计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200400130

深度伪造视频检测技术综述

暴雨轩　芦天亮　杜彦辉
中国人民公安大学警务信息工程与网络安全学院　北京 100038
(412851819@qq.com)

摘　要　深度伪造的滥用,给国家、社会和个人带来了潜在威胁。首先,介绍了深度伪造的概念和当前发展趋势,分析了基于生成对抗网络的深度伪造视频的生成原理和模型,并介绍了视频数据处理算法及主流的深度伪造数据集;其次,综述了基于视频帧内篡改特征的检测方法,针对深度伪造视频帧内的视觉伪影、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、模型;然后,针对深度伪造视频在帧间时空状态不一致的情形,阐述了相关时间序列算法和检测方法;接着,介绍了作为检测补充手段的基于区块链溯源的防篡改公共机制和数字水印、视频指纹等信息安全方法;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:深度伪造;深度学习;特征提取;视频帧;多媒体取证
中图法分类号　TP309;TP18

Overview of Deepfake Video Detection Technology

BAO Yu-xuan,LU Tian-liang and DU Yan-hui
College of Police Information Engineering and Network Security,People's Public Security University of China,Beijing 100038,China

Abstract　The abuse of deepfake brings potential threats to the country,society and individuals. Firstly,this paper introduces the concept and current trend of deepfake,analyzes the generation principle and models of deepfake videos based on generative adversarial networks,and introduces the video data processing algorithms and the mainstream deepfake datasets. Secondly,this paper summarizes the detection methods based on the tampering features in video frames. Aiming at the detection of visual artifacts and facial noise features in deepfake video frames,the classification algorithms and models related to machine learning and deep learning are introduced. Then,specific to inconsistency of time-space state between deepfake video frames,the relevant time series algorithms and detection methods are introduced. Then,the tamper-proof public mechanism based on blockchain tracing and information security methods such as digital watermark and video fingerprinting are introduced as supplementary detection means. Finally,the future research direction of deepfake video detection technology is summarized.

Keywords　Deepfake,Deep learning,Feature extraction,Video frame,Multimedia forensics

1　引言

深度伪造技术源自 deep-learning(深度学习)和 fake(造假)的组合,它可以将目标人物面部的图像叠加到视频原人物面部的相应位置,从而创建包含目标人物的视频,使目标人物说一些不曾说过的话,做一些不曾做过的动作,以达到混淆视听的目的[1]。尽管该技术可用于实现声音合成、视频分辨率修复和图像艺术风格迁移等[2],但总体而言,其弊端仍大于优势。深度伪造技术的快速发展使网络虚假视频数量呈快速上升趋势,荷兰网络安全公司 DeepTrace 在 2019 年发现 1.4 万多个深度伪造视频,较 2018 年增加了 84%[3]。由于深度伪造视频的制作门槛低、仿真度高、欺骗性强,因此该技术可能被滥用。就公民个人而言,包含其人脸的伪造视频在互联网上传播可能侵犯其名誉和隐私;就社会而言,深度伪造可通过制造虚假新闻引起金融市场的混乱;就国家而言,若深度伪造被用于制造政治矛盾、传播极端思想、煽动不安情绪,将会对国家安全造成巨大威胁[4]。

为了应对深度伪造技术带来的潜在风险,各国政府已纷纷着手准备。2018 年 9 月,欧盟主要的网络平台、社交媒体巨头、经营者代表等在布鲁塞尔发布了"*Code of Practice on Disinformation*"[5],用以解决近年来欧盟地区网络虚假信息泛滥的问题;美国一些联邦议员也呼吁政府警惕深度伪造相

到稿日期:2020-04-28　返修日期:2020-07-09　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家重点研发计划(20190178);中国人民公安大学基本科研业务费重大项目(2020JKF101)
This work was supported by the National Key R&D Program of China (20190178) and Fundamental Research Funds for the Central Universities of PPSUC(2020JKF101).
通信作者:芦天亮(lutianliang@ppsuc.edu.cn)

关技术可能给国家和社会带来的危害；2019 年 12 月，我国国家互联网信息办公室、文化和旅游部、国家广播电视总局联合印发了《网络音视频信息服务管理规定》，该规定明确指出：网络音视频信息服务提供者应当部署违法违规音视频以及非真实音视频鉴别的相关技术方案[6]。我国虽然在违法违规音视频鉴别方面已有相关技术，但在深度伪造视频检测技术方面的研究仍然欠缺，且没有成熟的检测技术可以运用。

本文从视频帧和视频完整性两个主要方面总结了现有的深度伪造视频检测技术。第 2 节对深度伪造视频生成原理进行详细综述；第 3 节介绍 deepfake 数据集和检测伪造视频的预处理手段；第 4 节和第 5 节分别介绍基于视频帧和视频完整性的深度伪造视频检测技术；最后总结全文，并对深度伪造视频检测技术的发展趋势进行展望。

2 深度伪造视频生成原理

不同于通过简单的复制、循环、压缩帧的方式对视频进行篡改，深度伪造视频的生成通常以深度学习模型为基础。前期的伪造视频生成主要依靠共享权重的自动编码器，生成对抗网络(Generative Adversarial Network，GAN)的出现克服了自编码器刻意逼近真实数据概率分布的缺点，使得生成的伪造视频更难以分辨。同时，一些团队还提出了多种改进的 GAN 算法，以实现更逼真的“伪造”效果。

2.1 自动编码器

自动编码器是一种神经网络模型，由编码器和解码器两部分组成。由于输入层和输出层拥有相同的节点数和数据维度，且均多于中间编码层，因此该网络能够通过降维的方式在输出层重构输入层的数据[7]。通常情况下，编码器将输入层数据 $\boldsymbol{x}\in X$ 映射到 $\boldsymbol{h}\in F$，如式(1)所示。

$$\boldsymbol{h}=\sigma(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{b}) \tag{1}$$

