但这种技术对于特殊需求者而言是具有高度使用价值的,如部分媒体希望自己发布的素材不被恶意使用,在这种场景下可以通过主动防御技术对自己的人脸素材进行保护后再发布,实现对 Deepfake 伪造的预防。

主动防御技术也可归纳为主动干扰和主动取证两类,主动干扰指的是对发布的人脸素材添加噪声干扰,如对抗攻击或数据毒化,使恶意使用者无法对该人脸进行伪造。如图 14 所示,在主动干扰方面,Huang 等人[94]提出了一种基于对抗攻击和数据毒化的人脸主动防御方法,来直接干扰人脸篡改模型的生成过程,来实现对人脸数据的主动保护。他们设计了一种二阶段的训练框架,凭借交替训练的策略,来训练一个针对于特定人脸属性编辑或表情迁移任务的加扰网络。使用者只需要用这一加扰网络对待保护的人脸数据进行处理,就可以让恶意伪造者在对这些加扰人脸数据进行篡改伪造时只能得到很糟糕的结果。各项实验都表明了该主动防御框架的有效性和在不同黑盒对抗环境下的鲁棒性。

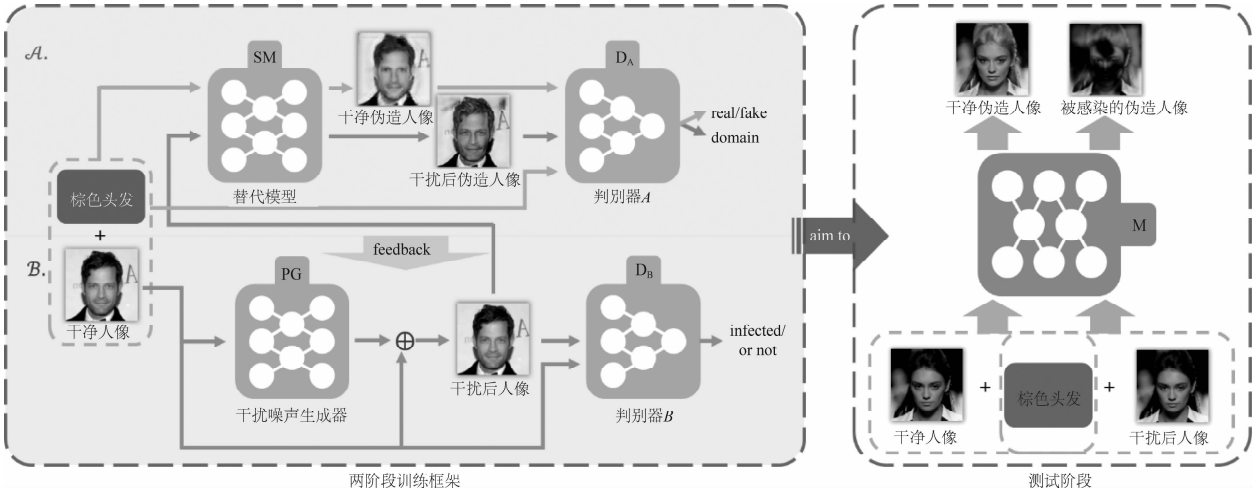


图 14 主动干扰方法原理框图

Fig. 14 Framework of initiative interference

在主动取证技术方面,Yu 等人[95]通过在模型中引入人工指纹,寻求一种对生成模型的演化不可知的、主动的、可持续的深度伪造检测方法。该方法首先将人工指纹嵌入到训练数据中,然后验证了其从训练数据到生成模型的可转移性,进而出现在生成的伪造结果中。此外,Kim 等人[96]提出了一个分散的属性方案,通过引入用户专属的参数化密钥,并通过该密钥来引导用户端模型的水印与无水印数据进行区分,使得人脸素材具有用户特定的属性,以此来实现伪造后的溯源追踪。

5 结论

Deepfake 人脸视频的深度伪造及其防御是人工智能安全领域中的热点问题,由于其在安全与伦理问题上的特殊性,受到了科研、政治、社会、经济等多个领域的广泛关注。这个领域的研究目前仍处于快速更新与进展期,需要不断地对现有研究进行阶段式总结,从而更好地明确未来的研究方向。而解决 Deepfake 相关问题,绝不仅仅是技术上的问题,更应该从立法监管、新闻宣传、知识科普等多个角度协同攻坚。

