计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202111119
60(6): 1396−1416, 2023

多模态深度伪造及检测技术综述

李泽宇[1] 张旭鸿[2] 蒲誉文[1] 伍一鸣[1] 纪守领[1]
[1](浙江大学计算机科学与技术学院 杭州 310007)
[2](浙江大学控制科学与工程学院 杭州 310007)
(22121026@zju.edu.cn)

A Survey on Multimodal Deepfake and Detection Techniques

Li Zeyu[1], Zhang Xuhong[2], Pu Yuwen[1], Wu Yiming[1], and Ji Shouling[1]
[1](*College of Computer Science and Technology*, *Zhejiang University*, *Hangzhou* 310007)
[2](*College of Control Science and Engineering*, *Zhejiang University*, *Hangzhou* 310007)

Abstract With the application of all kinds of deep learning generation models in various fields, the authenticity of their generated multimedia files has become increasingly difficult to distinguish, therefore, deepfake technology has been born and developed. Utilizing deep learning related techniques, the deepfake technology can tamper with the facial identity information, expressions, and body movements in videos or pictures, and generate fake voice of a specific person. Since 2018, when Deepfakes sparked a wave of face swapping on social networks, a large number of deepfake methods have been proposed, which had demonstrated their potential applications in education, entertainment, and some other fields. But at the same time, the negative impact of deepfake on public opinion, judicial and criminal investigations, etc. can not be ignored. As a consequence, more and more countermeasures have been proposed to prevent deepfake from being utilized by the criminals, such as the detection of deepfake and watermark. Firstly, a review and summary of deepfake technologies of different modal types and corresponding detection technologies are carried out, and the existing researches are analyzed and classified according to the research purpose and research method. Secondly, the video and audio datasets widely used in the recent studies are summarized. Finally, the opportunities and challenges for future development in this field are discussed.

Key words deepfake; deepfake detection; deep learning; face replacement; generative adversarial network

摘 要 随着各种深度学习生成模型在各领域的应用,生成的多媒体文件的真伪越来越难以辨别,深度伪造技术也因此得以诞生和发展.深度伪造技术通过深度学习相关技术能够篡改视频或者图片中的人脸身份信息、表情和肢体动作,以及生成特定人物的虚假语音.自2018年Deepfakes技术在社交网络上掀起换脸热潮开始,大量的深度伪造方法被提出,并展现了其在教育、娱乐等领域的潜在应用.但同时深度伪造技术在社会舆论、司法刑侦等方面产生的负面影响也不容忽视.因此有越来越多的对抗手段被提出用于防止深度伪造被不法分子所应用,如深度伪造的检测和水印.首先,针对不同模态类型的深度伪造技术以及相应的检测技术进行了回顾和总结,并根据研究目的和研究方法对现有的研究进行了分析和归类;其次,总结了近年研究中广泛使用的视频和音频数据集;最后,探讨了该领域未来发展面临的机遇和挑战.

关键词 深度伪造;深度伪造检测;深度学习;人脸替换;生成对抗网络

中图法分类号 TP391

收稿日期:2021-11-12;**修回日期**:2022-08-29

通信作者:纪守领(sji@zju.edu.cn)

随着深度学习(deep learning, DL)技术的飞速发展以及在各个领域的广泛应用，深度学习技术在视频以及图片的伪造编辑方面的应用也逐渐被人熟知，从早年ZAO APP提供的低成本换脸视频编辑服务促使相关视频在视频社交网络上风靡一时，到现在在教育、人机交互和艺术创作等领域中的广泛应用前景，深度伪造(deepfake)技术的应用都有一定的影响力，但是深度伪造具有不良目的的应用造成的负面影响远大于其积极影响. 2020年7月，麻省理工学院发布了一条尼克松宣布登月失败演讲的深度伪造视频，在视频中尼克松的面部表情以及语音都得到了还原，内容可以做到以假乱真的效果. 一些恶意用户可能会利用相关技术伪造政客、明星等公众人物的虚假视频内容，从而扩散谣言、引导舆论，并由此获利，同时伪造的视频可能在刑侦取证方面造成阻碍.

由于深度伪造产生较为严重的负面影响，社会各界已经开始采取相应的防护对策. 为了防止针对政治人物的伪造视频对国家安全造成影响，各国政府已经开始促进相关行业的标准和法律的制定. 同时，YouTube和TikTok等互联网公司也已经开始着手管制深度伪造视频，并举行了多次伪造视频检测比赛. 在学术界，研究者们针对深度伪造提出了适用于多种场景的大量检测技术作为深度伪造的技术治理手段. 针对近年来出现的伪造技术和检测技术，本文阐述了其中具有代表性的技术，与现有的其他综述[1]相比，更加系统地考虑了不同模态信息的深度伪造及检测技术，同时也介绍了深度伪造生成及检测模型的对抗攻击方法.

1 技术背景

1.1 深度伪造生成技术

因为深度伪造生成的各种技术之间存在一定的共通性，因此本节对深度伪造生成技术的生成模型进行总结，并介绍深度伪造技术中人脸伪造技术和语音伪造技术的基本步骤.

1.1.1 人脸伪造生成技术

针对人脸的深度伪造的生成技术一般包含4个步骤：1)使用人脸识别算法检测目标图片中的人脸；2)裁剪并预处理目标图片中的人脸；3)提取人脸中的身份和表情信息，并通过生成模型生成伪造人脸；4)将生成的人脸渲染到目标图像中人脸位置，重建图像.

1.1.2 语音伪造生成技术

针对语音模态的深度伪造技术一般需要音频和文字输入，用于指定目标语音的内容和音色. 语音伪造技术一般包含4个步骤：1)如果该方法接受文本输入，则将文本编码；2)提取输入音频的梅尔倒谱系数(mel-frequency cepstral coefficients, MFCCs)；3)将预处理后的数据输入生成模型，得到目标语音的帧级语音特征；4)通过声码器等方式得到目标语音.

1.1.3 深度伪造生成模型

1)自动编码器(auto encoder, AE). 自动编码器是一种无监督的神经网络模型，一般含有一个编码器E和一个解码器D，编码器会将输入的特征编码到维度较低的隐空间，并用解码器尝试将特征进行解码和重建. 在训练过程中，对于任一输入$\boldsymbol{x}$，使输出$D(E(\boldsymbol{x}))=\boldsymbol{x}_o$与输入$\boldsymbol{x}$尽量相近，即优化目标为尽量缩小重建损失. 在很多图像处理任务中可以利用自动编码器中的解码器来重建图像，或者通过编码器进行降维. 在深度伪造任务中可以通过添加解码器和编码器或者在编码中注入信息来实现身份或者动作篡改.

2)变分自动编码器(variational auto encoder, VAE)[2]. 变分自动编码器是由自动编码器发展而来的变种，拥有2个编码器E_1和E_2，分别拟合输入对应的隐空间中高斯分布的均值$\mu=E_1(\boldsymbol{x})$和方差$\mathrm{lb}\sigma=E_2(\boldsymbol{x})$. 解码器根据编码器拟合的后验分布取样并重建输入. 相较于传统的自动编码器，变分自动编码器有更好的耐噪声能力.

3)生成对抗网络(generative adversarial network, GAN)[3]. GAN是一种应用更为广泛的生成模型，其包含生成器G和分类器C. 生成器在高斯分布中采样生成图片，训练目标是能生成分类器无法分辨真伪的图片. 而分类器则学习辨认图片是否由生成器生成. 在训练过程中，交替固定生成器和分类器的参数，训练模型的另一部分. 在训练完生成器后，网络的损失函数$V(G,D)$逼近当前分布与真实分布的JS散度，再通过训练生成器最小化2个分布之间的距离. 另一种网络结构Wasserstein GAN采用Wasserstein距离来衡量生成图像分布和真实图像分布之间的距离，提高了GAN训练的稳定性和生成图像的多样性. 自从GAN诞生之后，陆续出现了CGAN, InfoGAN等变种，其中运用在图像生成领域中的主要有pix2pix[4]和CycleGAN.

pix2pix[4]是一种类似于条件性对抗生成网络的图片生成技术，可以实现图像在2个分布间的单向转换. 与CGAN相似，pix2pix会向生成器输入一个图

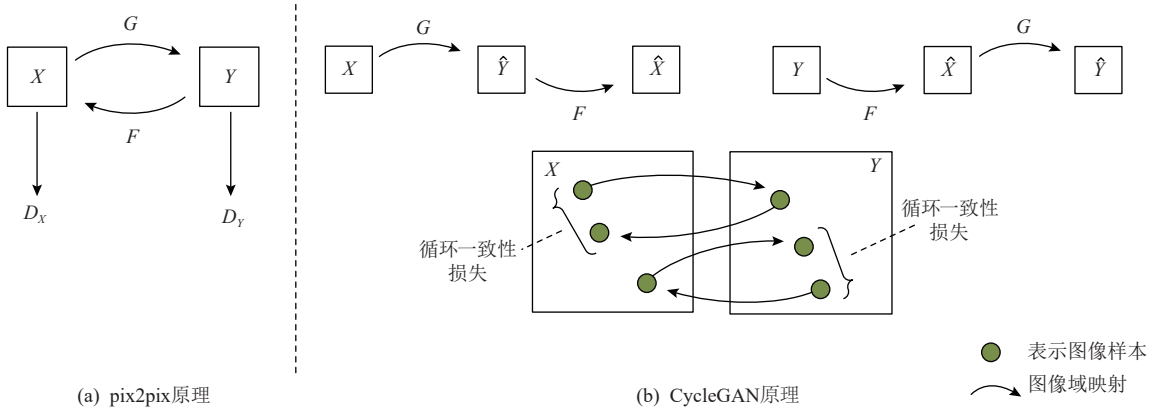
片 $\boldsymbol{x}$ 作为条件, 生成器根据 $\boldsymbol{x}$ 生成 $\boldsymbol{y}$, 而分类器则输入($\boldsymbol{x}$, $\boldsymbol{y}$), 鉴别 $\boldsymbol{y}$ 是否由生成器生成. 同时由于常规 GAN 输入的随机高斯噪音会被淹没在条件信息中, 因此, pix2pix 的生成器不需要输入噪音. pix2pix 的生成器使用了 U-net[5] 的编解码器结构, 判别器采用 PatchGAN 结构, 将图片分为 $N\times N$ 个区域并分别判别这些区域的真伪, 可以检测到更多的高频细节. 近年来也出现了 pix2pixHD[6], vid2vid[7] 等改进版本, 从模型和损失函数设计等方面进行优化以生成高分辨率的图片和视频.

CycleGAN[8] 是一种基于 pix2pix 改进的风格迁移技术, 与 pix2pix 需要成对的训练数据不同, CycleGAN 并不需要在源域和目标域间一对一成对地训练数据. 针对 2 个图像域 X 和 Y, CycleGAN 包含 2 个生成器 G 和 F, 分别处理 X 到 Y 的映射和 Y 到 X 的映射, 2 个分类器 D_X 和 D_Y 分别用来辨别图片是否来自 X 和 Y. pix2pix 和 CycleGAN 的原理对比如图 1 所示. CycleGAN 的损失函数主要包括对抗损失函数和循环一致性损失函数:

$$L_{\mathrm{ADV}}(G,F,D_X,D_Y,X,Y)=L_{GAN}(G,D_Y,X,Y)+L_{GAN}(F,D_X,X,Y), \tag{1}$$

$$L_{\mathrm{CYC}}(G,F)=E_{\boldsymbol{x}\sim Pdata(\boldsymbol{x})}[\|F(G(\boldsymbol{x}))-\boldsymbol{x}\|_1]+E_{\boldsymbol{y}\sim Pdata(\boldsymbol{y})}[\|G(F(\boldsymbol{y}))-\boldsymbol{y}\|_1], \tag{2}$$

其中 $Pdata(\boldsymbol{x})$ 和 $Pdata(\boldsymbol{y})$ 分别为图像域 X 和 Y 的样本分布, $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 为其中采集的样本.



(a) pix2pix原理 (b) CycleGAN原理

Fig. 1 Illustration of principle of pix2pix and CycleGAN

图 1 pix2pix 和 CycleGAN 的原理示意图

1.2 深度伪造分类

目前深度伪造在多媒体文件的合成和篡改中主要是针对图片和视频中的面部和肢体动作的篡改以及语音方面的修改. 面部篡改主要分为 2 类: 1)使用源图片人物身份替换目标图片人物身份的方法, 主要包括面部替换和面部转换 2 种; 2)保留目标图片人物身份的方法, 包括面部重现和面部特征编辑. 本文分别对这 4 种分类进行介绍.

1)面部替换(face swap). 通过将源图片的人脸身份信息注入目标图片完成换脸, 同时保留目标图片中的表情动作和背景等信息.

2)面部转换(face transfer). 将目标图片中的人脸完全替换为源图片的人脸, 包含身份信息、表情动作和面部朝向等.

3)面部重现(face reenactment). 不改变目标图片中的人脸身份, 将源图片中的人脸表情动作在目标图片上重现. 重现的部分包括表情、嘴型、面部朝向, 甚至是肢体动作.

4)面部特征编辑(face feature edit). 不改变人脸身份信息, 对人脸中的部分特征属性进行篡改, 如头发颜色、性别、是否佩戴眼镜等.

语音方面的深度伪造任务主要分为语音转换和语音合成 2 类.

1)语音转换(voice conversion). 转变输入音频的音色到目标人物的音色.

2)语音合成(text to speech). 根据输入文本, 输出相应音频.

