音视频深度伪造与鉴伪综述

魏方达[1]，刘森[1]，孙毅[2]，王晶[1]，赵胜辉[1]
（1.北京理工大学信息与电子学院，北京 100081；2.北京理工大学网络空间安全学院，北京 100081）

摘　要：近年来，深度学习在计算机视觉以及语音信号处理等领域取得了重大成功。然而，深度学习的飞速发展也带来了负面影响，各类伪造视频、语音在网络上泛滥成灾，一些不法分子利用深度学习技术替换原始视频的人脸、编辑面部属性、合成说话人语音、克隆语音，通过制作色情视频、虚假新闻、政治谣言等造成社会动荡、混乱，威胁个人利益、国家安全。为了消除这些负面影响，众多学者从不同的角度提出解决方案。早期伪造主要集中于单模态伪造，因此目前大多数解决方案侧重于单模态的伪造识别问题，未能充分考虑音频和视频之间的内在联系，然而现有的单模态鉴伪方式在处理音频与视频均被伪造的情况时常常表现出次优的识别性能。近期，随着研究的深入，部分学者开始探索使用多模态模型鉴伪，取得了显著的成果。本综述首先回顾了视频、语音伪造及鉴伪技术，收集整理了视频、语音、音视频伪造数据集，然后总结归纳了多模态鉴伪方法，最后对如今检测技术存在的问题和研究方向进行了分析并给出建议。

关键词：深度学习；伪造；鉴伪；多模态；音视频
中图分类号：TP391　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0070700

Review of Audio and Video Deepfakes and Counterfeit Detection

WEI Fangda[1], LIU Miao[1], SUN Yi[2], WANG Jing[1], ZHAO Shenghui[1]
(1.School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China;
2.School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

【Abstract】Deep learning has achieved significant success in computer vision and speech signal processing. However, its rapid development of deep learning has had negative effects. Many types of fake videos and voices are flooding the Internet. Some criminals use deep learning technology to replace the face in original video, edit facial attributes, and synthesize or clone the speaker's voice. Criminals can cause social unrest and chaos by producing pornographic videos, fake news, and political rumors, which threaten personal interests and national security. Many scholars have proposed solutions to eliminate these negative effects. Early forgeries focused primarily on single-modal forgeries. Therefore, most current solutions focus on single-modal counterfeit detection and fail to fully consider the intrinsic relationship between audio and video. Existing single-modal detection methods often exhibit suboptimal performance when both audio and video are forged. Recently, with the increase in research, some scholars have begun to explore the use of multimodal models for counterfeit detection and have achieved remarkable results. This survey reviews video and voice forgery and detection technologies, collects and sorts voice, audio, and video forgery datasets, and summarizes multimodal counterfeit detection methods. Finally, the existing problems and future research directions of the current detection technology are analyzed.

【Key words】deep learning; forgery; counterfeit detection; multimodal; audio and video

0　引言

近年来，深度伪造在网络上泛滥成灾，不法分子利用深度伪造技术合成视频和语音，对个人、社会、国家造成了严重威胁。伪造方式一般包括视频伪造及语音伪造。其中，视频伪造通过对面部操作生成合成视频，早期视频伪造多基于图形学建模，如今生成对抗网络(GAN)、扩散模型等深度学习技术成为主流，合成视频质量更高。语音伪造通过文本转语音或转换原始语音特征进行伪造。早期语音合成模型多为管道式，基于拼接合成法或参数法，如今端到端合成逐渐成为主流。早期语音转换(VC)多用平行数据，如今则多用非平行数据，可实现多对多转换。

为了应对深度伪造问题，众多学者提出了解决方案，这些方案涵盖视频、语音以及音视频多模态。视频鉴伪针对视频模态。早期方法多基于取证，使用传统信号处理技术提取频域和统计特征。现在人

基金项目：国家自然科学基金面上项目(62071039)。
作者简介：魏方达，男，硕士研究生，主研方向为音视频鉴伪；刘森、孙毅，博士研究生；王晶(通信作者)、赵胜辉，副教授。
收稿日期：2024-12-11　　**修回日期：**2025-02-17　**E-mail：**wangjing@bit.edu.cn

们逐渐采用不一致性分析以及数据驱动方法进行鉴伪。语音鉴伪针对语音模态。早期多为管道式模型,采用信号处理和支持向量机等方法,如今神经网络逐渐替代这些模块,提升了鉴伪效果。此外,为解决管道式结构易积累误差的问题,人们将前端和后端融合形成端到端结构,以提高鉴伪性能。

随着伪造技术的进步,单模态鉴伪难度升高,一些学者转向使用多模态技术采用不同的融合方式和损失函数处理音视频模态来鉴伪。此外,人们对泛化性能要求的提升,使得预训练式的多模态鉴伪方法被提出,其强调对真实视频的准确聚类以增强模型泛化性。