人脸视频的深度伪造及其防御技术的发展依旧几个关键的研究难点：

（1）伪造的智能化与真实感的权衡。智能化的伪造技术带来的是便捷性和自由度，但由于场景需求不同，统一的伪造框架无法很好的适应需求，因此结合人工干预的后处理手段可能是深度伪造更高级的形态。

（2）多维度协同防御。事后取证的检测技术难以抵消已经发生的危害，因此事前保护的主动防御技术是值得期待的发展方向。目前相关的研究较少，且由于数据来源不可控，这类方法的应用受到了极大制约。

（3）真实场景防御问题。当前 Deepfake 被动式检测技术的一大难点就是真实伪造场景下的鲁棒检测方法。由于传播场景复杂，对抗因素多，现有技术的泛化能力尚未达到实用，这一点也被 2020 年 Facebook 牵头举办的 DFDC 比赛结论所验证。为了解决这一难题，发展多样化且高度真实的数据集和伪造方法也是一大关键问题。

参考文献

[1] Deepfakes. Deepfakes github [EB/OL]. http://github.com/Deepfakes/faceswap, 2017. Accessed 2020-08-18. 2, 3, 5, 6.

[2] KINGMA D P, WELLING M. Auto-encoding variational Bayes [EB/OL]. https://arxiv.org/abs/1312.6114, arXiv preprint arXiv:1312.6114, 2013.

[3] MAKHZANI A, SHLENS J, JAITLY N, et al. Adversarial autoencoders [EB/OL]. https://arxiv.org/abs/1511.05644, arXiv preprint arXiv:1511.05644, 2015.

[4] DOERSCH C. Tutorial on variational autoencoders [EB/OL]. https://arxiv.org/abs/1606.05908, arXiv preprint arXiv:1606.05908, 2016.

[5] MIRZA M, OSINDERO S. Conditional generative adversarial nets[EB/OL]. https://arxiv.org/abs/1411.1784, arXiv preprint arXiv:1411.1784, 2014.

[6] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[EB/OL]. https://arxiv.org/abs/1511.06434, arXiv preprint arXiv:1511.06434, 2015.

[7] GOODFELLOW I, BENGIO Y, COURVILLE A, et al. Deep learning[M]. Cambridge: MIT press, 2016.

[8] Zao app. Retrieved from https://zao-app.com/ [EB/OL], 2019-12-01.

[9] QuYan app. Retrieved from https://www.quyan.ai/ [EB/OL], 2020-09-18.

[10] Facebook, Kaggle Deepfake detection challenge [EB/OL]. Retrieved from https://www.kaggle.com/c/Deepfake-detection-challenge, 2020-04-01.

[11] KORSHUNOV P, EBRAHIMI T. Using face morphing to protect privacy[C]//2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance. Krakow, Poland. IEEE, 2013: 208-213.

[12] RAGHAVENDRA R, RAJA K, VENKATESH S, et al. Face morphing versus face averaging: Vulnerability and detection[C]//2017 IEEE International Joint Conference on Biometrics (IJCB). Denver, CO, USA. IEEE, 2017: 555-563.

[13] FOLEY J D, VAN Dam A, FEINER S K, et al. Introduction to computer graphics[M]. Reading: Addison-Wesley, 1994.

[14] FOLEY J D, VAN F D, VAN Dam A, et al. Computer graphics: principles and practice[M]. Addison-Wesley Professional, 1996.

[15] ZHONG Shi, GHOSH J. Generative model-based document clustering: A comparative study[J]. Knowledge and Information Systems, 2005, 8(3): 374-384.

[16] BAU D, LIU S, WANG Tongzhou, et al. Rewriting a deep generative model[M]// Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 351-369.

[17] ABDAL R, QIN Yipeng, WONKA P. Image2StyleGAN: how to embed images into the StyleGAN latent space?[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South). IEEE, 2019: 4431-4440.

[18] ABDAL R, QIN Yipeng, WONKA P. Image2StyleGAN++: how to edit the embedded images?[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA. IEEE, 2020: 8293-8302.