1.3 深度伪造检测技术

1)卷积神经网络(convolutional neural network, CNN). CNN 除了在风格迁移方面的应用, 更多地用于图像分类和目标检测中. CNN 在传统全连接网络的基础上主要增加了卷积层和池化层. 后续出现的 CNN 架构有 DenseNet[9], ResNet[10] 和基于深度可分离卷积层的 Xception[11] 等, 而 XceptionNet[12] 也是深度伪造检测中常用的基础方法.

2)循环神经网络(recurrent neural network, RNN). RNN 常用于处理时间序列性的信息, 对于时刻 t 的输入 $\boldsymbol{x}_t$, 输出 $\boldsymbol{y}_t=f(\boldsymbol{x}_t,\boldsymbol{h}_{t-1})$, 其中 $\boldsymbol{h}_{t-1}$ 为时刻 $t-1$ 状态

下向时刻 t 的输出. 相比一般的前馈神经网络, RNN 能更好地处理序列变化的数据, 并记录过去状态对当前状态的影响. 长短期记忆网络(long short-term memory, LSTM)是一种特殊的循环神经网络, 它包含 2 种传递状态, 能够更好地解决 RNN 的梯度消失和梯度爆炸问题.

2 图像和视频伪造生成技术

现有的视频和图片的深度伪造技术主要是针对人脸信息或表情动作的篡改, 也有部分工作可以重现人物的肢体动作. 本节对人脸和身体的主流图像伪造手段进行分类阐述, 并简要介绍其发展历程.

2.1 面部替换伪造

传统的面部替换主要是基于图形学的伪造, 通过 3D 人脸模型的重建以及追踪等技术实现人脸的替换. 近十年的图形学图片人脸替换方法逐步实现了全自动, Dale 等人[13] 用 3D 多线性模型追踪源视频和目标视频的面部表现, 使用相应的 3D 图形将源视频匹配到目标视频中实现自动化人脸替换.

近年来, 随着深度学习的迅猛发展, 基于深度学习的换脸技术的时间成本和门槛逐渐降低, 以深度学习为基础的面部替换技术得到了更广泛的应用, 也推动了面部替换方法的研究. 早期的面部替换技术主要基于自动编码器. Deepfakes[14] 是一种被 reddit 用户使用的深度伪造换脸工具, 它基于自动编码器, 包含 1 个编码器和 2 个解码器, 组成 2 个自动编码网络, 分别用源人物和目标人物的面部图片同时训练, 以实现解码器分别重现人脸的能力. 在换脸过程中交换编码器, 实现在目标人物的图片编码中提取出源人物的人脸. Fast Face-swap[15] 是一种基于 CNN 的风格转换网络的换脸方法, 通过神经网络标注人脸特征点, 实现背景分割和人脸对齐, 采用 Texture networks[16] 的 CNN 结构实现风格迁移任务.

除了自动编码器, GAN 等生成模型也被应用到面部替换中, 极大提高了生成图片的图像质量. FaceswapGAN[17] 是 Deepfakes 融入 GAN 的产物, 引入了去噪自动编码器和面部交换注意力机制, 提高了图片的真实程度, 同时通过生成分割掩码解决图像的遮盖问题. Natsume 等人[18] 使用 VAE-GAN 网络结构, 引入 3 重损失函数验证身份信息损失, 实现了较为稳定的人脸替换; 之后又提出了人脸替换和人脸面部特征编辑的集成系统 RSGAN[19], 使用 2 个自动编码器在隐空间中分别表示头发区域和面部区域, 通过替换面部的隐空间表示来实现换脸并重建整个人脸图像, 能够解决之前换脸方法如 3D 变形人脸模型的人脸朝向和光照不匹配等问题. FSGAN[20] 通过基于 RNN 的方法将目标人脸的表情和面部动作重现给源人脸, 实现了较好的泛化能力, 并覆盖了表情迁移和身份替换 2 个任务, 可以使用较少的样本进行训练. Li 等人[21] 提出了一种新的 2 阶段的换脸方法 Faceshifter, 其模型结构如图 2 所示, 第 1 阶段在解码器中自适应注意力去正化(AAD)、自适应地集成人脸合成的特征和属性, 同时在第 2 阶段引入了启发式的错误承认细化网络(HEAR-Net), 以自监督的方式解决面部遮挡问题. Simswap[22] 使用身份注入模块解除身份限制, 并在损失函数中引入弱特征匹配损失.

随着小样本学习的方法的不断出现, 为了解决训练样本难以获得的问题, MegaFS[23] 通过分层表征人脸编码器, 提取更多人脸特征, 并在不经过特征解耦的情况下非线性地将身份信息从源图像迁移到目标图像. 同时 MegaFS 可以分模块训练, 可以适用于

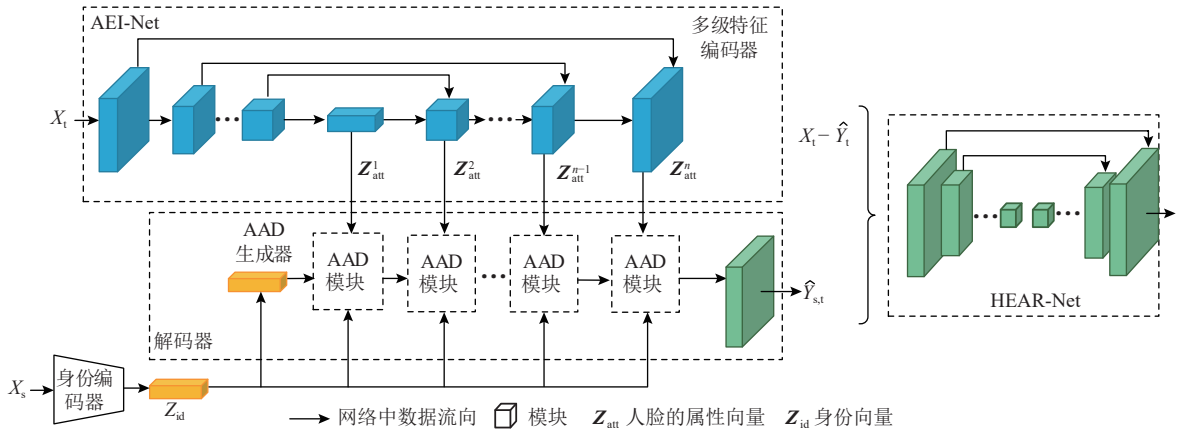


Fig. 2 Schematic diagram of the network structure of Faceshifter[21]

图 2 Faceshifter 网络结构示意图[21]

百万像素级图片的面部伪造.类似的小样本学习和单样本学习的应用可以降低深度伪造的应用门槛，提升相关工具的易用性.

2.2 面部转换

相较于面部替换，面部转换不保留目标图片的表情和面部动作信息，因此在伪造方面的灵活性较低.早期的面部转换主要采用面部可变形模型实现[24-25]，也有少量面部转换使用了深度学习模型. DepthNet[26]是一种根据检测源图像中人脸关键点深度来构建3D人脸模型，并使用孪生神经网络将3D人脸模型映射到目标图像的2D标志点模型的无监督方法，该方法对于输入图像中的遮挡部分较为敏感.

2.3 面部重现

面部重现相较于面部替换可以更加自由地将指定的面部动作迁移到特定身份的人脸中，因此对于公众人物的攻击更具威胁性.本节将主流的面部重现技术分为面部表情重现和嘴部动作驱动2类，并分类阐述其相关技术.

2.3.1 面部表情重现

与其他深度伪造技术相似，在深度学习获得大规模应用之前，表情重现的相关方法也围绕着3D人脸模型展开. 2015—2018年，Thies等人[27-29]提出了一系列基于3D模型重建的表情重现方法，逐步实现了实时表情迁移、视频级表情转移和包含头部移动的重现，都取得了较好的伪造效果，其中的Face2face方法[28]也常用于后来的伪造视频数据集的构建.

深度学习及相关生成模型的普及带来了更高效的表情重现方法.早期表情驱动深度伪造直接通过人脸的特征表示，如对3D模型或特征点等进行伪造，而神经网络模型负责后续的渲染以及真实度优化. Kim等人[30]提出了一种通过3D模型进行特征表示，并实现3D模型转换为人脸视频的伪造方法，且首次实现了将整个3D头部位置、角度、面部表情和眼睛动作全部转移到目标视频中.其他使用3D模型作为特征表示的方法还有PaGAN[31]，但其生成结果真实性欠佳. Geng等人[32]提出了一种基于人脸扭曲的伪造方法，使用Wg-GAN和Hrh-GAN来优化伪造结果的真实感和隐藏细节.随着深度生成模型的不断发展，2019年以后出现的伪造方法采用了更为复杂的模型结构，深度学习在整个方法中的比重不断增加. Imaginator[33]提出了时空特征融合机制，通过图片编码的空间特征和动作外加噪声可以解码出连续视频.同时，该方法还采用了2个判别器，分别判断生成图片中人脸外观是否真实、生成视频动作是否真实. Siarohin等人[34]提出的Monkey-Net将图像中的外观信息和动作信息解耦，实现了任意物体的动作驱动.该方法主要分为3个部分：动作转移网络、无监督关键点检测器和动作预测网络.该方法通过检测目标图像和源图像的关键点来预测每个关键点的视觉流图，并由此生成伪造图像.文献[35]是在Monkey-Net基础上的改进方法，引入关键点附近的局部仿射变换，针对Monkey-Net难以重现较大姿态变化的问题进行了优化. Qian等人[36]针对高分别率图像的伪造提出了附加焦点变分自动编码器(AF-VAE). Song等人[37]通过无监督的纹理合成实现表情的迁移，但该方法并非针对人脸. Pumarola等人[38]提出了一种基于动作单元标注的方法GANimation，采用无监督的训练策略，并通过注意力机制提高了模型鲁棒性. FACEGAN[39]使用运动单元表征面部表情，分别处理人脸和背景来提高伪造结果的质量，同时相较于已有方法减少了源图像和目标图像身份不同时的身份信息泄露. Gu等人[40]提出的FLNet可以利用多张源图片进行说话人脸合成，从而解决使用单张图片时容易出现的细节遮挡问题.

多数表情伪造方法的网络结构和损失函数设计都是基于CycleGAN[8]设计的. Xu等人[41]提出了一种基于CycleGAN的端到端面部重现方法，并通过实验验证多种规格的感受野对图像生成的影响，以及采用PatchGAN来提高生成图片的质量. RecycleGAN[42]是另一种基于CycleGAN的数据驱动的无监督视频重定向方法，可以实现源图像域的连续信息转移到另一个域中，从而实现表情驱动. RecycleGAN还提出只使用CycleGAN的空间循环一致性约束容易导致感知模式崩溃，因而通过设计周期性预测器及相应的损失函数来引入时空约束，达到更好的视频生成效果. ReenactGAN[43]使用一个编码器从图像中抽取人脸轮廓，并通过CycleGAN完成源人脸轮廓到目标人脸轮廓的映射，最后由pix2pix生成器重建图像.同时该方法全部采用前馈神经网络，可以实现实时的表情重现. FReeNet[44]包含统一标志点转换器和一个类似CycleGAN的几何感知生成器，转换器将源图像和目标图像编码合并得到移动标志点，并用其指导生成器生成伪造结果，实现针对任意身份的伪造. FaceSwapNet[45]与FReeNet采用类似结构，是一种多对多的伪造方法. Tripathy等人[46]提出的ICface方法分别使用头部姿态参数和面部动作单元表示头部姿态和表情动作，并提出了一个二段式的对抗神经网络用于摆正目标图像人脸和用新的头部动作和面部

表情对其进行重现，采用了类似 CycleGAN[8] 结构的生成器和类似 PatchGAN[41] 结构的判别器. 由于该方法训练过程中使用的源图像和目标图像是来自同一视频的不同帧，因此又提出了基于像素的重构损失函数. 与之类似的方法有 X2Face[47]，该方法训练过程分为 2 个阶段：第 1 阶段使用同一视频中的不同帧进行重现；第 2 阶段开始使用不同身份人脸图片并引入身份损失函数.

部分表情重现方法不通过驱动的方式重现表情，而是直接对目标人脸表情进行编辑，此类方法生成的伪造结果通常具有较高的多样性. Shen 等人[48] 提出的 Faceid-GAN 将伪造方法中常用的身份分类器引入到 GAN 的对抗性结构中，保证生成图片的真实性和身份一致，并用信息对称性的设计降低了训练的复杂度. Shen 等人[48] 提出的另一种伪造方法 FaceFeat-GAN[49] 分为 2 个生成阶段：1）生成人物身份、表情动作和一般特征等特征向量；2）通过这些特征向量生成伪造结果，在 Faceid-GAN 基础上提高了生成结果的多样性.