目前,针对本领域的综述多数侧重视频、语音单模态,只对早期的伪造以及鉴伪技术做了总结归纳[1-3],涉及音视频多模态鉴伪的综述较少并且不够全面[4]。首先,早期的伪造以及鉴伪总结侧重视频、语音单模态。但随着技术的进步,通过扩散模型等新技术,可以合成更逼真、分辨率更高的视频、语音。其次,一些新的合成方法克服了早期伪造方法存在的缺陷。例如,眨眼、头部转动、眼球转动等[1],而这些缺陷又是大多数早期的鉴伪工作所依赖的。因此,亟需对现有工作进行进一步的整理与归纳,并从音视频多模态的角度寻找解决方案。

本文第 1 章介绍视频伪造,第 2 章介绍语音伪造,第 3 章介绍伪造数据集,包括视频、语音及音视频多模态伪造数据集,第 4 章介绍视频鉴伪,第 5 章介绍语音鉴伪,第 6 章介绍音视频多模态鉴伪,第 7 章对现有鉴伪技术存在的问题及未来的研究方向做总结归纳。

1 视频伪造

近年来,深度学习在计算机视觉领域取得了重大成功。然而,这也导致了合成视频的泛滥。视频伪造包含基于面部交换(FS)的伪造、基于面部重建(FR)的伪造、基于面部属性编辑(FE)的伪造、基于面部生成(FG)的伪造 4 种类型。

1.1 基于面部交换的伪造

基于面部交换的伪造使用源视频的面部替换目标视频的面部,达到换脸目的。

1.1.1 基于图形学的面部交换

早期的面部交换方法基于图形学,通过 3D 模型对目标建模并渲染,达到换脸目的。例如,FaceSwap[5]通过 3D 模型对获取的人脸关键点进行建模和渲染,最后通过仿射变换和色彩校正实现换脸。

基于图形学的面部交换方法过于复杂,需要构建人脸的 3D 模型,时间开销大,门槛和成本都较高。此外,其对全局身份特征建模,容易忽略细节信息。

1.1.2 基于深度学习的面部交换

Deepfakes[6]为面部交换带来了新思路,如图 1 所示(彩色效果见《计算机工程》官网 HTML 版,下同),该模型使用两个共享权重的编码器训练提取特征,使两个解码器重建人脸,达到换脸效果。FaceShifter[7]使用 AEI-Net 产生初步换脸结果,HEAR-Net 对结果细化。XU 等[8]提出 RAFSwap 网络,可以生成更自然的人脸。

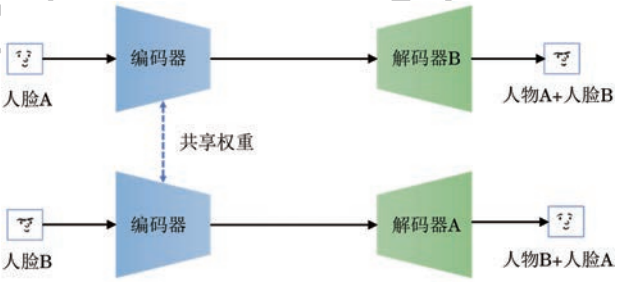

图 1 Deepfakes 面部交换架构[6]

Fig. 1 Deepfakes face swap architecture[6]

LIU 等[9]通过两个给定面孔的局部形状和纹理替换实现换脸。ZHAO 等[10]提出 DiffSwap 模型,依靠扩散模型生成目标视频。BALIAH 等[11]提出 REFace 模型,该模型将面部交换问题构建为自监督的面部修复任务,通过引入 CLIP 特征解缠,从目标图像中提取姿势、表情和照明信息,此外,使用多步去噪扩散模型加快计算速度。

基于面部交换的生成方法经历了从计算机图形学到生成对抗模型、扩散模型等深度学习算法的演变过程。随着技术的进步,面部交换算法开始关注照明、表情等局部细节信息的建模,并且针对身份保存和面部遮挡问题做出创新性的研究。

1.2 基于面部重建的伪造

基于面部重建的伪造方法使用源视频面部属性重建目标视频面部属性,包括面部表情重建、嘴部重建、眼部重建、头部重建等。

1.2.1 基于图形学的面部重建

早期面部重建方法多基于图形学,通过 3D 模型重建和渲染来实现面部重建。THIES 等[12]提出 Headon 模型,该模型修改视角和姿态独立的纹理进行视频级的渲染,实现了包括表情、眼睛、头部移动的完整重建。然而,基于图形学的面部重建方法过于复杂,依赖高精度的人脸 3D 模型,难以实现高质量的面部重建。

1.2.2 基于深度学习的面部重建

深度学习面部重建方法大多数使用其他模态驱动生成的结果，例如语音等。PRAJWAL 等[13]提出 Wav2Lip 模型，如图 2 所示，该模型将预训练的唇音同步模型作为判别器，用来监督模型生成自然、准确的唇部运动。LIANG 等[14]提出 GC-AVT 模型，从语音内容、头部姿态和情绪表征控制合成说话人面部，以互补的形式学习说话人面部情感的表达。