[19] FaceSwap. Faceswap github[EB/OL]. https://github.com/MarekKowalski/FaceSwap, 2016. Accessed 2020-08-18. 2, 3, 5, 6.

[20] AHLBERG J. CANDIDE-3—An updated parameterized face[EB/OL]. http://citeseerx.ist.psu.edu/viewdoc/summary? doi=10.1.1.33.5603, 2001.

[21] AVERBUCH-ELOR H, COHEN-OR D, KOPF J, et al. Bringing portraits to life[J]. ACM Transactions on

Graphics, 2017, 36(6): 1-13.

[22] SUWAJANAKORN S, SEITZ S M, KEMELMACHER-SHLIZERMAN I. Synthesizing obama[J]. ACM Transactions on Graphics, 2017, 36(4): 1-13.

[23] DENG Yu, YANG Jiaolong, XU Sicheng, et al. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, CA, USA. IEEE, 2019: 285-295.

[24] CHANG Fengju, TRAN A T, HASSNER T, et al. Deep, landmark-free FAME: Face alignment, modeling, and expression estimation[J]. International Journal of Computer Vision, 2019, 127(6/7): 930-956.

[25] NIRKIN Y, KELLER Y, HASSNER T. FSGAN: subject agnostic face swapping and reenactment[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South). IEEE, 2019: 7183-7192.

[26] LI Lingzhi, BAO Jianmin, YANG Hao, et al. Advancing high fidelity identity swapping for forgery detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA. IEEE, 2020: 5073-5082.

[27] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation [C]//Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015, 2015: 234-241. DOI:10.1007/978-3-319-24574-4_28.

[28] ZHU H, FU C, WU Q, et al. AOT: Appearance Optimal Transport Based Identity Swapping for Forgery Detection [EB/OL]. https://arxiv.org/abs/2011.02674, arXiv preprint arXiv:2011.02674, 2020.

[29] THIES J, ZOLLHÖFER M, STAMMINGER M, et al. Face2Face: real-time face capture and reenactment of RGB videos[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. IEEE, 2016: 2387-2395.

[30] THIES J, ZOLLHÖFER M, NIEßNER M. Deferred neural rendering: Image synthesis using neural textures[J]. ACM Transactions on Graphics (TOG), 2019, 38(4): 1-12.

[31] DOUKAS M C, KOUJAN M R, SHARMANSKA V, et al. Head2Head: deep facial attributes Re-targeting[J]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2021, 3(1): 31-43.

[32] DOUKAS M C, KOUJAN M R, SHARMANSKA V, et al. Head2HeadFS: video-based head reenactment with few-shot learning[EB/OL]. https://arxiv.org/abs/2103.16229, arXiv preprint arXiv:2103.16229, 2021.

[33] SIAROHIN A, LATHUILIÈRE S, TULYAKOV S, et al. First order motion model for image animation[J]. Advances in Neural Information Processing Systems, 2019, 32: 7137-7147.

[34] PRAJWAL K R, MUKHOPADHYAY R, NAMBOODIRI V P, et al. A lip sync expert is all You need for speech to lip generation in the wild[C]//Proceedings of the 28th ACM International Conference on Multimedia. Seattle WA USA. New York, NY, USA: ACM, 2020: 484-492.

[35] YI Ran, YE Zipeng, ZHANG Juyong, et al. Audio-driven talking face video generation with learning-based personalized head pose[EB/OL]. https://arxiv.org/abs/2002.10137, arXiv e-prints, 2020: arXiv: 2002.10137.

[36] SONG Linsen, WU W, FU Chaoyou, et al. Everything's talkin': Pareidolia face reenactment[EB/OL]. https://arxiv.org/abs/2104.03061, arXiv preprint arXiv:2104.03061, 2021.

[37] XIA Weihao, ZHANG Yulun, YANG Yujiu, et al. GAN inversion: A survey[EB/OL]. https://arxiv.org/abs/2101.05278, arXiv preprint arXiv:2101.05278, 2021.

[38] ZHU Jiapeng, SHEN Yujun, ZHAO Deli, et al. In-domain GAN inversion for real image editing[M]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 592-608.

[39] KARRAS T, LAINE S, AILA Timo. A style-based generator architecture for generative adversarial networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA. IEEE, 2019: 4396-4405.