为了解决面部重现的训练需要大量样本的问题，2019 年以后出现了一些小样本学习方法，可以针对只有少量目标样本的场景中进行伪造. Wang 等人[50] 提出了一种基于 vid2vid 的小样本学习方法，采用了注意力机制生成网络权重来增加泛化能力，可以通过目标的少量示例图像学习合成未见过的主题或者场景，并进行了广泛的实验验证. Zakharov 等人[51] 提出的方法包含一个编码器、生成器和判别器，针对视频 v_i，编码器从多个帧提取出人物信息向量 $\boldsymbol{e}_i$，由生成器结合目标表情的面部标志 L 以及 $\boldsymbol{e}_i$ 生成伪造结果 x_{g}，分类器辨认 x_{g} 中的表情和任务是否符合 L 和 v_i. 该模型首先在一个较大的视频数据集上进行元学习训练，此阶段只会更新生成器中的部分参数，用以学习人脸合成的一般特征. 随后在特定目标人脸的少量样本中进行训练，更新生成器的剩余参数，以实现对未见过人脸的小样本学习. Burkov 等人[52] 在文献 [51] 的基础上进行改进，使用隐空间向量代替面部标志点，并为模型添加了识别前景的语义分割能力. Ha 等人[53] 提出的 MarioNETte 通过使用标志点转换器，解决了表情重现任务中经常出现的源人物身份信息保留问题，并提出了不同于传统方法的使用图片注意力块和目标特征对齐的方法来提取人脸身份信息. Hao 等人[54] 提出了 SPADE（空间自适应归一化）模块来替代常用的表情信息注入方法 AdaIN（自适应实例标准化），并通过自注意力机制提升 GAN 生成图片的质量.

2.3.2 嘴部动作驱动

在面部表情重现中的大部分工作中都涉及到了嘴型的转移，因此本节介绍的针对嘴部的深度伪造生成技术主要是通过文字或者音频生成对应的嘴型. Fried 等人[55] 通过重组输入视频的语料完成指定词语的增删和更改，利用 Deep video portraits[30] 生成嘴型，并渲染回原视频中，从而更改视频语音内容. ObamaNet[56] 是最早的由输入样本生成相应视频和音频的伪造方法，其音频生成方法基于 Char2wav[57]，通过延时 LSTM 网络生成的嘴部关键点驱动基于 pix2pix 的生成网络生成相应的视频帧.

相比于使用文字驱动嘴型，更多研究选择使用语音驱动目标人物合成说话人脸. Jamaludin 等人[58] 提出了一种使用音频和目标人物静态图片合成视频的伪造方法 Speech2Vid，该方法接受音频和目标人物的若干静态图片作为输入，分别对其编码，合并后解码得到输出视频. Vougioukas 等人[59] 提出了一种基于 GAN 的使用音频驱动目标人物静态图片的端到端方法，该方法的 GAN 中包含 3 个分类器，分别检测生成结果的音视频同步、表情连贯和人脸身份信息损失. Suwajanakorn 等人[60] 通过 RNN 实现音频到嘴型的转换，再通过 3D 模型生成高质量的面部纹理和帧间流畅的纹理变换，最后将嘴型和音频生成伪造视频. Chen 等人[61] 采用了与文献 [60] 相似的结构，使用带卷积的 RNN 生成视频帧，并设计了一个基于注意力的动态损失函数，该函数和 GAN 的对抗性损失函数一起分别训练网络的 2 个部分. Zhou 等人[62] 提出了一种能够通过音频或者视频驱动目标人物嘴型的方法，通过将输入音视频中包含不同语义信息的域解耦，提取出目标人物身份信息和音频语言信息；通过时序 GAN 生成伪造结果. Thies 等人[63] 提出了一种根据任意音频生成说话人的人脸视频的方法，该方法首先通过 DeepSpeech[64] 循环神经网络提取语音特征，又训练了一个带有时域滤波器的语音表情生成网络驱动目标任务的 3D 模型，生成伪造结果.

2.4 面部特征编辑

面部特征编辑是较为传统的伪造类型，一些较为常用的基于 GAN 的图像处理方法，如 StyleGAN[65-67] 和 CycleGAN[8] 都可以用于编辑面部特征. StarGAN[68] 和 StarGAN v2[69] 具有在多个图像域之间转换的能力，具有更好的可扩展性. 其他的面部特征编辑工作有如 Sanchez 等人[70] 提出了一种用于 GAN 人脸编辑的 3 重连续性损失函数，并提出了一个直接编辑人脸表

情的合成方法 GANnotation. Kim 等人[71]在 CycleGAN 的循环一致性损失函数的基础上提出了 CAM 一致性损失函数，使得模型能够更好地保留与特征无关的位置的信息，并将其应用在 StarGAN 等现有生成模型上. Li 等人[72]为了解决人脸特征编辑的扩展性和多样性问题，提出了一种层次结构模型——HiSD (hierarchical style disentanglement)，将人脸的特征建模成标签和属性，并通过无监督的方法将其解耦，实现针对目标属性更加精准的篡改. 随着如 StyleGan3[67]这样的大型模型的提出，面部特征编辑任务得以向更注重细节纹理的方向发展.

2.5 人体动作伪造

部分深度伪造生成技术研究将源人物的肢体动作迁移到目标人物身上. Aberman 等人[73]提出了一种视频动作克隆技术，分别使用成对训练数据和非成对训练数据训练同一个生成网络；分别训练其根据指定动作生成静态帧和将动作转换成时序连续的帧序列的能力. Everybody Dance Now[74]是一种基于视频转换的动作迁移方法，使用动作探测器检测输入视频中的人物动作骨架图，再通过基于 pix2pix 的 GAN 网络将人物动作骨架图映射为目标人物的动作帧. 在训练过程中，要将生成视频和动作骨架图的连续 2 帧输入到 GAN 的判别器中，从而保证视频的时序连贯. 文献 [74] 的方法还包括一个针对面部的 pix2pix 网络，通过动作骨架和生成视频帧的人脸区域预测残差，增加伪造结果面部的细节和真实性. Liu 等人[75]对目标人物进行 3D 建模，从源人物视频中提取动作骨架，将其渲染到目标人物的 3D 模型中，最后根据条件性 GAN 得到预测结果. 该方法的损失函数计算用到了注意力图谱的加权方法，促进 GAN 注重于包含更多未学习特征的区域，使得该方法在生成结果真实性和性能方面有所提升. Monkey-Net[34]除了重现面部表情，更多地被利用在肢体动作的驱动.

3 语音伪造技术

3.1 语音生成技术

语音生成技术的实现主要基于 2 种方法：波形拼接和统计参数. 其中波形拼接方法是早期常用的方法，它首先分析文本以及韵律，再进行波形拼接. 虽然波形拼接方法使用了自然语音波形，可以合成出高自然度的合成语音，但是对于不同领域的文本合成，稳定性不强，在任意文本的语音合成中表现不佳. 统计参数方法可以分为基于隐马尔可夫模型(hidden Markov model, HMM)的早期方法以及基于神经网络的深度学习方法. 基于隐马尔可夫模型的方法，其相关工作可以参考文献 [76]；而基于深度学习的语音生成方法建模更加精确，统计参数更加平滑，近年来随着神经网络相关技术的发展其得到了更广泛的应用. 基于深度学习的语音合成主要可以分为管道式和端到端式. 其中传统的语音合成工作一般是管道式，需要对整个合成过程中的各个模块分别建模，使用多个模型流水线式地处理文本特征分析、声学特征分析和声音波形预测等任务.

端到端式语音合成与管道式语音合成相比不需要另外提取文本特征，可以直接输入未处理的文本，得到接近自然人物声音的合成结果. WaveNet[77]是一种早期的端到端的语音合成器，它使用扩展因果卷积层(dilated causal convolutional layers)，直接对采样值序列的映射进行学习，达到较好的语音合成效果，但由于 WaveNet 的输入是处理过的特征，并不是严格的端到端模型，因此目前一般作为声码器应用在音频伪造方法中. 其他的端到端的语音合成模型还有 Tacotron[78], Tacotron2[79], Char2wav[57] 等. Tacotron 使用一个包含一维卷积、高速网络、残差连接和双向 GRU 的 CBHG 模块提取输入文本的高层次特征，并用注意力解码器和输出解码器，分别生成语境向量和输出声谱. Tacotron2 将 Tacotron 中生成最后波形的 Griffin-Lim 算法优化成了深度学习模型，并使伪造结果更加接近自然人声，其系统结构如图 3 所示. Char2wav[57]包含阅读器和声码器，阅读器中的编码器是双向循环神经网络，用于提取文本特征；解码器是带有注意力机制的循环神经网络，用于生成声码器输入的声学特征. Fu 等人[80]在基于 Tacotron 的端到端语音合成模型的基础上，针对声学特征可能与

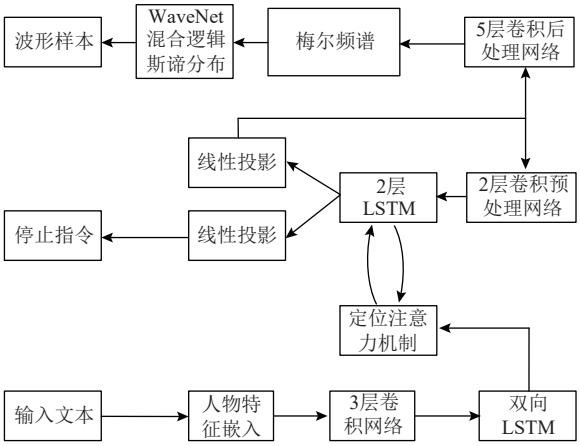


Fig. 3 Systematic structure of Tacotron2[79]

图 3 Tacotron2 系统结构[79]

文本不完全对齐的情况，提出了待反馈机制的时长控制器辅助注意力机制调整音素的方法，并使用一个自适应的优化算法用于识别标注效果较差的样本. MelGAN[81]是首个使用GAN生成语音的模型，与WaveNet等自回归的语音模型相比具有较快的速度，其改进模型[82]通过在合成中采用重构滤波器组合多个子频段的合成结果，简化了上采样层中的参数量，并引入了基于快速傅里叶变换的损失函数加速训练. 语音生成技术的不断成熟，促进了MelGAN在人机交互等领域的应用，但也使其能更容易地参与伪造视频的制作.

3.2 语音转换技术

语音转换技术使用的模型可以分为统计模型和深度学习模型. 利用统计模型的语音转换任务一般可以通过其训练的数据类型分为使用并行训练数据和非并行数据. 与语音生成技术相似，语音转换技术近年来的技术趋势也在由统计模型向深度学习转变，一些基于GAN的生成模型如CycleGAN和StarGAN也被应用到语音转换领域. CycleGAN-VC[83]在CycleGAN的基础上使用门控CNN提取连续性和层次性特征，并引入身份映射损失函数防止语义信息的流失. CycleGAN-VC2[84]较前一代的CycleGAN更新了对抗损失函数：

$$L_{adv\_2}(G_{X\to Y}, G_{Y\to X}, D_X) = E_{\boldsymbol{x}\sim P_X(\boldsymbol{x})}[lb(\boldsymbol{x})] + E_{\boldsymbol{x}\sim P_X(\boldsymbol{x})}[lb(1-D_X(G_{Y\to X}(G_{X\to Y}(\boldsymbol{x}))))]. \quad (3)$$

同时还在生成器中加入一维卷积，并将判别器改为PatchGAN结构，缩短了生成结果和真实人物语音的差距. CycleGAN-VC3[85]在前2代的基础上引入了时频自适应归一化，增加了对梅尔频谱转换的支持. StarGAN-VC[86]是另一种基于GAN的音色转换技术，能实现多对多非平行数据的语音转换任务，较传统的StarGAN加入了一致性损失，采用了GLU激活函数，并使用了类似CycleGAN-VC的判别器，在性能上优于CycleGAN-VC. StarGAN-VC2[87]是StarGAN-VC的改进版本，针对训练策略和网络结构进行优化，引入了条件性对抗损失函数和条件性实例正则化层. 近期有部分研究致力于将训练好的语音合成模型应用在语音转换任务中，如文献[88–91]一般利用迁移学习或者语音合成的预训练模型来简化非并行语音转换模型的训练.

4 深度伪造检测技术

随着深度伪造生成技术的不断发展，网络上各种相关的开源工具以及商用软件层出不穷，对司法和隐私等领域造成了严重的威胁. 为了应对深度伪造生成技术的负面影响，近年来有关深度伪造的对抗策略的研究也逐渐增多. 本节依次对图像以及视频伪造检测的主流方法进行分类介绍.

4.1 图像与视频伪造检测

根据是否针对特定的伪造痕迹进行检测，深度伪造检测可以分为数据驱动的检测方法和针对特定伪造痕迹取证的检测方法.

4.1.1 伪造痕迹的检测

当前的主流深度伪造方法产生的伪造图像和视频中，可以进行检测的伪造痕迹主要有：图像处理取证、生物信息、融合痕迹、时序连贯和模型指纹等. 早期的深度伪造检测主要是基于传统的图像取证方法，Matern等人[92]使用逻辑回归等方法，针对Deepfakes和Face2face伪造结果在全局连续性、照明估计和几何估计3个方面的视觉伪造痕迹进行检测. Zhou等人[93]提出了一种双流网络识别伪造痕迹的方法：双流网络的2个网络分别对人脸进行分类以及捕获噪声残留和相机特征，并通过2个网络的得分判断图像整体是否经过伪造. Nataraj等人[94]提出了一种根据3个颜色通道的共现矩阵检测GAN生成图片的方法. Li等人[95]设计了一种单中心损失函数以扩大真实图像和伪造图像特征分布的距离，并提出了一个根据频域特征分类的伪造检测方法. Luo等人[96]针对现有CNN网络泛化能力不强的问题，提出了针对图像高频噪声的检测方法用以提高跨数据集能力，并使用一个分别检测高频噪声流和RGB颜色流的双流检测网络，其包含3个根本组件：多尺度高频特征提取、残差引导空间注意力模块和双重交叉注意力模块. Shang等人[97]提出的像素-区域关系网络（PRRNet）通过像素关系模块和区域关系模块检测图像中空间关联和不一致伪造痕迹.