其中，$\boldsymbol{h}$ 表示从输入层得到的潜在特征向量，σ 表示激活函数，$\boldsymbol{W}$ 表示权重矩阵，$\boldsymbol{b}$ 表示偏置向量。权重和偏置通常随机初始化，并在训练过程中迭代更新。解码过程中，解码器通过 $\boldsymbol{h}$ 在输出层输出 $\boldsymbol{x}'$，实现重构输入数据 $\boldsymbol{x}$，如式(2)所示。

$$\boldsymbol{x}'=\sigma'(\boldsymbol{W}'\boldsymbol{h}+\boldsymbol{b}') \tag{2}$$

自动编码器的训练通过误差的反向传播进行，反复训练两组$(\boldsymbol{W},\boldsymbol{b})$，直至误差取得全局最小值时收敛，如式(3)所示。

$$L(\boldsymbol{x},\boldsymbol{x}')=\|\boldsymbol{x},\boldsymbol{x}'\|^2=\|\boldsymbol{x}-\sigma'(\boldsymbol{W}'(\sigma(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{b}))+\boldsymbol{b}')\|^2 \tag{3}$$

图 1 给出了共享权重的自动编码器应用于深度伪造视频的训练和生成过程。在训练阶段使用两组人脸图像，第一组是原视频中将要被替换的人脸图像，第二组是将要替换到视频中的目标人脸图像。由于对两组图像使用不同的编码器进行训练会使模型无法习得其共同潜在特征，因此 Nguyen 等[8]在训练阶段使两组图像在编码阶段共享权重，再使用含不同权重的解码器分别对两组图像完成数据重构。在生成阶段，对于新视频中将被替换的人脸图像，在编码阶段获得其面部特征潜在表示，再通过解码器 B 将第二组假人脸图像替换到新的人脸上，实现深度伪造。

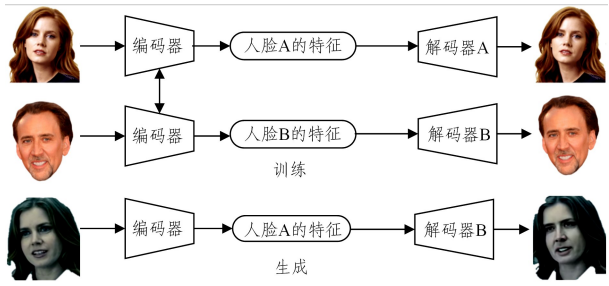

图 1 基于共享权重自编码器的深度伪造视频生成原理

Fig. 1 Deepfake video generation principle based on autoencoder for sharing weights

2.2 生成对抗网络

共享权重的自编码器为了使生成模型性能达到最佳，有时会刻意逼近真实数据，导致泛化能力不足。生成对抗网络以隐式概率分布函数为基础，更为深入地学习真实样本的分布，较自动编码器更优秀[9]。式(4)和式(5)为 GAN 模型的目标函数。

$$V(D,G)=E_{x\sim p_{\text{data}}}\log D(x)+E_{x\sim p_z}\log(1-D(G(z))) \tag{4}$$

$$\arg\min_G\max_D V(D,G) \tag{5}$$

其中，G 是生成器，D 是判别器。生成器从输入空间将满足随机分布的输入数据 z 映射到生成空间，记为 $x=G(z)$，再通过判别器将 x 映射为 $D(x)$，通过取期望值的方式求解目标函数。G 的代价函数希望 $V(D,G)$ 尽量小，D 希望 $V(D,G)$ 尽量大，以形成两者之间的博弈。在训练过程中，通过将期望转为积分并引入 KL 散度和 JS 散度，使 G 和 D 各自代价函数最小，通过模型收敛达到平衡。最终，D 无法准确辨认 G 生成的数据，即 $D(G(z))=0.5$，生成数据和真实数据具有相同分布，使得生成的深度伪造视频能够欺骗大多数人的眼睛。

2.3 其他改进的 GAN 算法

在实际运用中，改进的 GAN 模型能够达到比 GAN 模型更好的生成效果。

(1)DCGAN

DCGAN 去除了所有池化层，仅通过卷积层使网络自身学习空间上采样和下采样，并对每一层网络进行批归一化处理[10]。批归一化的关键步骤即尺度变换和偏移，如式(6)所示。

$$y_i\leftarrow\gamma\hat{x}_i+\beta=BN_{\gamma,\beta}(x_i) \tag{6}$$

其中，γ 为尺度因子，β 为平移因子，均由网络在训练时习得。批归一化通过求取每个训练批次数据的均值和方差，并对其做归一化，以线性变换的方式高度还原数据原始输入的分布。同时，DCGAN 模型使用 Tanh 激活函数提高模型学习效率，极大地提升了 GAN 训练的稳定性以及生成结果的质量。

(2)Wasserstein GAN

采用总变差距离、KL 距离和 JS 距离来刻画真实数据和生成数据分布的相似度时无法产生有效的梯度，Wasserstein GAN[11]引入了 EM(Earth-Mover)距离来解决这一问题。EM 距离表达式如式(7)所示。

$$W(p_1,p_2)=\inf_{\gamma\in\Pi(p_1,p_2)}E_{(x,y)\sim\gamma}[\|x-y\|] \tag{7}$$

其中，$\Pi(p_1,p_2)$ 为真实数据与生成数据的联合概率分布。通过求取联合分布样本对的期望距离可以缓解模型的梯度消失问题，从而更有效地优化模型在反向传播中的参数。

(3)CycleGAN

匹配图像的转换需要相似度极高的数据集，因而非匹配图像的互相转化更贴合深度伪造的实现过程。iGAN 团队提出的 CycleGAN[12] 使用两个不同领域图像的 GAN，使各自生成器生成的对方领域的图像尽全力“骗过”对方的判别器。为避免出现各 GAN 独立训练导致生成器直接从对方领域生成图像的情况，CycleGAN 引入 Cycle 连续性损失函数，其公式如式(8)所示。

$$L_{\mathrm{cyc}}(G,F)=E_{x\sim p_{\mathrm{data}}}(x)[\|F(G(x))-x\|_1]+E_{y\sim p_{\mathrm{data}}}(x)[\|G(F(y))-y\|_1] \tag{8}$$