[40] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of StyleGAN[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA. IEEE, 2020: 8107-8116.

[41] GU Jinjin, SHEN Yujun, ZHOU Bolei. Image processing using multi-code GAN prior[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA. IEEE, 2020: 3009-3018.

[42] GUAN Shanyan, TAI Ying, NI Bingbing, et al. Collaborative learning for faster StyleGAN embedding[EB/OL]. https://arxiv.org/abs/2007.01758, arXiv preprint arXiv:2007.01758, 2020.

[43] NITZAN Y, BERMANO A, LI Yangyan, et al. Face identity disentanglement via latent space mapping[J]. ACM Transactions on Graphics, 2020, 39(6): 1-14.

[44] RICHARDSON E, ALALUF Y, PATASHNIK O, et al. Encoding in style: A StyleGAN encoder for image-to-image translation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 2287-2296.

[45] SHEN Yujun, GU Jinjin, TANG Xiaoou, et al. Interpreting the latent space of GANs for semantic face editing[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA. IEEE, 2020: 9240-9249.

[46] AFCHAR D, NOZICK V, YAMAGISHI J, et al. MesoNet: a compact facial video forgery detection network[C]//2018 IEEE International Workshop on Information Forensics and Security (WIFS). Hong Kong, China. IEEE, 2018: 1-7.

[47] CHOLLET F. Xception: deep learning with depthwise separable convolutions[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. IEEE, 2017: 1800-1807.

[48] CIFTCI U A, DEMIR I, YIN L. Fakecatcher: Detection of synthetic portrait videos using biological signals[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[49] NGUYEN H H, YAMAGISHI J, ECHIZEN I. Capsule-forensics: Using capsule networks to detect forged images and videos[C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK. IEEE, 2019: 2307-2311.

[50] DANG Hao, LIU Feng, STEHOUWER J, et al. On the detection of digital face manipulation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA. IEEE, 2020: 5780-5789.

[51] YANG Xin, LI Yuezun, LYU Siwei. Exposing deep fakes using inconsistent head poses[C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK. IEEE, 2019: 8261-8265.

[52] AGARWAL S, FARID H, FRIED O, et al. Detecting deep-fake videos from phoneme-viseme mismatches[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, WA, USA. IEEE, 2020: 2814-2822.

[53] AGARWAL S, FARID H, EL-GAALY T, et al. Detecting deep-fake videos from appearance and behavior[C]//2020 IEEE International Workshop on Information Forensics and Security (WIFS). New York, NY, USA. IEEE, 2020: 1-6.

[54] FEI Jianwei, XIA Zhihua, YU Peipeng, et al. Exposing AI-generated videos with motion magnification[J]. Multimedia Tools and Applications, 2020: 1-14.

[55] LI Yuezun, LYU Siwei. Exposing Deepfake videos by detecting face warping artifacts[EB/OL]. https://openaccess.thecvf.com/content_CVPRW_2019/papers/Media%20Forensics/Li_Exposing_DeepFake_Videos_By_Detecting_Face_Warping_Artifacts_CVPRW_2019_paper.pdf, arXiv preprint arXiv:1811.00656, 2018.

[56] NGUYEN H H, FANG Fuming, YAMAGISHI J, et al. Multi-task learning for detecting and segmenting manipulated facial images and videos[C]//2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS). Tampa, FL, USA. IEEE, 2019: 1-8.

[57] COX I J, MILLER M L, BLOOM J A, et al. Steganography[M]//Digital Watermarking and Steganography. Amsterdam: Elsevier, 2008: 425-467.

[58] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB/OL]. https://arxiv.org/abs/1312.6199, arXiv preprint arXiv: 1312.6199, 2013.

[59] AKHTAR N, MIAN A. Threat of adversarial attacks on deep learning in computer vision: A survey[J]. IEEE Access, 2018, 6: 14410-14430.

[60] BÖHME R, FREILING F C, GLOE T, et al. Multimedia forensics is not computer forensics[M]//Computational Forensics. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009: 90-103.

[61] PEVNÝ T, FRIDRICH J. Benchmarking for steganography[M]//Information Hiding. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008: 251-267.