除了直接针对检测图像进行图像取证的方法，部分研究者还针对深度伪造步骤中将生成模型输出结果和外围背景的融合步骤产生的伪造痕迹进行检测. Li等人[98]针对深度伪造过程中的仿射造成的人造痕迹进行检测，并采用简化的图像处理过程而非常规的深度伪造方法生成伪造样本，减少时间成本并避免了过拟合. Li等人[99]提出了根据深度伪造面部融合步骤检测伪造边界的检测方法Face x-ray，该方法使用真实人脸生成融合人脸数据集，并通过其中的伪造边界标注训练了一个卷积神经网络，用以根据是否可以检测出边界判断真伪. 该方法的局限

在于假设了伪造生成模型都有人脸融合步骤，且会被对抗样本欺骗. Li 等人[100]提出了一种多任务学习的检测方法，使用 2 个网络分支分别检测面部区域伪造痕迹以及面部和背景的不一致性. Nguyen 等人[101]提出了一种同时鉴别目标图片或视频真伪和定位伪造区域的多任务学习方法，该方法包含一个编码器和一个 Y 型解码器，分别用于鉴别真伪和生成伪造区域. 其中解码器的训练使用半监督学习方法增强其生成能力. Nirkin 等人[102]针对深度伪造图像人脸区域和背景区域的差异，使用 3 个基于 Xception 结构的编码器分别对预处理后的整张图片、人脸区域和背景区域编码的特征向量结合后通过一个浅层分类器进行分类.

基于时序连贯性的检测方法主要是使用 RNN 对帧间的伪造特征进行检测. Amerini 等人[103]引入视频解码时运动补偿的预测误差衡量视频的局部连贯性，分别使用 CNN 和 LSTM 网络对其进行处理，在 FF++ 数据集上分别达到了 91% 和 94% 的准确率. Amerini 等人[104]通过分析目标视频的光流图检测帧间的伪造痕迹，并通过 VGG16 和 ResNet50 构建了 2 种检测模型. Guera 等人[105]通过 CNN 提取帧级特征，并将特征输入到循环神经网络中得到分类结果. Sun 等人[106]提出了一个与文献 [105] 相似的方法，将视频以帧的形式提取出关键点和特征向量序列，并输入到一个双流循环神经网络中，提取时序特征并得到分类. Sabir 等人[107]使用循环卷积网络检测伪造视频中的时序信息，并实验总结了多种 CNN 结构和 RNN 训练策略.

深度伪造生成模型的对抗训练一般更注重全局连续性和时序连贯性，但对于人脸生理信息的模拟可能更依赖于源视频和图像，因此有较多工作根据特定的人脸生理信息检测深度伪造. Agarwal 等人[108]针对公众人物的深度伪造提供了一种检测方法，使用皮尔森相关系数转换视频中人脸动作单元的特征信息，并使用支持向量机对其进行分类. Agarwal 等人[108]还提出了一种检测音频音素和唇形不同步的方法[109]. Yang 等人[110]对人物的头部姿态做 3D 模拟，分别用人脸的中心区域和外部区域计算出一个方向向量，并以此用支持向量机检测伪造视频中的头部姿态不一致. FakeCatcher[111]是一种针对生物信息的空间连贯性和时间连续性的检测方法，通过生物信息图谱重建，分别使用传统机器学习分类器和 CNN 分类器对目标视频分类，可以实现自然环境下的深度伪造检测. Fernandes 等人[112]提出了一种基于属性置信度标准[113]的检测方法.

为了增强深度伪造检测技术在不同数据集和伪造方法的泛化能力，很多检测方法针对 GAN 网络生成图片的一般性伪造痕迹进行检测，即检测生成模型的指纹. McCloskey 等人[114]提出了 GAN 生成图片和真实图片在颜色构成和饱和度方面的差异，并以此作为深度伪造检测的依据. Guarnera 等人[115]根据图像中像素关联性与生成模型之间的关系，使用期望最大化算法提取出伪造过程中生成模型的转置卷积层的特征，并用支持向量机等无监督聚类方法分类真实图像和伪造图像. Qian 等人[116]提出了一种基于频域特征的检测模型 F_3-Net，提出频率感知图像分解和局部频率统计 2 个频域特征，并设计了一种融合模块综合利用了这 2 种特征进行深度伪造检测. Masi 等人[117]使用一个双流的编码器网络，分别编码颜色域和频域加强特征.2 个编码器的结果经过 DenseNet 块输入到双向 LSTM 网络中得到分类结果. Liu 等人[118]针对伪造模型中常用的上采样步骤，提出了空间相位浅层学习（spatial-phase shallow learning）方法，验证了经过上采样生成的伪造图像和真实图像在频域中的相位频谱差异，通过浅层神经网络对两者进行区分.

近年来也出现了较多利用人物参考图片进行伪造检测的方法，该类方法可以通过比对参考图像和待测图像的身份差异帮助提高检测精确度. Agarwal 等人[119]提出了一种通过生物信息进行伪造检测的方法，使用 VGG 和 FAb-Net[120]网络分别提取动态的表情动作和静态的外观特征，根据提取到的特征分别在参考数据集中找到对应人脸，若 2 个人脸不是同一个人脸则检测视频为伪造视频，反之则为真实视频. Cozzolino 等人[121]在给定多个参考视频的人物场景中，利用自监督学习判断视频是否为伪造. 该方法通过 3D 可变形模型提取人脸信息，并用时序身份网络编码，通过判断目标视频和参考视频的编码相似性是否超过阈值来确定目标视频的真伪. Dong 等人[122]提出了一种基于身份验证的伪造检测算法 OuterFace，通过比对检测人脸和参考人脸通过面具遮罩得到的外侧人脸中的身份信息确定检测人脸是否为伪造；同时提出了一个伪造视频数据集 Vox-DeepFake，该数据集较现有数据集增加了参考身份信息，且视频内容和身份有较高的多样性. Jiang 等人[123]提出了一种基于身份空间约束（identity spatial constraints, DISC）的检测框架，它包含一个主干网络和一个身份语义编码器，身份语义编码器同时输入目标图像和参考图像，并在特定阶段以身份空间约束融入主干网络中.

随着多模态学习的普及，基于多模态特征融合的深度伪造检测方法开始出现. Lewis 等人[124]提出了一种混合深度学习方法，分别处理目标视频的图像和音频，并用多个 LSTM 网络对视频进行分类. Lomnitz 等人[125]使用 Xception 网络和双向 LSTM 网络建立图像的多帧处理模型，使用 SincNet[126]处理视频的音频信息，并结合了图像的频域特征共同进行伪造视频的分类. Mittal 等人[127]提出了一种基于孪生神经网路架构的多模态检测方法，除了利用视频的图像和音频信息，还实现了从中提取情感信息用于检测. 同样利用情感信息进行多模态检测的还有文献 [128]，该文献方法分别从视频和音频中提取面部和语音的低级描述，通过 LSTM 网络预测其中的情感语义信息，并根据两者之间的差距进行分类.

4.1.2 基于数据驱动的检测方法

随着图像处理中各种 CNN 网络架构运用的不断成熟，在深度伪造检测领域也出现了很多基于数据驱动的分类方法. MesoNet[129]针对常用的人脸伪造方法 Deepfakes 和 Face2face，提出了基于卷积神经网络的轻量级检测网络，并使用 Inception 网络结构对其进行改良. Jain 等人[130]提出一个带残差连接的卷积神经网络结构用于深度伪造检测. FakeSpotter[131]在深度伪造检测方法中引入神经元覆盖率判据，通过计算目标图像输入到深度人脸识别网络中每层的激活神经元数目得到特征向量，并训练一个浅层神经网络，根据特征向量判断输入图像是否经过伪造. Dang 等人[132]提出了一种利用注意力机制处理主干网络提取的特征并进行伪造检测分类的方法，该方法的网络输入为提取的特征 $\boldsymbol{F}$，通过伪造外观模型和直接回归 2 种方式得到注意力图谱 $\boldsymbol{M}_{\text{att}}$，并根据注意力模型的输出 $\boldsymbol{F}' = \boldsymbol{F} \odot \text{Sigmoid}(\boldsymbol{M}_{\text{att}})$ 进行伪造分类. 该方法还提供了监督学习、弱监督学习和非监督学习 3 种训练方式. Hsu 等人[133]提出了一种基于成对学习的检测方法，该方法包含一个基于 DenseNet 的一般伪造特征网络和一个分类器，前者通过基于对比学习的孪生神经网络进行自监督训练，后者进行正常的 2 分类训练. Khalid 等人[134]将深度伪造检测视为异常检测问题，在真实图片数据集上训练了一个单类变分自动编码器，通过重建图像分数判断其是否为伪造图像；还提出了一个改进结构，在原模型后新加一个变分编码器，将编码结果均值和原模型编码器输出均值的均方根差距作为损失函数和重建分数. Rana 等人[135]提出了一种集成学习方法 DeepfakeStack，并在 XceptionNet, InceptionV3 多种预训练 CNN 模型的基础上训练了一个元学习器，通过并行使用多种分类模型在 FF++数据集上达到了 99% 以上的准确率. 与之相似的有文献 [136] 的工作，该文献引入了注意力机制和孪生式训练方式. Kim 等人[137]提出了一种基于表示学习和知识蒸馏的检测方法 FReTA，该方法利用原有的预训练模型，通过迁移学习快速学习针对性能伪造方法的检测模型. Aneja 等人[138]提出了一种基于迁移学习的小样本学习方法，该方法将深度伪造检测视为不同分布，先在源域上训练一个编码器，再在目标域的较少样本上进行微调(fine-tuning)，在测试时根据测试样本在隐空间中训练得到的真实分布和伪造分布的距离进行分类. Wang 等人[139]提出了一种利用注意力机制提高卷积神经网络训练效果的框架 RFM，该方法首先用原始检测图片生成主干网络的伪造注意力图谱，再以抹除可疑区域的方式做数据增强，重新训练主干网络. Zhao 等人[140]提出一种利用多重注意力机制的细粒度分类方法，该方法同时利用了图像中的纹理特征和高层次语义特征，并设计了区域独立损失函数和注意力引导的数据挖掘方法，如图 4 所示. Kumar 等人[141]提出了一种针对 face2face 重现方法的多流神经网络方法，通过多个残差网络提取图片局部特征进行分类，并设

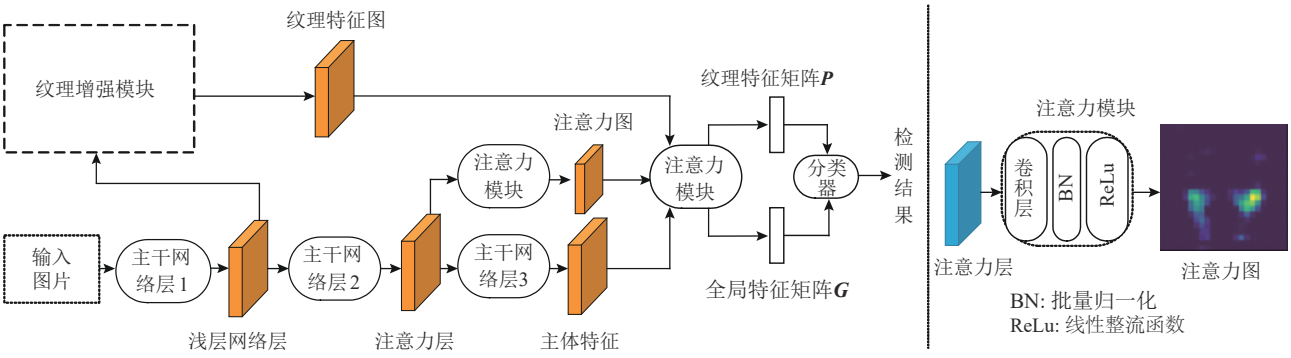


Fig. 4 Structure diagram of multi-attention deepfake detection method[140]

图 4 多注意力深度伪造检测方法结构图[140]

计了一个损失函数用来平衡多流神经网络的训练. Jeon 等人[142]提出了一种使用预训练模型 fine-tuning 的小样本学习方法，该方法提出了使用自注意力的 fine-tune 变换器，并用其预测的结果和预训练模型与 MobileNet 的组合体输出相加，输入到带全局平均池化的分类器中得到最终结果. Wang 等人[143]提出了一种 CNN 生成深度伪造的检测方法，将目前主流深度伪造模型归类为基于 CNN 的图片生成模型，试图寻找它们之间共同存在的伪造缺陷，并在单个生成模型产生的数据集上训练出可以检测更多模型伪造结果的分类器. Wang 等人[143]在一个 ProGAN 生成的数据集上实施了数据增强，并用其训练了一个 ResNet-50 网络，在其未见过的伪造模型生成的测试样本上表现出了较好的检测能力. 基于数据驱动的伪造检测容易受到训练集中伪造方法种类的影响，因此其在不同伪造方法上的泛化能力还有一定的提升空间. Liu 等人[144]提出的 Gram-Net 通过检测输入图片的纹理特征来检测伪造，该方法通过提取 ResNet 网络中不同语义级别的特征，使用基于格拉姆矩阵的 Gram-block 提取全局纹理特征. 该方法相比其他方法具有更好的鲁棒性和泛化性.