在使用自身领域图像的 GAN 生成器生成图像后，通过对方领域图像的 GAN 生成器还原原领域的数据 x'。通过损失函数的惩罚让 x 与 x' 尽可能接近，从而使生成内容尽可能拟合对方领域的图像，实现非匹配图像的转化。

(4)E-GAN

E-GAN[13] 将进化策略引入模型的优化过程，通过变异算子产生生成器种群，增加启发式损失函数和最小二乘损失函数，使模型在每次训练中自适应选择最佳目标函数进行更新，增强训练的稳定性。E-GAN 模型使用式(9)评估迭代过程中生成器的质量，使用式(10)评估生成样本的多样性，最终选择出拟合真实样本分布最佳的生成器，有效避免了模式崩塌。

$$F_q=E_z[D(G(z))] \tag{9}$$

$$F_d=-\log\|\nabla_D-E_x[\log D(x)]-E_z[\log(1-D(G(z)))]\| \tag{10}$$

(5)StyleGAN[14]

StyleGAN 向传统生成器的输入加入由 8 个全连接层组成的映射网络，通过该映射网络生成一个不必遵循训练数据分布的向量，从而降低网络习得的图像特征之间的相关性，减少各类特征之间的干扰，以实现对生成数据的控制。同时，在每次对控制向量卷积后，采用一次自适应实例归一化(Adaptive Instance Normalization，AdaIN)来添加噪声，使得生成的人脸更逼真。

2.4 深度伪造开源项目及工具

一些团队利用自动编码器和生成对抗网络开发出了相应的深度伪造项目和工具。Faceswap[15] 是一个开源的 AI 换脸项目，使用两个共享参数的编解码器对生成深度伪造图像和视频；Faceswap-GAN[16] 引入感知对抗损失函数，从多角度惩罚伪造图像与原图像的差异，在与输入面部保持一致的基础上使生成人物眼睛的运动更加逼真，有效消除了深度伪造过程中分割掩膜时出现的图像噪声问题，从而生成了更高质量的视频；DFaker[17] 通过 SSIM 算法评估伪造视频与原视频的相似性，从亮度、对比度、结构 3 个维度构造相似测量函数，以高斯加权函数为加权窗口生成映射矩阵来实现视频帧局部质量检测。

国内的深度伪造工具以 ZAO[18] 为主，其运用 CycleGAN 模型完成人脸伪造，并通过模糊边缘处理，以 Canny 边缘检测算法通过对图像降噪、像素点梯度求解和双阈值检测提取伪造的人脸边缘，再以像素的权重矩阵对提取边缘进行高斯模糊，最终获得更自然的伪造人脸。

深度伪造项目和工具的盛行虽起到一定的娱乐作用，但也存在因协议不规范而泄露用户相关隐私数据的风险，为诈骗等违法犯罪提供可能。因此，深度伪造视频检测技术的提出极为必要。

3 深度伪造数据集及伪造视频预处理

3.1 深度伪造数据集

获取深度伪造视频数据集并建立完备的对应数据库利于统一化管理相关的数据，是实施检测的前提。目前，深度伪造数据集大多源于国外，通常利用 GAN 模型对来自 YouTube 等平台的源视频人脸进行替换生成。常见的深度伪造数据集如表 1 所列。

表 1 深度伪造数据集

Table 1 Deepfake datasets

Datasets	Number of true/false videos	link
UADFV	49/49	https://github.com/danmohaha/WIFS2018_In_Ictu_Oculi
Celeb-DF	590/5639	https://github.com/danmohaha/celeb-deepfakeforensics
FF++	1000/1000	https://github.com/ondyari/FaceForensics
DF-TIMIT	320/640	https://www.idiap.ch/dataset/deepfaketimit
DFDC	1131/4113	https://www.deepfakedetectionchallenge.ai/

(1)FaceForensics++[19]

FF++在 FaceForensics 数据集的基础上进行扩展，选取来自 YouTube 的 509 个视频，采用 GAN 模型 Pix2Pix 弥补视频低分辨率的细节信息，并通过 Face2Face 开源换脸项目重新合成视频中的人脸，最终生成包含 1 000 个深度伪造视频的数据集。

(2)DF-TIMIT[20]

DeepFake-TIMIT 数据集以来自 VidTIMIT 视频数据库的 16 对具有相似特征的人物视频为基础，使用 faceswap-GAN 模型交换人脸，在 Tesla P40 处理器训练 12 h 和 4 h 分别生成高低质量视频各 320 个，分辨率为 128 * 128 和 64 * 64，并利用直方图归一化和改变光照条件的方式调整深度伪造视频的噪声。

(3)Celeb-DF[21]

该数据集收集了 YouTube 上 59 位不同性别、年龄和种族的名人的 590 个真实视频剪辑合成了多达 5 639 个深度伪造视频，每个视频的长度约为 13 s。该数据集使用层数更多的自编码器提高视频的分辨率，通过颜色迁移算法减少伪造视频中颜色不匹配的情况，并利用更平滑的掩膜覆盖目标视频中的人物面部，很大程度地去除了视频中的视觉伪影，在所

有深度伪造视频数据集中质量最高。

3.2 伪造视频预处理技术

通过总结现有文献，对视频帧进行有效的数据处理，如提取视频中的关键帧并进行裁剪，实施人脸关键点检测并将人脸对齐到同一坐标系，可减少数据冗余和噪声，有助于提升分类模型检测伪造视频的准确率。

3.2.1 提取关键帧

关键帧指能够体现视频主要视觉内容的帧图像集合，精简的关键帧集合不仅能够传达视频的主要内容，还可以大幅降低视频分析过程的时空复杂度和计算量。提取关键帧的算法主要有3类：基于镜头的方法、基于视频聚类的方法、基于运动分析的方法。目前尚没有具有绝对优势的关键帧提取算法。

基于镜头的方法首先对视频文件按镜头变化分割，然后在每个镜头选择首尾两帧或中间帧作为关键帧[22]，或者对镜头的帧序列进行等间隔取样[23]。这种方法的优点是实施简单，但是当视频内容复杂、人物运动剧烈时效果不佳。