[62] FORSYTH D A, PONCE J. Computer vision: A modern approach[M]. Cambridge: Pearson, 2012.

[63] SZELISKI R. Computer vision: algorithms and applications[M]. Springer Science & Business Media, 2010.

[64] ZHOU Peng, HAN Xintong, MORARIU V I, et al. Two-stream neural networks for tampered face detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, HI, USA. IEEE, 2017: 1831-1839.

[65] MATERN F, RIESS C, STAMMINGER M. Exploiting visual artifacts to expose Deepfakes and face manipulations[C]//2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). Waikoloa, HI, USA. IEEE, 2019: 83-92.

[66] TAN Mingxing, LE Q V. EfficientNet: rethinking model scaling for convolutional neural networks [C]// International Conference on Machine Learning. PMLR, 2019: 6105-6114.

[67] MASI I, KILLEKAR A, MASCARENHAS R M, et al. Two-branch recurrent network for isolating Deepfakes in videos[M]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 667-684.

[68] JOZEFOWICZ R, ZAREMBA W, SUTSKEVER I. An empirical exploration of recurrent network architectures [C]// International Conference on Machine Learning. PMLR, 2015: 2342-2350.

[69] QIAN Yuyang, YIN Guojun, SHENG Lu, et al. Thinking in frequency: Face forgery detection by mining frequency-aware clues[C]//Computer Vision - ECCV 2020, 2020: 86-103. DOI:10.1007/978-3-030-58610-2_6.

[70] CHEN S, YAO T, CHEN Y, et al. Local eelation learning for face forgery detection[EB/OL]. https://www.aaai.org/AAAI21Papers/AAAI-1964.ChenS.pdf, arXiv preprint arXiv:2105.02577, 2021.

[71] LIU Honggu, LI Xiaodan, ZHOU Wenbo, et al. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 772-781.

[72] ZHAO Hanqing, ZHOU Wenbo, CHEN Dongdong, et al. Multi-attentional Deepfake detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 2185-2194.

[73] DOLHANSKY B, BITTON J, PFLAUM B, et al. The Deepfake detection challenge dataset[EB/OL]. https://ui.adsabs.harvard.edu/abs/2020arXiv200607397D/abstract, arXiv preprint arXiv:2006.07397, 2020.

[74] LIN Jiaying, ZHOU Wenbo, LIU Honggu, et al. Lip Forgery Detection via Multi-phoneme selection[C]//Proceedings of International Workshop on Safety & Security of Deep Learning, 2021.

[75] CHUGH K, GUPTA P, DHALL A, et al. Not made for each other-audio-visual dissonance-based deepfake detection and localization[C]//Proceedings of the 28th ACM International Conference on Multimedia, 2020: 439-447.

[76] CIFTCI U A, DEMIRİ, YIN Lijun. How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals[C]//2020 IEEE International Joint Conference on Biometrics (IJCB). Houston, TX, USA. IEEE, 2020: 1-10.

[77] DONG Xiaoyi, BAO Jianmin, CHEN Dongdong, et al. Identity-driven Deepfake detection[EB/OL]. https://arxiv.org/abs/2012.03930, arXiv preprint arXiv:2012.03930, 2020.

[78] COZZOLINO D, ROSSLER A, THIES J, et al. Id-reveal: Identity-aware deepfake video detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 15108-15117.

[79] LI Lingzhi, BAO Jianmin, ZHANG Ting, et al. Face X-ray for more general face forgery detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA. IEEE, 2020: 5000-5009.

[80] ZHAO Tianchen, XU Xiang, XU Mingze, et al. Learning to recognize patch-wise consistency for Deepfake detection [EB/OL]. https://arxiv.org/abs/2012.09311, arXiv preprint arXiv:2012.09311, 2020.

[81] HALIASSOS A, VOUGIOUKAS K, PETRIDIS S, et al. Lips don't lie: A generalisable and robust approach to face forgery detection [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 5039-5049.

[82] LOPES A T, DE AGUIAR E, DE SOUZA A F, et al. Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order[J]. Pattern Recognition, 2017, 61: 610-628.

[83] DONG Xunde, SI Wenjie, HUANG Weiyi. ECG-based identity recognition via deterministic learning[J]. Biotechnology & Biotechnological Equipment, 2018, 32(3): 769-777.