近些年随着 transformer 技术在自然语言处理和计算机视觉领域中的应用和发展，利用视觉 transformer 的深度伪造检测方法不断出现. Wodajo 等人[145]提出的 CviT 方法使用 CNN 和视觉 transformer 作为骨干网络提取特征，实现对输入人脸的分类. Wang 等人[146]提出了一种多尺度的虚假人脸检测方法 M2tr，该方法通过卷积神经网络分别提取图像空间域和经过离散余弦变换的频域特征，其中空间域特征通过基于多尺度注意力的视觉 transformer 得到新的特征；将上述 2 种特征通过多头注意力机制进行特征融合，再输入到分类器得到检测结果. 该方法由于同时考虑了空域和频域 2 个模态，并使用了多个尺度的注意力机制，具有较好的检测效果.Heo 等人[147]在视觉 transformer 为主干网络的检测模型中引入了知识蒸馏方法，并使用基于卷积神经网络的 EfficientNet 作为教师模型进行了验证.

表 1 对近年来主流的图像和视频伪造检测方法的特点和性能等进行了总结.

4.2 深度伪造模型及其检测的对抗攻击

随着深度伪造模型及其对应的检测技术不断趋于完善，部分深度伪造模型及其深度伪造检测的工

Table 1 Summary of Image and Video Fake Detection Methods

表 1 图像和视频伪造检测方法总结

检测方法	特点	适用场景	实验数据集	检测性能	模型主干网络
Exploiting Visual Artifact[92]	通过提取牙齿，眼睛及脸部轮廓等特征进行伪造检测	使用 Deepfakes 方法和 face2face 方法生成的深度伪造视频	FaceForensics	0.866（AUC）	逻辑回归、多层感知机
FDFL[95]	使用频域特征，优化难度小	检测面部替换，面部重现等伪造图片和视频	FaceForensics++[12]	0.994（ACC） 0.997（AUC）	CNN
Generalizing Face Forgery Detection[96]	利用图像的高频噪声，泛化能力较强	针对未知伪造方法生成图像的检测，需要高泛化性检测方法的场景	FaceForensics++[12]	0.994（AUC）	CNN、注意力机制
Face x-ray[99]	较高的泛化性	需要高泛化性检测方法的场景	FaceForensics++[12]，DFDC[148]，celebDF[149]	0.985（AUC，FF++），0.806（泛化 AUC，celebDF），0.809（泛化 AUC，DFDC）	CNN
LRNet[106]	通过帧间时序特征识别伪造视频，同时有较强的鲁棒性	针对存在压缩和破损等情况的深度伪造视频检测	FaceForensics++[12] celebDF[149]	0.957（AUC，FF++，c40 压缩）0.554（AUC，celebDF，c40 压缩）	CNN+RNN
Exposing Inconsistent Head Poses[110]	通过检测人物头部姿态判断是否为伪造视频	深度伪造视频检测	自建数据集	0.974（AUC）	SVM
F_3-Net[116]	基于频域特征的深度伪造检测	被压缩的伪造视频检测	FaceForensics++[12]	0.958（AUC）	CNN
Two-branch Recurrent Network[117]	融合了 RGB 域信息和频域的高频信息	深度伪造视频检测	FaceForensics++[12]，DFDC[148]，celebDF[149]	0.987（AUC，单帧），0.991（AUC，视频）	CNN+LSTM
Id-reveal[122]	通过比对待测视频和参考视频中人脸身份信息判断伪造	拥有指定人物参考视频的深度伪造视频检测	DFD[150]	0.86（AUC）	CNN
Emotions Don’t Lie[127]	通过提取多模态情感信息之间的差异来检测伪造	带有音频的深度伪造视频检测	DF-TIMIT[151]，DFDC[148]	0.844（AUC，DFDC）	CNN
Fakespotter[131]	通过神经网络可解释性方法检测伪造视频	针对 GAN 等生成模型的深度伪造检测	Celeb-DF v2[152]	0.668（AUC）	深度人脸识别模型

检测方法	特点	适用场景	实验数据集	检测性能	模型主干网络
On the Detection of Digital Face Manipulation [132]	基于注意力机制的深度伪造检测	需要可视化伪造区域的检测场景	自建数据集	0.997（AUC）	CNN、注意力机制
FReTal[137]	通过知识蒸馏和迁移学习，解决针对新出现的伪造方法的检测	适用于检测较新的伪造生成方法	FaceForensics++[12]	0.925（泛化 AUC）	CNN
Multi-attentional deepfake detection[140]	聚合高维的语义信息和低维的纹理信息	图像和视频深度伪造检测	FaceForensics++[12]，DFDC[148]，celebDF[149]	0.993（AUC，FF++）	CNN、注意力机制
CviT[145]	引入视觉 transformer 检测深度伪造	图像和视频深度伪造检测	FaceForensics++[12]，DFDC[148]	0.915（ACC，DFDC）	CNN+视觉 transformer

作选择针对模型进行对抗攻击，从而促使深度伪造模型误分类或者深度伪造结果的真实性降低.

4.2.1 针对伪造模型的对抗攻击

Ruiz 等人[153] 提出了一种通过在目标图像施加扰动使深度伪造模型失效的攻击方法，他们将对抗攻击模型视为图片翻译模型，提出了条件性和非条件性的图片翻译干扰，并给出了基于对抗训练的初步防御方式. Huang 等人[154] 提出了一种针对深度伪造的主动防御方法，该方法分为替代模型和扰动生成器 2 个部分，其中替代模型首先使用正常人脸训练，目标是模拟深度伪造模型，扰动生成器随后训练在原始人脸上施加扰动并以替代模型的反馈作为损失函数. 该方法还针对特定的伪造方法提出了加强训练方案. Dong 等人[155] 提出了 3 种针对深度伪造的对抗攻击方法，即转移对抗攻击、孪生对抗攻击和隐式孪生对抗攻击.

4.2.2 针对检测模型的对抗攻击

由于现有的深度伪造模型的鲁棒性还有提升空间，因此部分研究者针对深度鉴伪模型进行对抗攻击，从而使伪造图片和视频能够绕过鉴伪模型的检测. 但由于该类工作有潜在的不良社会影响，因此相关论文数量较少. Neves 等人[156] 训练了一个自动编码器用于抹除 GAN 生成深度伪造结果中残留的模型指纹，且该模型只需要在真实人脸数据集上训练. Carlini 等人[157] 选择了 3 种深度检测分类器，在黑盒和白盒 2 种场景对其进行对抗攻击，均能有效降低检测模型分类的 AUC 值. Hussain 等人[158] 针对在黑盒场景和白盒场景下对 XceptionNet[11] 和 MesoNet[129] 2 种深度伪造模型进行对抗攻击，证明了现有的检测模型的对抗鲁棒性较为脆弱，其网络结构如图 5 所示.

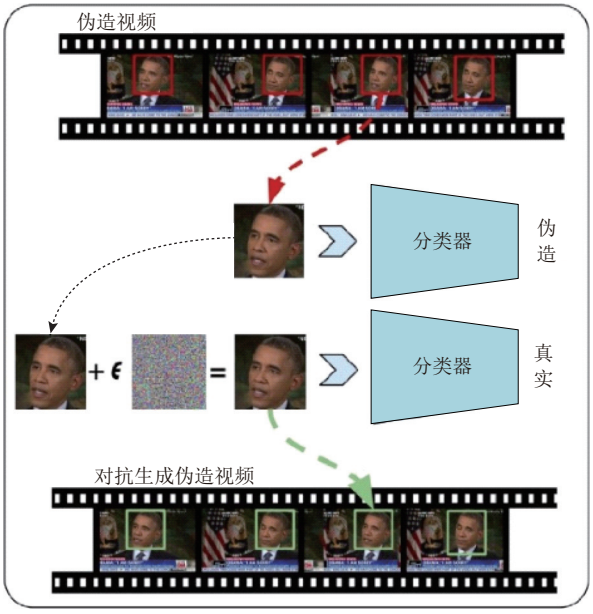


Fig. 5 Network structure of deepfake adversarial examples generation method[158]

图 5 深度伪造对抗样本生成方法的网络结构[158]

4.3 语音伪造检测

随着各种语音深度伪造模型的伪造水平不断提升，现有的技术已经可以较好地模拟目标人物的音调音色，甚至配合视觉深度伪造模型生成一个完整的伪造视频，相应地，音频的深度伪造检测工作也逐渐受到了更多研究者的关注. 语音深度伪造检测一般分为前端和后端，分别从音频中提取声学特征，以及利用高斯混合模型、神经网络等分类模型根据声学特征对目标音频进行分类. 部分研究者选择从声学特征的角度开始研究，Patel 等人[159] 在 ASVspoof 2015 竞赛中基于耳蜗过滤器倒谱系数（cochlear filter cepstral coefficients, CFCC）和瞬时频率（instantaneous frequency, IF）提出了新的声学特征 CFCCIF，并在论文中提出了更新的特征 CFCCIFS. Tom 等人[160] 提出了群体延迟图（GD-gram）作为重放检测的判断依据，分别将 GD-gram 和使用注意力遮罩的 GD-gram 输入 ResNet-18 得到判断结果. Das 等人[161] 提出了 2 种用于检测模型前端的声学特征 eCQCC 和 CQSPIC.

其他的研究者更注重于分类模型. Lavrentyeva 等人[162] 在 ASVspoof 2017 比赛中针对录音重放挑战提出了一种基于轻量级 CNN 检测的方法，该方法的前

端基于赛事举办方提出的常数 Q 变换倒谱系数(CQCC)提取系统，通过常数 Q 变换和快速傅里叶变换得到归一化对数功率谱，并将其输入到轻量级 CNN[163] 或者 CNN 和双向 RNN 的串联网络中. 同样采用轻量级 CNN 的有参加了 ASVspooof 2019 比赛的 Lavrentyeva 等人[164]. Mittal 等人[127]提出了基于 CNN 的 SincNet 直接对音频进行处理，与标准 CNN 不同，SincNet 的第 1 层使用了预定义的滤波函数，仅有少量几个参数可以从数据中学习，这大幅度减少了网络中的可训练参数数量. Cai 等人[165]提出了一种语音级的神经网络框架，并使用了多种声学特征表示作为模型的输入进行实验. 在 ASVspooof 2019 比赛中，Lai 等人[166]提出了一种基于残差网络和挤压刺激网络(squeeze excitation network, SENet)的检测方法挖掘不同通道之间的关系，使其关注更具判别性的特征图. Lai 等人[166]考虑了对数功率频谱和 CQCC 2 种声学特征，通过使用统一特征图或整条语句的方式将其输入到神经网络模型中. 与其工作类似的有 Parasu 等人[167]提出的轻量级残差网络模型. Ma 等人[168]使用基于规范化方法的持续学习，在损失函数中添加了 LwF[169] 约束和正样本对其约束，其模型结构采用类似于文献 [164] 的轻量级卷积神经网络.

5 深度伪造数据集

5.1 深度伪造视频数据集

近年来，针对深度伪造的检测方法逐渐完善，学术界和工业界也发布了大量的数据集用于训练和评价检测模型. 表 2 总结了近年来发布的主流数据集. 如谷歌提出的 DFD(detection)[150] 和 FaceForensics++[12] 数据集、商汤科技提出的 Deeper Forensics 1.0[173] 和 ForgeryNet[176] 数据集等. 随着深度伪造及其检测技术的不断发展，近年来发布的深度伪造数据集中生成方法更加多样化，包含伪造样本容量更多，也逐渐出现了更多的检测任务，如 ForgeryNet[176] 的伪造区域定位(见表 3)和 FFIW-10K Dataset[175] 中可能有多张人脸被篡改的鉴别任务.

5.2 深度伪造语音数据集

ASVspoofing 是近几年影响力较大的语音类伪造检测比赛，自 2015 年举办第一届以来至今已经举办了 4 届. 其中前 3 届的 ASVspoofing 包含了语音合成和转换以及录音重放的鉴别任务，ASVspoofing 2021

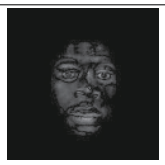
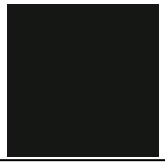
Table 2 Deepfake Video and Image Datasets

表 2 深度伪造视频和图片数据集

数据集	发布年份	伪造方法	数据集描述	数据集大小	真伪样本数量比
DFD[150]	2019	Deepfakes	篡改视频均使用 C_0，C_{23}，C_{40} 这 3 种压缩方式	363 个原始视频、3 068 个篡改视频、28 个演员和 16 个不同场景	1∶8.45
Deepfake-TIMIT[151]	2018	FaceSwap-GAN	从 VidTIMIT 数据库中选取相近人脸伪造构建	320 个视频、每个视频有高清（128×128）和低清（64×64）版本	1∶1
DFDC（deepfake detection challenge）Preview[148]	2019	未知	DFDC 预赛中使用的数据集	5 214 个视频	1∶3.57
DFDC[170]	2020	8 种伪造方法	DFDC 比赛中使用的数据集	119 154 个视频	1∶5.26
FaceForensics++（FF++）[12]	2019	Deepfakes，FaceSwap，Face2face，Neuraltexture，faceshifter	Google 推出的另一个数据集，前身为 FaceForensics，目前仍在持续更新	6 000 个视频	1∶5
Celeb-DF[149]	2020	Deepfakes	视频数量较少，已有后续版本 Celeb-DF v2[152] 和 DFGC（deepfake game competition）[171]	590 个真实视频、5 639 个伪造视频	1∶9.56
Wild Deepfake[172]	2020	网络途径获取	通过网络获取的伪造数据集，效果较好	707 个伪造视频、100 个演员	
DeeperForensics 1.0[173]	2020	deepfake-VAE	大型深度伪造数据集，包含多种灯光条件和面部角度，同时使用了改进的生成方法，较之前数据集更为真实	60 000 个视频、1 760 万帧	1∶5
Video Forensics HQ[174]	2020	Neural Textures	高清视频伪造数据集		
FFIW-10K[175]	2021	3 种合成方法	同一个视频片段中出现多个可能被篡改的人脸，平均每帧 3.15 个人脸	10 000 个真实视频和 10 000 个篡改视频	1∶1
ForgeryNet[176]	2021	15 种合成方法（7 种图像级方法、8 种视频级方法）	支持多种任务的超大数据集（630 万个分类标签、290 万个操纵区域标注和 221 247 个时空伪造段标签）	290 万张图像、221 247 个视频	视频 1∶1.22，图片 1∶1.01
FakeAVCeleb[177]	2021	5 种伪造方法	多模态数据集、伪造视频包含音频	25 500 个视频	1∶51.02

Table 3 Spatial Forgery Localization Detection Task in ForgeryNet[176]

表 3 ForgeryNet 中的定位伪造区域的检测任务[176]

伪造方法	图片样本	检测结果标注
面部替换		
面部重现		
面部特征编辑		
真实人脸		

在前 3 届的基础上增加了没有说话人验证的语音深度伪造鉴别任务. ASVspoofing 比赛发布的相关数据集也是当前最为常用的深度伪造语音数据集. 前 3 届 ASVspoofing 数据集详情如表 4 所示.