基于视频聚类的方法通常在各视频簇中初始化聚类中心帧，并用颜色直方图计算该聚类中心帧的颜色通道中心，再迭代其他视频帧到该中心的距离，最终选择距离各簇聚类中心最近的帧为关键帧。文献[24]利用模糊K均值聚类算法提取关键帧；文献[25]通过改进的K均值算法将初始聚类中心交由视频数据本身的分布来决定，且聚类数 K 根据视频内容以自适应方式获得最佳取值，从而增强了结果的稳定性。该方法能充分反映视频内容，但对视频帧的顺序欠缺考虑。

基于运动分析的方法通过结合物体和人物的运动特征来提取关键帧。Ejaz等[26]提出一种结合相对运动强度和相对运动方向的动态视觉注意方法，较静态的视觉注意机制取得了更好的效果。Wolf等[27]通过光流法来计算镜头中人物、物体的运动量，并在其局部最小值处提取关键帧。该方法可以表达出视频的运动特征，但展现出的鲁棒性较差。

3.2.2 人脸对齐

通过关键帧提取后得到的视频人脸图像因角度差异，没有划归到统一坐标轴下，因此需要使用人脸对齐定位面部特征点的位置，然后通过几何变换来减小不同帧的人脸之间的姿态差异，以提升后续特征提取阶段对各种人脸姿态变化的鲁棒性。人脸对齐算法主要分为3类：基于优化的方法、基于回归的方法和基于深度学习的方法。

基于优化的算法通过模型的拟合达到最优结果。ASM(Active Shape Models)算法利用全局形状模型来匹配人脸初始形状，将人脸初始帧序列的68个特征点所表征的形状向量进行归一化、PCA(Principal Components Analysis)降维形成初始模型，然后通过局部特征点匹配计算最优模型[28]。AAM(Active Appearance Models)算法针对图像的纹理信息建立形状和灰度结合的模型，在图片表情或者光线环境发生变化时较ASM算法更健壮[29]。该方法在简单特征点定位方面具有优势，但依赖于人脸形状的初始化参数。

基于回归的算法通过向量操作降低计算复杂度。ERT算法[30]首先建立级联的残差回归树，依据图像特征进行节点分裂，再通过叠加叶子节点的图像残差使人脸从当前形状逐步回归到真实形状，实现人脸对齐。ESR算法[31]以直接学习向量回归函数而不是整个人脸模型的方式，有效克服了模型训练过程中参数复杂的问题。RCPR算法在ESR算法的基础上增加了对面部遮挡变化的处理手段和面部特征的选择种类，增强了人脸对齐在遮挡情况下的鲁棒性[32]。

基于深度学习的人脸对齐方法以Sun等[33]提出的CNN三级级联网络为代表，第一层级网络用以提取人脸特征坐标点并平均网络输出结果，第二、三层级网络分别依据第一层坐标点对原始图像进行裁剪并做精准预测。Zhou等[34]提出的并行分组网络以由粗到精的级联方式计算人脸特征点位置坐标，取得了良好的对齐效果。基于深度学习的方法能够更全面地学习人脸特征，但因模型复杂而较基于优化的算法通用性差。

4 基于视频帧的检测方法

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。深度伪造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，为检测深度伪造视频提供了依据。

4.1 基于帧内差异的检测方法

深度伪造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。

4.1.1 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2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。

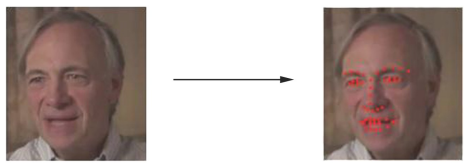

图2 面部关键点提取

Fig. 2 Extraction of facial key points

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3、图4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[35]。

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[35]。

图 3 伪造人脸眼睛颜色差异

Fig. 3 Eye color differences between deepfake faces

图 4 人眼颜色特征的提取

Fig. 4 Extraction of eye color features

(2)3D 头部姿态差异

Yang 等[36]用旋转矩阵 $\boldsymbol{R}$ 和平移向量 $\boldsymbol{t}$ 评估伪造视频和真实视频人物头部的方向、姿态差异。式(11)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 $\boldsymbol{R}$ 和 $\boldsymbol{t}$，其中$[U,V,W]^{\mathrm{T}}$ 和$[x,y]^{\mathrm{T}}$ 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 5 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 5(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 5(n))。选取真假人脸的中心区域(见图 5(i)、图 5 (l))和整个面部区域(见图 5(j)、图 5(m))的旋转矩阵差$\boldsymbol{R}_a-\boldsymbol{R}_c$和平移向量差$\boldsymbol{t}_a-\boldsymbol{t}_c$放入 SVM 分类器进行分类训练，能够检测出伪造视频。

$$\min_{\boldsymbol{R},\boldsymbol{t},s}\sum_{i=1}^{n}\left\|s\begin{pmatrix}x_i\\y_i\\1\end{pmatrix}-\begin{pmatrix}f_x&0&c_x\\0&f_y&c_y\\0&0&1\end{pmatrix}\left(\boldsymbol{R}\begin{pmatrix}U_i\\V_i\\W_i\end{pmatrix}+\boldsymbol{t}\right)\right\|^2 \tag{11}$$

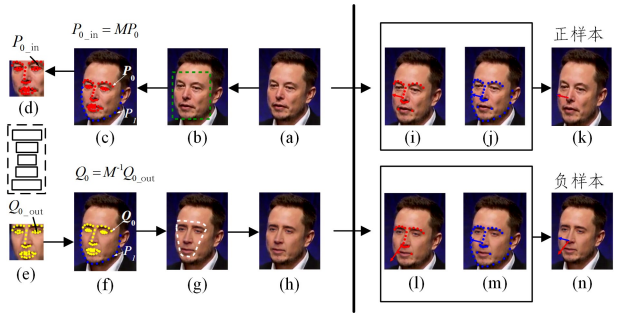

图 5 基于 SVM 的 3D 头部姿态差异检测

Fig. 5 3D head poses difference detection based on SVM

(3)人脸关键点的概率分布差异

如图 6 所示，Yang 等[37]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。

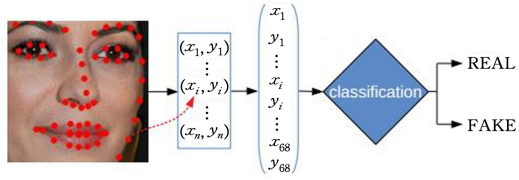

图 6 基于人脸关键点的概率分布差异检测

Fig. 6 Probability distribution difference detection based on facial key points

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。Wadhwa 等[38]提出一种欧拉影像放大算法，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、振幅来鉴别深度伪造视频。