[84] KORSHUNOV P, MARCEL S. Deepfakes: a new threat to face recognition? assessment and detection[EB/OL]. https://arxiv.org/abs/1812.08685, arXiv preprint arXiv:1812.08685, 2018.

[85] RÖSSLER 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.

[86] LI Yuezun, YANG Xin, SUN Pu, et al. Celeb-DF: A large-scale challenging dataset for Deepfake forensics[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA. IEEE, 2020: 3204-3213.

[87] JIANG Liming, LI Ren, WU W, et al. DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA. IEEE, 2020: 2886-2895.

[88] ZI Bojia, CHANG Minghao, CHEN Jingjing, et al. WildDeepfake: A challenging real-world dataset for Deepfake detection[C]// Proceedings of the 28th ACM International Conference on Multimedia. Seattle WA USA. New York, NY, USA: ACM, 2020: 2382-2390.

[89] HE Yinan, GAN Bei, CHEN Siyu, et al. ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 4360-4369.

[90] KWON P, YOU J, NAM G, et al. KoDF: A large-scale korean DeepFake detection dataset[J]. arXiv preprint arXiv:2103.10094, 2021.

[91] WU Xi, XIE Zhen, GAO Yutao, et al. SSTNet: detecting manipulated faces through spatial, steganalysis and temporal features[C]// ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain. IEEE, 2020: 2952-2956.

[92] SUN Zekun, HAN Yujie, HUA Zeyu, et al. Improving the efficiency and robustness of Deepfakes detection through precise geometric features [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 3609-3618.

[93] WANG Junke, WU Zuxuan, CHEN Jingjing, et al. M2TR: multi-modal multi-scale transformers for Deepfake detection [EB/OL]. https://arxiv.org/abs/2104.09770, arXiv preprint arXiv:2104.09770, 2021.

[94] HUANG Q, ZHANG J, ZHOU W, et al. Initiative Defense against Facial Manipulation[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35 (2): 1619-1627.

[95] YU Ning, SKRIPNIUK V, ABDELNABI S, et al. Artificial GAN fingerprints: Rooting Deepfake attribution in training data[EB/OL]. https://openreview.net/forum?id=tzfpltOnsJZ, arXiv e-prints, 2020: arXiv: 2007.08457.

[96] KIM C, REN Yi, YANG Yezhou. Decentralized attribution of generative models[EB/OL]. https://arxiv.org/abs/2010.13974, arXiv preprint arXiv:2010.13974, 2020.

作者简介

周文柏　男，1992 年生，安徽合肥人。中国科学技术大学网络空间安全学院特任副研究员，2019 年获得中国科学技术大学博士学位，主要研究方向为信息隐藏与人工智能安全。中国科大“墨子杰出青年”。现任中国图象图形学学会数字媒体取证与安全专委会委员、青年组组长。

E-mail: welbeckz@ustc.edu.cn

张卫明(通讯作者)　男，1976 年生，河北保定人。中国科学技术大学网络空间安全学院副院长，教授，博导。图象图形学学会多媒体取证与安全专家委员会副秘书长。主要研究兴趣包括信息隐藏和人工智能安全。

E-mail: zhangwm@ustc.edu.cn

俞能海　男，1964 年生，安徽芜湖人。中国科学技术大学网络空间安全学院执行院长，教授，博导。国家重点研发计划“网络空间安全”专项首席科学家。主要研究领域为视觉计算与安全、图像视频处理与分析、信息隐藏与媒体内容安全。

E-mail: ynh@ustc.edu.cn

赵汉卿　男，1998 年生，黑龙江哈尔滨人。中国科学技术大学网络空间安全学院博士研究生，主要研究方向为人工智能安全与计算机视觉。

E-mail: zhq2015@mail.ustc.edu.cn

刘泓谷　男，1997 年生，四川巴中人。中国科学技术大学网络空间安全学院博士研究生，主要研究方向为人工智能安全与计算机视觉。

E-mail: lhg9754@mail.ustc.edu.cn

韦天一　男，1997 年生，山东菏泽人。中国科学技术大学网络空间安全学院博士研究生，主要研究方向为人工智能安全与计算机视觉。

E-mail: bestwty@mail.ustc.edu.cn