6 挑战与未来发展

6.1 深度伪造生成技术

深度伪造生成技术经过近几年的发展，各种任务的生成方法已经越来越成熟，但针对面部替换和面部重现的深度伪造方法在某些特定场景下仍然会有较为明显的伪造痕迹. 其原因主要在于现实环境下取得的图片或者视频样本可能有部分面部信息缺失，以及伪造过程中融合背景和人脸时效果较差. 针对这一缺陷，可以尝试在视频对数据预处理时用生成模型补全缺失的人脸信息，并在融合背景和人脸的步骤使用专门设计的网络进行优化. 在视频伪造方面，由于目前的伪造方法，如文献 [21−22] 等方法一般是单帧合成，在帧间时序信息中容易留下较多的伪造痕迹. 针对这一缺陷，可以尝试在视频生成后专门针对帧间差异进行平滑处理. 针对语音的深度伪造生成技术目前在伪造结果自然度的方面还有待提升.

未来的深度伪造研究会在解决信息缺失、增加伪造结果真实性的同时，致力于使用更少样本生成伪造结果，并提高伪造模型在实际应用过程中的运行性能. 同时一些其他形式的深度伪造潜力也在不断被研究者们挖掘，如 Facebook 公司提出的文字版本深度伪造[181].

6.2 深度伪造检测技术

目前针对单个数据集或者伪造方式的深度伪造检测方法已经较为成熟，但现有的深度鉴伪模型仍存在较多问题. 目前深度伪造检测相关方法的跨模型实验能够明显体现出泛化性能较差，而诸多针对检测模型的对抗攻击测试也显示现有检测模型的鲁棒性有待提升，且难以应对现实世界中不同压缩率、噪声等复杂条件. 同时由于在真实应用场景下难以得知伪造生成算法的类型，无法使用已有的预训练模型，也给深度伪造检测实际应用增添了不确定性.

深度伪造检测研究在公共舆论以及个人隐私安全保护等方面有重大的社会意义，针对现有方法的缺点提出了深度伪造检测技术的未来发展方向. 首先，针对检测模型泛化能力差的缺点，以后的伪造检测应该更加注重深度伪造模型的共性，如利用 GAN 模型指纹，研究泛化能力强的检测方法，或者通过使用目标人物参考图片或视频的方法，保证检测模型的高准确率. 其次，未来的深度伪造检测需要更加注重鲁棒性，可以在构建数据集时兼顾更多压缩率和噪声干扰，使模型在可以适应现实世界复杂环境的同时也能防止部分伪造视频通过添加扰动而绕过检测. 最后，未来的研究工作会聚焦到更多检测任务，而不是简单的二分类任务中. 现有的数据集已经定义了一些新的检测任务，如视频样本中会有多个可能被篡改的人脸[175]，以及要实现更细粒度的分类和

Table 4 Deepfake Audio Datasets

表 4 深度伪造语音数据集

数据集	发布年份	描述	数据集大小
ASVspoof 2015[178]	2015	语音合成与转换	16 651 段原始音频、246 500 段合成转换音频
ASVspoof 2017[179]	2017	录音重放	3 566 段非重放音频、14 466 段重放音频
ASVspoof 2019[180]	2019	录音重放、语音合成与转换	15 928 原始音频、117 996 合成转换音频

定位深度伪造的时间和空间位置[176]等.

7 总 结

随着深度学习在图片处理领域应用的不断成熟，各种针对人脸等部位的伪造技术层出不穷，使得相关技术在教育和娱乐等领域得到广泛应用的同时，也对现有的检测技术产生了巨大的挑战.虽然并非所有的深度伪造的出发点都是恶意篡改，但我们目前还无法预估相关技术被不法分子利用后产生的不良影响，因此也迫切需要相关法律的制定或者完善检测体系，以促使深度伪造相关技术能在更多场景合法应用.总结了近些年来深度伪造及其检测的主流技术，并对其进行了分类探讨.同时还总结了目前常用的深度伪造视频及音频数据集，并分析了深度伪造及检测的技术难点和未来发展方向.我们希望通过本文能让更多人了解深度伪造相关技术，防止其产生不良的社会影响，并促进其在更多领域的合法应用.

作者贡献声明：李泽宇负责文献资料的整理分析及论文撰写；张旭鸿负责方案设计和论文修订；蒲誉文和伍一鸣负责论文修订；纪守领负责对调研提出意见并指导论文修改.

参 考 文 献

[1] Mirsky Y, Lee W. The creation and detection of deepfakes: A survey[J]. ACM Computing Surveys, 2021, 54(1): 264−263

[2] Kingma D P, Welling M. Auto-encoding variational Bayes[J]. arXiv preprint, arXiv: 1312.6114, 2013

[3] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C] //Proc of the 27th Int Conf on Neural Information Processing Systems. La Jolla, CA : NIPS, 2014: 2672−2680

[4] Isola P, Zhu Junyan, Zhou Tinghui, et al. Image-to-image translation with conditional adversarial networks[C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 1125−1134

[5] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C] //Proc of the 18th Int Conf on Medical Image Computing and Computer-assisted Intervention. Berlin: Springer, 2015: 234−241

[6] Wang Tingchun, Liu Mingyu, Zhu Yanjun, et al. High-resolution image synthesis and semantic manipulation with conditional GANs[C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 8798−8807

[7] Wang Tingchun, Liu Mingyu, Zhu Yanjun, et al. Video-to-video synthesis[J]. arXiv preprint, arXiv: 1808.06601, 2018

[8] Zhu Junyan, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C] //Proc of the 30th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 2223−2232

[9] Huang Gao, Liu Zhuang, Van Der Maaten L, et al. Densely connected convolutional networks[C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 4700−4708

[10] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770−778

[11] Chollet F. Xception: Deep learning with depthwise separable convolutions [C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 1251−1258

[12] Rossler A, Cozzolino D, Verdoliva L, et al. FaceForensics++: Learning to detect manipulated facial images [C] //Proc of the 17th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 1−11

[13] Dale K, Sunkavalli K, Johnson M K, et al. Video face replacement [J]. ACM Transactions on Graphics, 2011, 30(6): 8: 1−8: 10

[14] torzdf. Deepfakes [CP/OL] 2017 [2021-10-15]. https://github.com/deepfakes/face swap

[15] Korshunova I, Shi Wenzhe, Dambre J, et al. Fast face-swap using convolutional neural networks[C] //Proc of the 16th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 3677−3685

[16] Ulyanov D, Lebedev V, Vedaldi A, et al. Texture networks: Feed-forward synthesis of textures and stylized images[C] //Proc of the 33rd Int Conf on Machine Learning. New York: PMLR, 2016: 1349−1357

[17] Shaoanlu. Fceswap-GAN [CP/OL]. 2017 [2021-10-15]. https://github.com/shaoa nlu/faceswap-GAN

[18] Natsume R, Yatagawa T, Morishima S. FsNet: An identity-aware generative model for image-based face swapping[C] //Proc of the 14th Asian Conf on Computer Vision. Berlin: Springer, 2018: 117−132

[19] Natsume R, Yatagawa T, Morishima S. RSGAN: Face swapping and editing using face and hair representation in latent spaces[J]. arXiv preprint, arXiv: 1804.03447, 2018.

[20] Nirkin Y, Keller Y, Hassner T. FSGAN: Subject agnostic face swapping and reenactment[C] //Proc of the 17th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 7184−7193

[21] Li Lingzhi, Bao Jianmin, Yang Hao, et al. Faceshifter: Towards high fidelity and occlusion aware face swapping[J]. arXiv preprint, arXiv: 1912.13457, 2019

[22] Chen Renwang, Chen Xuanhong, Ni Bingbing, et al. Simswap: An efficient framework for high fidelity face swapping[C] //Proc of the 28th ACM Int Conf on Multimedia. New York: ACM, 2020: 2003−2011

[23] Zhu Yuhao, Li Qi, Wang Jian, et al. One shot face swapping on megapixels [C] //Proc of the 18th IEEE/CVF Conf on Computer

Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 4834−4844

[24] Lin Yuan, Lin Qian, Tang Feng, et al. Face replacement with large-pose differences[C] //Proc of the 20th ACM Int Conf on Multimedia. New York: ACM, 2012: 1249−1250

[25] Min Feng, Sang Nong, Wang Zhefu. Automatic face replacement in video based on 2D morphable model[C] //Proc of the 20th Int Conf on Pattern Recognition. Piscataway, NJ: IEEE, 2010: 2250−2253

[26] Moniz J R A, Beckham C, Rajotte S, et al. Unsupervised depth estimation, 3D face rotation and replacement[J]. arXiv preprint, arXiv: 1803.09202, 2018

[27] Thies J, Zollhofer M, Niessner M, et al. Real-time expression transfer for facial reenactment[J]. ACM Transactions on Graphics, 2015, 34(6): 183: 1−183: 4

[28] Thies J, Zollhofer M, Stamminger M, et al. Face2Face: Real-time face capture and reenactment of rgb videos[C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2387−2395

[29] Thies J, Zollhofer M, Theobalt C, et al. Headon: Real-time reenactment of human portrait videos[J]. ACM Transactions on Graphics, 2018, 37(4): 164: 1−164: 13

[30] Kim H, Garrido P, Tewari A, et al. Deep video portraits[J]. ACM Transactions on Graphics, 2018, 37(4): 163: 1−163: 14

[31] Nagano K, Seo J, Xing Jun, et al. PaGAN: Real-time avatars using dynamic textures[J]. ACM Transactions on Graphics (TOG), 2018, 37(6): 258: 1−258: 12

[32] Geng Jiahao, Shao Tianjia, Zheng Youyi, et al. Warp-guided GANs for single-photo facial animation[J]. ACM Transactions on Graphics, 2018, 37(6): 231: 1−231: 12

[33] Wang Yaohui, Bilinski P, Bremond F, et al. Imaginator: Conditional spatio-temporal GAN for video generation[C] //Proc of the 20th IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2020: 1160−1169

[34] Siarohin A, Lathuiliere S, Tulyakov S, et al. Animating arbitrary objects via deep motion transfer[C] //Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 2377−2386

[35] Siarohin A, Lathuiliere S, Tulyakov S, et al. First order motion model for image animation[C] //Proc of the 32nd Int Conf on Neural Information Processing Systems. La Jolla, CA : NIPS, 2019: 7137−7147

[36] Qian Shengju, Lin K Y, Wu W, et al. Make a face: Towards arbitrary high fidelity face manipulation[C] //Proc of the 32nd IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 10033−10042

[37] Song Linsen, Wu W, Fu Chaoyou, et al. Pareidolia face reenactment[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 2236−2245

[38] Pumarola A, Agudo A, Martinez A M, et al. GANimation: Anatomically-aware facial animation from a single image[C] //Proc of the 15th European Conf on Computer Vision (ECCV). Berlin: Springer, 2018: 818−833

[39] Tripathy S, Kannala J, Rahtu E. FACEGAN: Facial attribute controllable reenactment gan[C] //Proc of the 21st IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2021: 1329−1338

[40] Gu Kuangxiao, Zhou Yuqian, Huang T. FLNet: Landmark driven fetching and learning network for faithful talking facial animation synthesis[C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 10861−10868

[41] Xu Runze, Zhou Zhiming, Zhang Weinan, et al. Face transfer with generative adversarial network[J]. arXiv preprint, arXiv: 1710.06090, 2017

[42] Bansal A, Ma Shugao, Ramanan D, et al. RecycleGan: Unsupervised video retargeting[C] //Proc of the 15th European Conf on Computer Vision (ECCV). Berlin: Springer, 2018: 119−135

[43] Wu W, Zhang Yunxuan, Li Cheng, et al. ReenactGAN: Learning to reenact faces via boundary transfer[C] //Proc of the 15th European Conf on Computer Vision (ECCV). Berlin: Springer, 2018: 603−619

[44] Zhang Jiangning, Zeng Xianfang, Wang Mengmeng, et al. FReeNet: Multi-identity face reenactment[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 5326−5335

[45] Zhang Jiangning, Zeng Xianfang, Pan Yusu, et al. FaceSwapNet: Landmark guided many-to-many face reenactment[J]. arXiv preprint, arXiv: 1905.11805, 2019

[46] Tripathy S, Kannala J, Rahtu E. ICface: Interpretable and controllable face reenactment using GANs[C] //Proc of the 20th IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2020: 3385−3394