4.1.2 基于 CNN 的检测方法

深度伪造过程由于受到计算资源和生成时间的限制，只能通过对面部中心区域进行仿射变换、压缩像素来实现篡改，从而导致面部中心伪造部分和脸部边缘的分辨率不一致。Li 等[39]提取原视频帧中裁剪的人脸，并使用高斯模糊生成不同分辨率、亮度、锐度、对比度的负样本，以模拟深度伪造过程中出现的各种差异。在训练阶段，分别使用 VGG16[40]，ResNet50，ResNet101，ResNet152[41]4 种 CNN 模型对负样本的相关特征进行学习，用 UADFV 和 DF-TIMIT 数据集进行验证。其中，ResNet 由于引入了残差学习而比 VGG16 的准确率更高，ResNet50 在 ResNet 网络类中因网络层数加深而效果最佳。

马尔兰州大学的 Adobe 团队提出一种了更快的 R-CNN(Region-CNN)双流网络[42]。双流指 RGB 流和由 RGB 流经 SRM 滤波器获得的图像噪声流。如图 7 所示，该方法在两类卷积层和以 RGB 特征为输入的 RPN 层的共同作用下获取视频帧的被篡改区域，经 ROI 池化层提取固定维度的 RGB 特征和噪声特征流，最后用双线性池化层实现两类特征的融合，以检测出伪造视频。由于引入的噪声流能够依据帧的灰度值差异寻找伪造区域，因此该方法比仅使用 RGB 流检测深度伪造视频的效果更好。

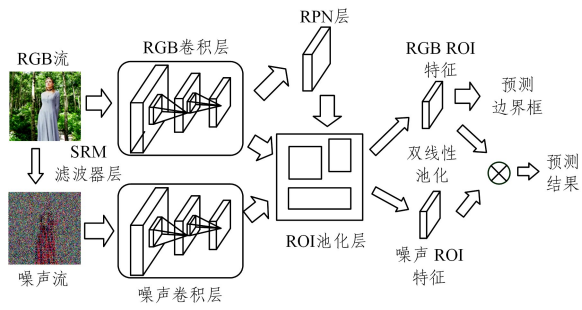

图 7 R-CNN 双流网络检测方法

Fig. 7 R-CNN two stream network detection method

此外，Afchar 等[43]提出的 MesoNet 网络结合 Inception

模块，以基于中层的语义的方式对深度伪造视频进行检测；Zhang 等[44]使用轻量级流模块来代替卷积神经网络中的全局平均池化层和全连接层，从而在提取帧的噪声特征时极大地减少了计算代价；Bayar 等[45]提出的 MISLnet 通过引入受约束的卷积层来减少特征受非线性运算的破坏，以更好地学习篡改特征。

4.1.3 基于“胶囊网络”的检测方法

Nguyen 等[46]将结合动态路由算法的“胶囊网络”用于深度伪造检测。该算法通过计算前后胶囊网络神经元向量的内积来动态更新权重，再由权重的变化改变耦合系数，最后经耦合系数与胶囊网络权重的线性加权求和来实现网络的参数更新。

“胶囊网络”由于采用向量模来衡量伪造视频视觉伪影特征出现的概率，相比卷积神经网络用标量表示神经元及其权重的方式，有效降低了视频人物因扭转角度、方向差异而带来的检测误差，同时该模型能使用更少的训练数据最大化地保留有价值的信息，且其传输和运算逻辑更符合人脑神经元的工作方式，因此在特征提取、抵御噪声方面优于 CNN。此外，“胶囊网络”由于模型的特殊性，能够更全面地检测各种视频伪造手段，且模型的适用不局限于计算机视觉领域。

4.2 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。

4.2.1 CNN 和 RNN 结合的检测方法

如图 8 所示，Güera 等[47]提出了一种 CNN 和长短期记忆网络(Long Short-Term Memory，LSTM)[48]相结合的方法来检测深度伪造视频。

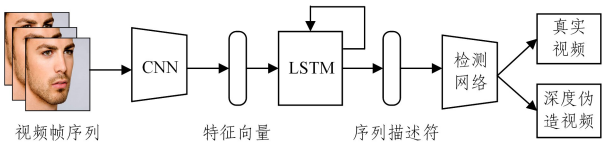

图 8 基于 CNN 和 RNN 结合的检测方法

Fig. 8 Detection method based on CNN and RNN

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 HOHA 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。

Sabir 等[49]在上述方法的基础上采用 ResNet50 和 DenseNet 提取帧序列的特征。ResNet50 运用残差学习有效解决了网络加深而导致的训练集准确率下降的问题。DenseNet 在减少参数的基础上，以鼓励特征重用的方式有效地缓解了梯度消失的问题。实验中，GRU 模型仅使用一个门控即可同时实现遗忘和选择记忆，较 LSTM 有效提高了训练效率。在训练方式上，该方法利用端到端的方法减少人工干预，让模型尽可能自己学习帧序列特征，从而使结果更准确。如表 2 所列，实验在 Deepfake，Face2Face，FaceSwap 3 种开源的深度伪造项目上对 3 种方法进行验证：基于 CNN，基于 CNN 和人脸对齐，基于 CNN、RNN 和人脸对齐。结果证明了数据处理中采用人脸对齐后识别准确率更高，采用 DenseNet 和 GRU 结合的模型训练效果最佳。