[47] Wiles O, Koepke A, Zisserman A. X2Face: A network for controlling face generation using images, audio, and pose codes[C] //Proc of the 15th European Conf on Computer Vision (ECCV). Berlin: Springer, 2018: 670−686

[48] Shen Yujun, Luo Ping, Yan Junjie, et al. Faceid-GAN: Learning a symmetry three-player GAN for identity-preserving face synthesis[C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 821−830

[49] Shen Yujun, Zhou Bolei, Luo Ping, et al. FaceFeat-GAN: A two-stage approach for identity-preserving face synthesis[J]. arXiv preprint, arXiv: 1812.01288, 2018

[50] Wang Tingchun, Liu Mingyu, Tao A, et al. Few-shot video-to-video synthesis[J]. arXiv preprint, arXiv: 1910.12713, 2019

[51] Zakharov E, Shysheya A, Burkov E, et al. Few-shot adver-sarial learning of realistic neural talking head models[C] //Proc of the 32nd IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 9459−9468

[52] Burkov E, Pasechnik I, Grigorev A, et al. Neural head reenactment with latent pose descriptors[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 13786−13795

[53] Ha S, Kersner M, Kim B, et al. MarioNETte: Few-shot face reenactment preserving identity of unseen targets[C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 10893−10900

[54] Hao Hanxiang, Baireddy S, Reibman A R, et al. Far-GAN for one-shot face reenactment[J]. arXiv preprint, arXiv: 2005.06402, 2020

[55] Fried O, Tewari A, Zollhofer M, et al. Text-based editing of talking-head video[J]. ACM Transactions on Graphics, 2019, 38(4): 68: 1−68: 14

[56] Kumar R, Sotelo J, Kumar K, et al. ObamaNet: Photo-realisticlip-sync from text[J]. arXiv preprint, arXiv: 1801.01442, 2017

[57] Sotelo J, Mehri S, Kumar K, et al. Char2wav: End-to-end speech synthesis[C] //Proc of the ICLR 2017 Workshop. 2017: 24−26

[58] Jamaludin A, Chung J S, Zisserman A. You said that?: Synthesising talking faces from audio[J]. International Journal of Computer Vision, 2019, 127(11): 1767−1779

[59] Vougioukas K, Petridis S, Pantic M. Realistic speech-driven facial animation with GANs[J]. International Journal of Computer Vision, 2020, 128(5): 1398−1413

[60] Suwajanakorn S, Seitz S M, Kemelmacher-shlizerman I. Synthesizing Obama: Learning lip sync from audio[J]. ACM Transactions on Graphics, 2017, 36(4): 95: 1−95: 13

[61] Chen Lele, Maddox R K, Duan Zhiyao, et al. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss[C] //Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 7832−7841

[62] Zhou Hang, Liu Yu, Liu Ziwei, et al. Talking face generation by adversarially disentangled audio-visual representation[C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 9299−9306

[63] Thies J, Elgharib M, Tewari A, et al. Neural voice puppetry: Audio-driven facial reenactment[C] //Proc of the 16th European Conf on Computer Vision (ECCV). Berlin: Springer, 2020: 716−731

[64] Hannun A, Case C, Casper J, et al. DeepSpeech: Scaling up end-to-end speech recognition[J]. arXiv preprint, arXiv: 1412.5567, 2014

[65] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks[C] //Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 4401−4410

[66] Karras T, Laine S, Airtala M, et al. Analyzing and improving the image quality of StyleGAN[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 8110−8119

[67] Karras T, Aittala M, Laine S, et al. Alias-free generative adversarial networks[J]. arXiv preprint, arXiv: 2106.12423, 2021

[68] Choi Y, Choi M, Kim M, et al. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 8789−8797

[69] Choi Y, Uh Y, Yoo J, et al. StarGAN v2: Diverse image synthesis for multiple domains[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 8188−8197

[70] Sanchez E, Valstar M. Triple consistency loss for pairing distributions in GAN-based face synthesis[J]. arXiv preprint, arXiv: 1811.03492, 2018

[71] Kim D, Khan M A, Choo J. Not just compete, but collaborate: Local image-to-image translation via cooperative mask prediction[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 6509−6518

[72] Li Xinyang, Zhang Shengchuan, Hu Jie, et al. Image-to-image translation via hierarchical style disentanglement[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 8639−8648

[73] Aberman K, Shi Mingyi, Liao Jing, et al. Deep video-based performance cloning[J]. Computer Graphics Forum, 2019, 38(2): 219−233

[74] Chan C, Ginosar S, Zhou Tinghui, et al. Everybody Dance Now [C] //Proc of the 32nd IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 5933−5942

[75] Liu Lingjie, Xu Weipeng, Zollhofer M, et al. Neural rendering and reenactment of human actor videos[J]. ACM Transactions on Graphics, 2019, 38(5): 139: 1−139: 14

[76] Tokuda K, Nankaku Y, Toda T, et al. Speech synthesis based on hidden Markov models[J]. Proceedings of the IEEE, 2013, 101(5): 1234−1252

[77] Oord A, Dieleman S, Zen H, et al. WaveNet: A generative model for raw audio[J]. arXiv preprint, arXiv: 1609.03499, 2016

[78] Wang Yuxuan, Skerry-ryan R, Stanton D, et al. Tacotron: A fully end-to-end text-to-speech synthesis model[J]. arXiv preprint, arXiv: 1703.10135, 2017

[79] Shen J, Pang Ruoming, Weiss R J, et al. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions[C] //Proc of the 43rd IEEE Int Conf on Acoustics, Speech and Signal Processing(ICASSP). Piscataway, NJ: IEEE, 2018: 4779−4783

[80] Fu Ruibo, Tao Jianhua, Wen Zhengqi, et al. Focusing on attention: Prosody transfer and adaptative optimization strategy for multi-speaker end-to-end speech synthesis[C] //Proc of the 45th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2020: 6709−6713

[81] Kumar K, Kumar R, de Boissiere T, et al. MelGAN: Generative adversarial networks for conditional waveform synthesis[J]. arXiv preprint, arXiv: 1910.06711, 2019.

[82] Yang Geng, Yang Shan, Liu Kai, et al. Multi-band melgan: Faster waveform generation for high-quality text-to-speech[C] //Proc of the 8th IEEE Spoken Language Technology Workshop (SLT). Piscataway, NJ: IEEE, 2021: 492−498

[83] Kaneko T, Kameoka H. CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks[C] //Proc of the 27th European Signal Processing Conf (EUSIPCO). Piscataway, NJ: IEEE, 2018: 2100−2104

[84] Kaneko T, Kameoka H, Tanaka K, et al. CycleGAN-VC2: Improved cyclegan-based non-parallel voice conversion[C] //Proc of the 44th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2019: 6820−6824

[85] Kaneko T, Kameoka H, Tanaka K, et al. CycleGAN-VC3: Examining and improving CycleGAN-VCs for mel-spectrogram conversion[J]. arXiv preprint, arXiv: 2010.11672, 2020

[86] Kameoka H, Kaneko T, Tanaka K, et al. StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks[C] //Proc of the 7th IEEE Spoken Language Technology Workshop (SLT). Piscataway, NJ: IEEE, 2018: 266−273

[87] Kaneko T, Kameoka H, Tanaka K, et al. StarGAN-VC2: Rethinking conditional methods for StarGAN-based voice conversion[J]. arXiv preprint, arXiv: 1907.12279, 2019

[88] Liu Ruolan, Chen Xiao, Wen Xue. Voice conversion with transformer network[C] //Proc of the 45th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2020: 7759−7759

[89] Luong H T, Yamagishi J. Bootstrapping non-parallel voice conversion from speaker-adaptive text-to-speech[C] //Proc of the 16th IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Piscataway, NJ: IEEE, 2019: 200−207

[90] Zhang Mingyang, Zhou Yi, Zhao Li, et al. Transfer learning from speech synthesis to voice conversion with non-parallel training data[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29(1): 1290−1302

[91] Huang Wenqin, Hayashi T, Wu Yiqiao, et al. Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining[J]. arXiv preprint, arXiv: 1912.06813, 2019

[92] Matern F, Riess C, Stamminger M. Exploiting visual artifacts to expose deepfakes and face manipulations[C] //Proc of the 20th IEEE Winter Applications of Computer Vision Workshops (WACVW). Piscataway, NJ: IEEE, 2019: 83−92

[93] Zhou Peng, Han Xintong, Morariu V I, et al. Two-stream neural networks for tampered face detection[C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway, NJ: IEEE, 2017: 1831−1839

[94] Nataraj L, Mohammed T M, Manjunath B, et al. Detecting GAN generated fake images using co-occurrence matrices[J]. Electronic Imaging, 2019 : 1−7

[95] Li Jiaming, Xie Hongtao, Li Jiahong, et al. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 6458−6467

[96] Luo Yuchen, Zhang Yong, Yan Junchi, et al. Generalizing face forgery detection with high-frequency features[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 16317−16326

[97] Shang Zhihua, Xie Hongtao, Zha Zhengjun, et al. PrrNet: Pixel-region relation network for face forger1y detection[J/OL]. Pattern Recognition, 2021, 116 [2021-10-15]. https://doi.org/10.1016/j.patcog.2021.107950

[98] Li Yuezun, Lyu Siwei. Exposing deepfake videos by detecting face warping artifacts[J]. arXiv preprint, arXiv: 1811.00656, 2018

[99] Li Lingzhi, Bao Jianmin, Zhang Ting, et al. Face x-ray for more general face forgery detection[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 5001−5010

[100] Li Xurong, Yu Kun, Ji Shouling, et al. Fighting against deepfake: Patch&pair convolutional neural networks (PPCNN)[C] //Proc of the 29th the Web Conf . New York: ACM, 2020: 88−89

[101] Nguyen H, Fang Fuming, Yamagishi J, et al. Multi-task learning for detecting and segmenting manipulated facial images and videos[J]. arXiv preprint, arXiv: 1906.06876, 2019

[102] Nirkin Y, Wolf L, Keller Y, et al. Deepfake detection based on the discrepancy between the face and its context[J]. arXiv preprint, arXiv: 2008.12262, 2020

[103] Amerini I, Caldelli R. Exploiting prediction error in consistencies through LSTM-based classifiers to detect deepfake videos[C] //Proc of the 8th ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2020: 97−102

[104] Amerini I, Galteri L, Caldelli R, et al. Deepfake video detection through optical flow based CNN[C] //Proc of the 32nd IEEE/CVF Int Conf on Computer Vision Workshops. Piscataway, NJ: IEEE, 2019: 1205−1207

[105] Guera D, Delp E J. Deepfake video detection using recurrent neural networks[C/OL] //Proc of the 15th IEEE Int Conf on Advanced Video and Signal Based Surveillance (AVSS). Piscataway, NJ: IEEE, 2018 [2021-10-15]. https://doi.org/10.1109/AVSS.2018.8639163

[106] Sun Zekun, Han Yujie, Hua Zeyu, et al. Improving the efficiency and robustness of deepfakes detection through precise geometric features[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 3609−3618

[107] Sabir E, Cheng Jiaxin, Jaiswal A, et al. Recurrent convolutional strategies for face manipulation detection in videos[C] //Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2019: 80−87

[108] Agarwal S, Farid H, Gu Yuming, et al. Protecting world leaders against deep fakes [C] //Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2019: 38−45

[109] Agarwal S, Farid H, Fried O, et al. Detecting deep-fake videos from phoneme-viseme mismatches[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2020: 660−661

[110] Yang Xin, Li Yuezun, Lyu Siwei. Exposing deep fakes using inconsistent head poses[C] //Proc of the 44th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2019: 8261−8265

[111] Ciftci U A, Demir I, Yin Lijun. FakeCatcher: Detection of synthetic portrait videos using biological signals[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020 [2021-10-15]. https://doi.org/10.1109/T PAMI.2020.3009287

[112] Fernandes S, Raj S, Ewetz R, et al. Detecting deepfake videos using attribution-based confidence metric[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2020: 308−309

[113] Jha S, Raj S, Fernandes S, et al. Attribution-based confidence metric for deep neural networks[C] //Proc of the 32nd Int Conf on Neural Information Processing Systems. La Jolla, CA : NIPS, 2019: 11826−11837

[114] McCloskey S, Albright M. Detecting GAN-generated imagery using color cues[J]. arXiv preprint, arXiv: 1812.08247, 2018

[115] Guarnera L, Giudice O, Battiato S. Deepfake detection by analyzing convolutional traces[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2020: 666−667.