表 2 检测准确率

Table 2 Detection accuracy

Detecting method	CNN	CNN+face alignment	CNN+RNN+face alignment
Deepfake	94.5～94.8	96.0～96.7	96.9
Face2Face	89.8～90.6	90.1～93.2	94.3
FaceSwap	90.9～93.1	93.8～96.1	96.3

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。Li 等[50]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。

4.2.2 光流法和 CNN 结合的检测方法

光流法利用视频帧的像素点在时域上的变化来寻找前后相邻两帧存在的对应关系，从而计算出相邻帧的光流矢量。Lucas-Kanade 算法是运用两帧差分来计算光流矢量的经典算法，其通过泰勒公式展开某像素点光强度在相邻两帧变化的约束方程，将该像素点的邻域点集考虑在内，以解决未知量过多的问题；利用最小二乘法使 n 个像素点方程组偏移量平方和最小，从而求出方程的最优解；以矩阵运算求解出该像素点的光流矢量。Amerini 等[51]使用 LK 算法和 PWC-Net 将伪造视频帧估算为光流。PWC-Net 将同一图像在不同维度

下求得的特征组合起来，得到可反映多维度信息的特征集，再使用上采样对特征集进行扭曲变换并计算代价函数，最后用光流评估器将原始 RGB 帧转化为光流矢量。真实视频人物面部周围逐帧形成的光流矢量与伪造视频在方向、大小、数量上存在的差异，可被 CNN 捕获。实验中，VGG16 模型通过减少超参数而专注于构建卷积层的简单网络，极大地减少了提取特征的工作量；同时，其使用 Adam 算法以 1×10^{-4} 的学习率更新神经网络权重，在包含 1 000 个原始视频序列的 FaceForensics++数据集上进行训练、验证；最后，通过 sigmod 函数和全连接层对真伪视频逐帧做出二分类，以检测出深度伪造视频，准确率达 82%。

5 基于视频完整性的检测方法

基于视频帧的检测方法在面对高质量的伪造视频时检测效率会大幅下降，因此需要通过建立公共机制或使用数据完整性校验等手段对原始视频的传播进行溯源及做出篡改检测。此类方法识别准确率高且适用范围广，不局限于深度伪造视频，也可以检测其他篡改类型的伪造视频及数据文件（如文字、图像、音频），但需要在视频发布前准予授权，或对数据文件进行预处理，如提取视频关键信息或在其中嵌入水印，无法适用于未提前掌握原始视频的情况。

5.1 基于区块链等公共机制的检测方法

区块链为视频等作品的鉴权提供途了径，能保证权属的真实性。相关媒体数据在区块链上被授权后，其后续修改都会被实时记录，以实现对数字媒体版权的保护。HASAN 等[52]提出了区块链和智能合约结合的视频真实性验证机制，由存储视频及其元数据的星际文件系统 IPFS、存储作者身份信息的以太坊域名系统 ENS 和作者声誉系统组成。如图 9 所示，视频发布时会配置智能合约，若之后有用户想对原视频进行编辑、修改，则需要在 IPFS 上向原作者发出获取权限请求并提交子合约，原作者审核通过后与原合约建立父子合约关系。如果能通过现有视频合约追溯到原视频合约，则说明原视频没有被篡改。

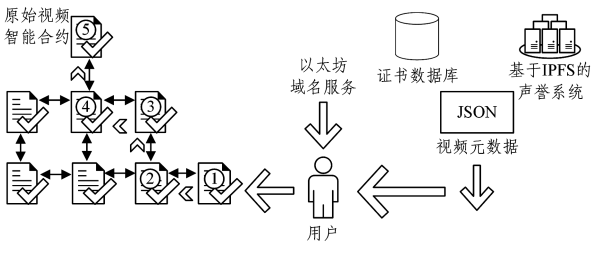

图 9 区块链和智能合约结合的检测方法

Fig. 9 Blockchain and smart contract detection method

5.2 基于水印与视频指纹的检测方法

在原始视频中嵌入水印或者提取视频指纹是传统信息安全领域的防篡改技术，具有通用性，但需要预先对视频进行处理。

5.2.1 基于水印的视频篡改检测

数字水印（Digital Watermarking）在不被知觉系统感知和不影响原数字载体使用价值的前提下，将只能通过特定检测器提取的标识信息嵌入原载体，通过隐藏信息的改变实现防篡改。其可在视频发布前通过向其特定区域添加水印，在接收方验证水印的存在后，确定视频的真实性。视频水印构造方法可分为基于空间域、基于压缩域和基于变换域的方法。

此外，感光元件硅的薄厚差异会导致相机的 PRNU 模式不同，可被当作检测视频伪造的水印。Koopman 等[53]从视频中提取关键帧并分组，通过比对各组的 PRNU 返回一个归一化互相关分数（Normalised Cross Correlation Scores），证明了原视频的 NCCS 明显高于深度伪造视频，由此实现了对深度伪造视频的检测。

5.2.2 基于视频指纹的检测方法

视频指纹指根据视频生成的能够唯一标识其内容的特征向量，常用于媒体数据的版权保护。可通过对比真伪视频指纹的相似性，来确定两者的同源性。

传统的视频指纹算法提取帧的时空特征作为视频指纹。Lee 等[54]将视频帧重采样后转化为灰度图像并分块，提取分块后连续帧的梯度方向中心（Centroids of Gradient Orientations，CGO）作为视频指纹，以阈值匹配衡量视频相似度；Tang 等[55]利用提出的时空极余弦变换（Spatial-Temporal Polar Cosine Transform，ST-PCT）提取视频帧的时空域特征，并将其压缩为二进制哈希码作为待识别的指纹特征。

此外，深度学习模型在视频指纹提取方面也具有很大优势。Li 等[56]利用三维卷积神经网络（Three-Dimension CNN，3D-CNN）获取视频连续帧的信息通道，提取其局部运动信息，最终将多通道特征融合作为视频指纹。Ma 等[57]考虑了连续帧之间的时间相关性，通过 CNN 和 RNN 分别提取视频的时间特征和空间特征，再将每个视频的帧级别特征表征为该视频的特征，作为视频指纹，减少了时间信息的丢失。

6 总结和展望

图 10 所示为本文提出的深度伪造视频检测技术。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。

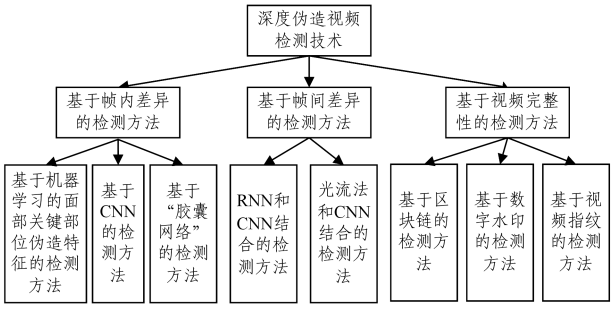

图 10 深度伪造视频检测技术

Fig. 10 Deepfake video detection technology

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，可以分别提取深度伪造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[58]，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RGB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。

结束语 本文分析了国内外深度伪造视频检测的前沿研究进展。从检测方法来看，深度伪造视频检测的主要方法可分为两类：1)利用人工智能算法、模型对视频帧进行特征提取，以进行分类检测；2)利用视频内容完整性，结合区块链、传统信息安全等手段对原始发布视频进行溯源或防篡改检测。前者能够对特定的伪造手段提出具有针对性的检测方法，但是往往滞后于伪造方法的出现，不能全面、广泛地打击深度伪造视频；后者适用范围广，不局限于深度伪造视频，亦能有效检测其他媒体数据和篡改手段，但是需要提前掌握原始数据并做预处理。从研究成果来看，针对深度伪造及相关检测技术的主要研究人员集中在国外各研究机构，我国相关算法和技术研究相对滞后。从数据集来看，国外已经拥有十余个开源的换脸数据集，而我国没有用于研究和测试的深度伪造数据集。总体来说，我国在基于人工智能技术的视频合成和检测领域落后于欧美，亟需开展相关研究，建立以基于人工智能算法的伪造视频帧内、帧间特征检测方法为主，以区块链公共机制、传统信息安全对视频进行完整性校验为辅的检测机制，掌握合成伪造视频的关键技术和检测核心技术，以维护我国国家安全和社会稳定。

参 考 文 献

[1] LONG K, MA Y, ZHU Q C. How Will Deepfake Technology Influence National Security: Emerging Challenges and Policy Implications [J]. China Information Security, 2019(10): 21-34.

[2] CHENG X Y, XIE L, ZHU J X, et al. Review of Generative Adversarial Network[J]. Computer Science, 2019, 46(3): 74-81.

[3] 深圳英鹏信息技术股份有限公司. 向 Deepfake 宣战！[EB/OL]. (2020-01-14)[2020-04-05]. https://baijiahao.baidu.com/s?id=1655671833927886540&wfr=spider&for=pc.

[4] CHESNEY R, CITRON D. Deepfakes and the New Disinformation War: The Coming Age of Post-truth Geopolitics[J]. Foreign Aff., 2019, 98: 147.

[5] NEWS ARTICLE. Code of Practice on Disinformation [EB/OL]. (2018-09-26)[2020-04-05]. https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation.

[6] 国家互联网信息办公室等三部门规范网络音视频信息服务[J]. 中国广播, 2019(12): 39.

[7] HINTON G E, SALAKHUTDINOV R. Reducing the Dimensionality of Data with Neural Networks [J]. Science, 2006, 313(5786): 504-507.

[8] NGUYEN T T，NGUYEN C M，NGUYEN D T，et al. Deep Learning for Deepfakes Creation and Detection[J]. arXiv：1909.11573，2019.

[9] GOODFELLOW I，POUGET-ABADIE J，MIRZA M，et al. Generative Adversarial Nets[C]// Advances in Neural Information Processing Systems. 2014：2672-2680.

[10] RADFORD A，METZ L，CHINTALA S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks[J]. arXiv：1511.06434，2015.

[11] ARJOVSKY M，CHINTALA S，BOTTOU L. Wasserstein GAN[J]. arXiv：1701.07875，2017.

[12] ZHU J Y，PARK T，ISOLA P，et al. Unpaired Image-to-image Translation Using Cycle-consistent Adversarial Networks[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017：2223-2232.

[13] WANG C，XU C，YAO X，et al. Evolutionary Generative Adversarial Networks[J]. IEEE Transactions on Evolutionary Computation，2019，23(6)：921-934.

[14] KARRAS T，LAINE S，AILA T. A Style-based Generator Architecture for Generative Adversarial Networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019：4401-4410.

[15] Faceswap. Faceswap is the leading free and Open Source multi-platform Deepfakes software [EB/OL]. (2019-08-15)[2020-04-23]. https：//faceswap. dev/.

[16] Shaoanlu. Faceswap-GAN [EB/OL]. (2019-10-04) [2020-04-23]. https：//github. com/shaoanlu/faceswap-GAN.

[17] dfaker. DFaker[EB/OL]. (2018-02-24)[2020-04-23]. https：//github. com/dfaker/df.

[18] SU M L. ZAO Privacy Protection Saves "big hole"[J]. Computers & Networks，2019，45(17)：8-10.

[19] ROSSLER A，COZZOLINO D，VERDOLIVA L，et al. Faceforensics＋＋：Learning to Detect Manipulated Facial Images [C]// Proceedings of the IEEE International Conference on Computer Vision. 2019：1-11.

[20] KORSHUNOV P，MARCEL S. Deepfakes：a New Threat to Face Recognition? Assessment and Detection[J]. arXiv：1812.08685，2018.

[21] LI Y，YANG X，SUN P，et al. Celeb-df：A new Dataset for Deepfake Forensics[J]. arXiv：1909.12962，2019.

[22] WANG X，DING H，CHEN H. A Shot Clustering Based Approach for Scene Segmentation[J]. Journal of Image and Graphics，2007，12(1)：2127-2131.

[23] ZOLFAGHARI M，SINGH K，BROX T. Eco：Efficient Convolutional Network for Online Video Understanding[C]// Proceedings of the European Confe-rence on Computer Vision (ECCV). 2018：695-712.