[116] Qian Yuyang, Yin Guojun, Sheng Lu, et al. Thinking in frequency: Face forgery detection by mining frequency-aware clues[C] //Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 86−103

[117] Masi I, Killekar A, Mascarenhas R M, et al. Two-branch recurrent network for isolating deepfakes in videos[C] //Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 667−684

[118] Liu Honggu, Li Xiaodan, Zhou Wenbo, et al. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 772−781

[119] Agarwal S, Farid H, EL-Gaaly T, et al. Detecting deepfake videos from appearance and behavior[C/OL] //Proc of the 12th IEEE Int Workshop on Information Forensics and Security (WIFS). Piscataway, NJ: IEEE, 2020 [2021-10-15]. https://doi.org/10.1109/WIFS49906.2020.9360904

[120] Wiles O, Koepke A, Zisserman A. Self-supervised learning of a facial attribute embedding from video[J]. arXiv preprint, arXiv: 1808.06882, 2018

[121] Cozzolino D, Rossler A, Thies J, et al. Id-reveal: Identity-aware deepfake video detection[J]. arXiv preprint, arXiv: 2012.02512, 2020

[122] Dong Xiaoyi, Bao Jianmin, Chen Dongdong, et al. Identity-driven deepfake detection[J]. arXiv preprint, arXiv: 2012.03930, 2020

[123] Jiang Jun, Wang Bo, Li Bing, et al. Practical face swapping detection based on identity spatial constraints[C] //Proc of the 7th IEEE Int Joint Conf on Biometrics (IJCB). Piscataway, NJ: IEEE, 2021: 1−8

[124] Lewis J K, Toubal I E, Chen Helen, et al. Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multi-modal deep learning[C/OL] //Proc of the 49th IEEE Applied Imagery Pattern Recognition Workshop (AIPR). Piscataway, NJ: IEEE, 2020 [2021-10-15]. https://doi.org/10.1109/AIPR50011.2020.9425167

[125] Lomnitz M, Hampel-arias Z, Sandesara V, et al. Multimodal approach for deepfake detection[C/OL] //Proc of the 49th IEEE Applied Imagery Pattern Recognition Workshop (AIPR). Piscataway, NJ: IEEE, 2020 [2021-10-15]. https://doi.org/10.1109/AIPR50011.2020.9425192

[126] Ravanelli M, Bengio Y. Speaker recognition from raw waveform with SincNet[C] //Proc of the 7th IEEE Spoken Language Technology Workshop(SLT). Piscataway, NJ: IEEE, 2018: 1021−1028

[127] Mittal T, Bhattacharya U, Chandra R, et al. Emotions don't lie: An audio-visual deepfake detection method using affective cues [C] //Proc of the 28th ACM Int Conf on Multimedia. New York: ACM, 2020: 2823−2832

[128] Hosler B, Salvi D, Murray A, et al. Do deepfakes feel emotions? A semantic approach to detecting deepfakes via emotional inconsistencies[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 1013−1022

[129] Afchar D, Nozick V, Yamagishi J, et al. MesoNet: A compact facial video forgery detection network[C/OL] //Proc of the 10th IEEE Int Workshop on Information Forensics and Security (WIFS). Piscataway, NJ: IEEE, 2018 [2021-10-15]. https://doi.org/10.1109/WIFS.2018.8630761

[130] Jain A, Singh R, Vatsa M. On detecting GANs and retouching based synthetic alterations[C/OL] //Proc of the 9th Int Conf on Biometrics Theory, Applications and Systems (BTAS). Piscataway, NJ: IEEE, 2018 [2021-10-15]. https://doi.org/10.1109/BTAS.2018.8698545

[131] Wang Run, Xu Juefei, Ma Lei, et al. FakeSpotter: A simple yet robust baseline for spotting ai-synthesized fake faces[J]. arXiv preprint, arXiv: 1909.06122, 2019

[132] Dang Hao, Liu Feng, Stehouwer J, et al. On the detection of digital face manipulation[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 5781−5790

[133] Hsu C C, Zhuang Yixiu, Lee C Y. Deep fake image detection based on pairwise learning[J/OL]. Applied Sciences, 2020 [2021-10-15]. https://doi.org/10.3390/app10010370

[134] Khalid H, Woo S S. Oc-fakedect: Classifying deepfakes using one-class variational autoencoder[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2020: 656−657

[135] Rana M S, Sung A H. DeepfakeStack: A deep ensemble-based learning technique for deepfake detection[C] //Proc of the 7th IEEE Int Conf on Cyber Security and Cloud Computing(CSCloud)/IEEE Int Conf on Edge Computing and Scalable Cloud (EdgeCom). Piscataway, NJ: IEEE, 2020: 70−75

[136] Bonettini N, Cannas E D, Mandelli S, et al. Video face manipulation detection through ensemble of CNNs[C] //Proc of the 31st Int Conf on Pattern Recognition (ICPR). Piscataway, NJ: IEEE, 2021: 5012−5019

[137] Kim M, Tariq S, Woo S S. FReTal: Generalizing deepfake detection using knowledge distillation and representation learning[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 1001−1012

[138] Aneja S, Niessner M. Generalized zero and few-shot transfer for facial forgery detection[J]. arXiv preprint, arXiv: 2006.11863, 2020

[139] Wang Chengrui, Deng Weihong. Representative forgery mining for fake face detection[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 14923−14932

[140] Zhao Hanqing, Zhou Wenbo, Chen Dongdong, et al. Multi-attentional deepfake detection[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 2185−2194

[141] Kumar P, Vatsa M, Singh R. Detecting face2face facial reenactment in videos[C] //Proc of the 20th IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2020: 2589−2597

[142] Jeon H, Bang Y, Woo S S. FdftNet: Facing off fake images using

fake detection fine-tuning network[C] //Proc of the 35th IFIP Int Conf on ICT Systems Security and Privacy Protection. Berlin: Springer, 2020: 416−430

[143] Wang Shengyu, Wang O, Zhang R, et al. CNN-generated images are surprisingly easy to spot for now[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 8695−8704

[144] Liu Zhengzhe, Qi Xiaojuan, Torr P. Global texture enhancement for fake face detection in the wild[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 8060−8069

[145] Wodajo D, Atnafu S. Deepfake video detection using convolutional vision transformer[J]. arXiv preprint, arXiv: 2102.11126, 2021

[146] Wang Junke, Wu Zuxuan, Chen Jingjing, et al. M2tr: Multi-modal multi-scale transformers for deepfake detection[J]. arXiv preprint, arXiv: 2104.09770, 2021

[147] Heo Y, Choi Y, Lee Y, et al. Deepfake detection scheme based on vision transformer and distillation[J]. arXiv preprint, arXiv: 2104.01353, 2021

[148] Dolhansky B, Howes R, Pflaum B, et al. The deepfake detection challenge (DFDC) preview dataset[J]. arXiv preprint, arXiv: 1910.08854, 2019

[149] Li Yuezun, Yang Xin, Sun Pu, et al. Celeb-DF: A large-scale challenging dataset for deepfake forensics[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 3207−3216

[150] Ondyari. Deepfake detection (DFD) dataset [DB/OL]. 2018 [2021-10-15]. https://github.com/ondyari/FaceForensics

[151] Koeshunov P, Marcel S. Deepfakes: A new threat to face recognition? assessment and detection[J]. arXiv preprint, arXiv: 1812.08685, 2018

[152] Li Yuezun, Yang Xin, Sun Pu, et al. Celeb-DF (v2): A new dataset for deepfake forensics[J]. arXiv preprint, arXiv: 1909.12962, 2019

[153] Ruiz N, Bargal S A, Sclaroff S. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems[C] //Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 236−251

[154] Huang Qidong, Zhang Jie, Zhou Wenbo, et al. Initiative defense against facial manipulation[C] //Proc of the 35th AAAI Conf on Artificial Intelligence. New York: ACM, 2021: 1619−1627

[155] Dong Junhao, Xie Xiaohua. Visually maintained image disturbance against deepfake face swapping [C/OL] //Proc of the 22nd IEEE Int Conf on Multimedia and Expo (ICME). Piscataway, NJ: IEEE, 2021 [2021-10-15]. https://doi.org/10.1109/ICME51207.2021.9428173

[156] Neves J C, Tolosana R, Vera-rodriguez R, et al. Real or fake? Spoofing state-of-the-art face synthesis detection systems[J]. arXiv preprint, arXiv: 1911.05351, 2019

[157] Carlini N, Farid H. Evading deepfake-image detectors with white- and black-box attacks[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2020: 658−659

[158] Hussain S, Neekhara P, Jere M, et al. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples [C] //Proc of the 21st IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2021: 3348−3357

[159] Patel T B, Patil H A. Cochlear filter and instantaneous frequency based features for spoofed speech detection[J]. IEEE Journal of Selected Topics in Signal Processing, 2016, 11(4): 618−631

[160] Tom F, Jain M, Dey P. End-to-end audio replay attack detection using deep convolutional networks with attention.[C] //Proc of the 20th Interspeech. 2018 [2021-10-15]. https://www.isca-speech.org/archive_v0/Interspeech_2018/abstracts/2279.html

[161] Das R K, Yang Jichen, Li Haizhou. Assessing the scope of generalized counter-measures for anti-spoofing[C] //Proc of the 45th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2020: 6589−6593

[162] Lavrentyeva G, Novoselov S, Malykh E, et al. Audio replay attack detection with deep learning frameworks[C] //Proc of the 19th Interspeech.2017[2021-10-15].https://www.isca-speech.org/archive_v0/Interspeech_2017/abstracts/0360.html

[163] Wu Xiang, He Ran, Sun Zhenan, et al. A light CNN for deep face representation with noisy labels[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(11): 2884−2896

[164] Lavrentyeva G, Novoselov S, Tseren A, et al. Stc anti-spoofing systems for the ASVspoof 2019 challenge[J]. arXiv preprint, arXiv: 1904.05576, 2019

[165] Cai Weicheng, Wu Haiwei, Cai Danwei, et al. The DKU replay detection system for the ASVspoof 2019 challenge: On data augmentation, feature representation, classification, and fusion[J]. arXiv preprint, arXiv: 1907.02663, 2019

[166] Lai C I, Chen Nanxin, Villalba J, et al. Assert: Anti-spoofing with squeeze-excitation and residual networks[J]. arXiv preprint, arXiv: 1904.01120, 2019

[167] Parasu P, Epps J, Sriskandaraja K, et al. Investigating light-resnet architecture for spoofing detection under mismatched conditions [C] // Proc of the 22nd Interspeech. 2020 [2021-10-15]. https://www.isca-speech.org/archive_v0/Interspeech_2020/abstracts/2039.html

[168] Ma Haoxin, Yi Jiangyan, Tao Jianhua, et al. Continual learning for fake audio detection[J]. arXiv preprint, arXiv: 2104.07286, 2021

[169] Li Zzhizhong, Hoiem D. Learning without forgetting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(12): 2935−2947

[170] Dolhansky B, Bitton J, Pflaum B, et al. The deepfake detection challenge (DFDC) dataset[J]. arXiv preprint, arXiv: 2006.07397, 2020

[171] Peng Bo, Fan Hongxing, Wang Wei, et al. DFGC 2021: A deepfake game competition[J]. arXiv preprint, arXiv: 2106.01217, 2021

[172] Zi Bojia, Chang Minghao, Chen Jingjing, et al. Wild Deepfake: A challenging real-world dataset for deepfake detection[C] //Proc of the 28th ACM Int Conf on Multimedia. New York: ACM, 2020: 2382−2390

[173] Jiang Liming, Li Ren, Wu W, et al. DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 2889−2898

[174] Fox G, Liu Wentao, Kim H, et al. Video ForensicsHQ: Detecting high-quality manipulated face videos[C/OL] //Proc of the 22nd IEEE Int Conf on Multimedia and Expo (ICME). Piscataway, NJ: IEEE, 2021 [2021-10-15]. https://doi.org/10.1109/ICME51207.2021.9428101

[175] Zhou Tianfei, Wang Wenguan, Liang Zhiyuan, et al. Face forensics in the wild[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 5778−5788

[176] He Yinan, Gan Bei, Chen Siyu, et al. ForgeryNet: A versatile benchmark for comprehensive forgery analysis[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 4360−4369

[177] Khalid H, Tariq S, Woo S S. FakeAVCeleb: A novel audio-video multimodal deepfake dataset[J]. arXiv preprint, arXiv: 2108.05080, 2021

[178] University of Edinburgh, the Centre for Speech Technology Research (CSTR). ASVspoof 2015 database[DB/OL]. 2015 [2021-10-15]. https://datasha re.ed.ac.uk/handle/10283/853

[179] University of Edinburgh, the Centre for Speech Technology Research (CSTR). ASVspoof 2017 database [DB/OL]. 2017 [2021-10-15]. https://datashar e.ed.ac.uk/handle/10283/3055.

[180] University of Edinburgh, the Centre for Speech Technology Research (CSTR). ASVspoof 2019 database [DB/OL]. 2019 [2021-10-15]. https://datashar e.ed.ac.uk/handle/10283/3336.

[181] Krishnan P, Kovvuri R, Pang Guan, et al. Textstyle brush: Transfer of text aesthetics from a single example[J]. arXiv preprint, arXiv: 2106.08385, 2021

Li Zeyu, born in 1999. Master candidate. His main research interests include AI security and computer vision.

李泽宇, 1999 年生. 硕士研究生. 主要研究方向为人工智能安全和计算机视觉.

Zhang Xuhong, born in 1988. PhD, associate professor. Member of CCF. His main research interests include AI security, data driven software and system security, big data systems and analytics.

张旭鸿, 1988 年生. 博士, 副研究员. CCF 会员. 主要研究方向为人工智能安全、数据驱动软件与系统安全、大数据系统与分析.

Pu Yuwen, born in 1993. PhD. His main research interests include privacy computing and AI security.

蒲誉文, 1993 年生. 博士. 主要研究方向为隐私计算和人工智能安全.

Wu Yiming, born in 1996. PhD candidate. Her main research interests include data driven security, black industry mining, and cybercrime research.

伍一鸣, 1996 年生. 博士研究生. 主要研究方向为数据驱动安全、黑灰产业挖掘、网络犯罪研究.

Ji Shouling, born in 1986. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include data-driven security and privacy, AI security, big data mining and analytics.

纪守领, 1986 年生. 博士, 教授, 博士生导师. CCF 高级会员. 主要研究方向为数据驱动安全和隐私、人工智能安全、大数据挖掘与分析.