[24] YU X D，WANG L，TIAN Q，et al. Multilevel Video Representation with Application to Keyframe Extraction[C]// Proceedings of the 10th International Multimedia Modelling Conference. IEEE，2004：117-123.

[25] YIN Y，JIANG H N. Key Frame Extraction Based on Clustering of Optimizing Initial Centers[J]. Computer Engineering and Applications，2007(21)：165-167.

[26] EJAZ N，BAIK S W，MAJEED H，et al. Multi-scale Contrast and Relative Motion-based Key Frame Extraction[J]. EURASIP Journal on Image and Video Processing，2018，2018(1)：40.

[27] WOLF W. Key Frame Selection by Motion Analysis[C]// Proceedings of the 1996 IEEE International Conference on Acoustics，Speech，and Signal Processing Conference. IEEE，1996：1228-1231.

[28] COOTES T F，TAYLOR C J，LANITIS A. Active Shape Models：Evaluation of a Multi-Resolution Method for Improving Image Search[C]// BMVC. 1994：327-336.

[29] COOTES T F，EDWARDS G J，TAYLOR C J. Active Appearance Models[C]// European Conference on Computer Vision. Berlin：Springer，1998：484-498.

[30] KAZEMI V，SULLIVAN J. One Millisecond Face Alignment with an Ensemble of Regression Trees[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014：1867-1874.

[31] CAO X，WEI Y，WEN F，et al. Face Alignment by Explicit Shape Regression[J]. International Journal of Computer Vision，2014，107(2)：177-190.

[32] BURGOS-ARTIZZU X P，PERONA P，DOLLÁR P. Robust Face Landmark Estimation under Occlusion[C]// Proceedings of the IEEE International Conference on Computer Vision. 2013：1513-1520.

[33] SUN Y，WANG X，TANG X. Deep Convolutional Network Cascade for Facial Point Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013：3476-3483.

[34] ZHOU E，FAN H，CAO Z，et al. Extensive Facial Landmark Localization with Coarse-to-fine Convolutional Network Cascade [C]// Proceedings of the IEEE International Conference on Computer Vision Workshops. 2013：386-391.

[35] MATERN F，RIESS C，STAMMINGER M. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations[C]// Proceedings of 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). IEEE，2019：83-92.

[36] YANG X，LI Y，LYU S. Exposing Deepfakes Using Inconsistent Head Poses[C]// Proceedings of 2019 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). IEEE，2019：8261-8265.

[37] YANG X，LI Y，QI H，et al. Exposing Gan-synthesized Faces Using Landmark Locations[C]// Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. 2019：113-118.

[38] WADHWA N，WU H Y，DAVIS A，et al. Eulerian Video Magnification and Analysis[J]. Communications of the ACM，2016，60(1)：87-95.

[39] LI Y，LYU S. Exposing Deepfake Videos by Detecting Face Warping Artifacts[J]. arXiv：1811.00656，2018.

[40] SIMONYAN K，ZISSERMAN A. Very Deep Convolutional Net-

works for Large-scale Image Recognition[J]. arXiv:1409.1556, 2014.

[41] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.

[42] ZHOU P, HAN X, MORARIU V I, et al. Learning Rich Features for Image Manipulation Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:1053-1061.

[43] 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). IEEE, 2018:1-7.

[44] ZHANG P, ZOU F, WU Z, et al. FeatherNets: Convolutional Neural Networks as Light as Feather for Face Anti-spoofing [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019:1574-1583.

[45] BAYAR B, STAMM M C. Constrained Convolutional Neural Networks: A New Approach Towards General Purpose Image Manipulation Detection[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(11):2691-2706.

[46] NGUYEN H H, YAMAGISHI J, ECHIZEN I. Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos[C]// Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019:2307-2311.

[47] GÜERA D, DELP J. Deepfake Video Detection Using Recurrent Neural Networks[C]// 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2018:1-6.

[48] HOCHREITER S, SCHMIDHUBER J. Long Short-term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[49] SABIR E, CHENG J, JAISWAL A, et al. Recurrent Convolutional Strategies for Face Manipulation Detection in videos[J]. Interfaces (GUI), 2019, 3:1.

[50] LI Y, CHANG M C, LYU S. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking[C]// 2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2018:1-7.

[51] AMERINI I, GALTERI L, CALDELLI R, et al. Deepfake Video Detection through Optical Flow based CNN[C]// Proceedings of the IEEE International Conference on Computer Vision Workshops. 2019:1205-1207.

[52] HASAN H R, SALAH K. Combating Deepfake Videos Using Blockchain and Smart Contracts[J]. IEEE Access, 2019, 7:41596-41606.

[53] KOOPMAN M, RODRIGUEZ A M, GERADTS Z. Detection of Deepfake Video Manipulation[C]// The 20th Irish Machine Vision and Image Processing Conference (IMVIP). 2018:133-136.

[54] LEE S, YOO C D. Video Fingerprinting Based on Centroids of Gradient Orientations[C]// Proceedings of 2006 IEEE International Conference on AcousticsSpeech and Signal Processing. IEEE, 2006, 2:1-4.

[55] TANG W, WO Y, HAN G. Geometrically Robust Video Hashing Based on ST-PCT for Video Copy Detection[J]. Multimedia Tools and Applications, 2019, 78(15):21999-22022.

[56] LI J, ZHANG H, WAN W, et al. Two-class 3D-CNN Classifiers Combination for Video Copy Detection[J]. Multimedia Tools and Applications, 2018, 3:1-13.

[57] MA C, GU Y, GONG C, et al. Unsupervised Video Hashing via Deep Neural Network[J]. Neural Processing Letters, 2018, 47(3):877-890.

[58] CHEN P, LIANG T, DAI J, et al. Forged Facial Video Detection Based on Global Temporal and Local Spatial Feature[J]. Journal of Cyber Security, 2020, 5(2):73-83.

BAO Yu-xuan, born in 1997, master. His main research interests include cyber security and artificial intelligence.

LU Tian-liang, born in 1985, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include cyber security and artificial intelligence.