面部重建从图形学方法发展到深度学习方法，经历了目标视频驱动到音频、情绪等多个模态驱动的演变，生成的视频中人物越来越逼真，并且感情丰富，对鉴伪算法提出了更高的要求。

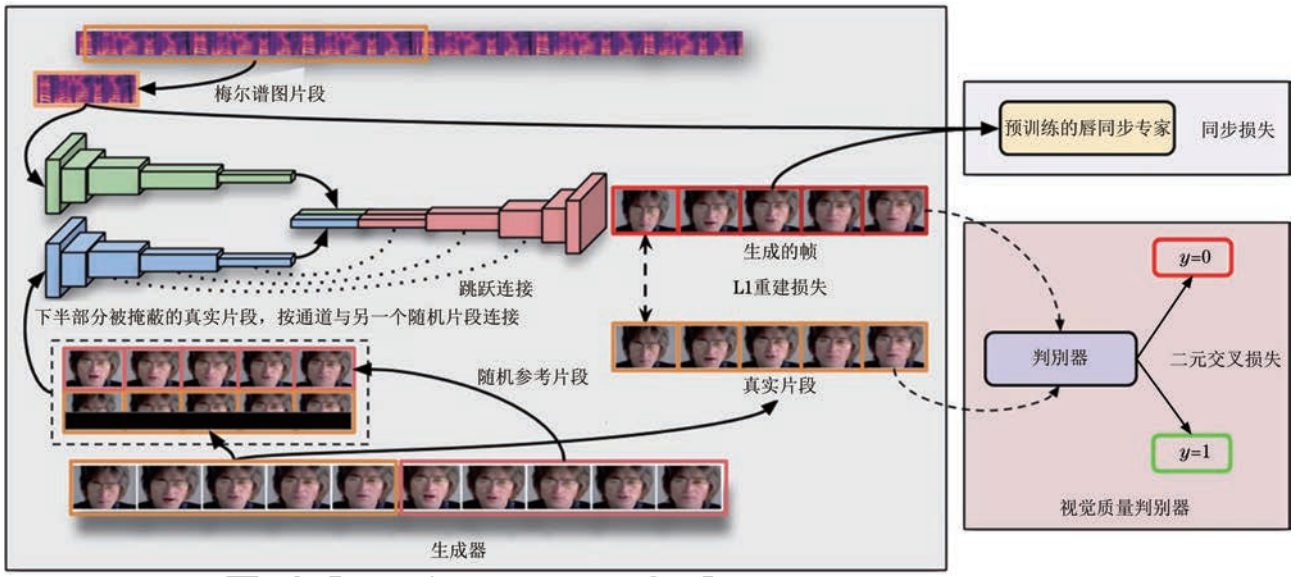

图 2 Wav2Lip 架构[13]

Fig. 2 Wav2Lip architecture[13]

1.3 基于面部属性编辑的伪造

基于面部属性编辑的伪造方法指添加、更改或删除目标身份属性。例如，更换目标对象的发型、胡须、年龄、体重、颜值、眼镜和种族等属性。

CHOI 等[15]提出 StarGAN 模型。传统的 GAN 在多属性编辑时，对每一属性需要单独训练一个生成器，如图 3(a)所示，而该模型可以在多个属性编辑条件下使用单个生成网络生成目标，如图 3(b)所示。GAO 等[16]提出 HifaFace 缓解生成器在合成丰富细节时的压力。此后，XU 等[17]提出 TransEditor 实现高可控性的复杂人脸属性编辑任务。

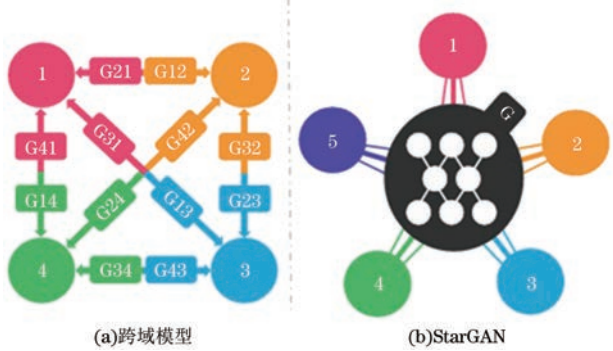

图 3 StarGAN 对比[15]

Fig. 3 StarGAN comparison[15]

早期面部属性编辑更多关注可控性问题，要求对目标属性的正确编辑，然而高可控性可能带来低可扩展性和过度约束问题。于是，研究者针对这些问题提出新的模型。随着技术的进步，人们逐渐将注意力放在精准度上，要求在高可控性的基础上，生成的图像具有丰富细节。

1.4 基于面部生成的伪造

基于面部生成的伪造方法在没有目标身份作为基础的情况下创建伪造角色。

KARRAS 等[18]提出 ProGAN 模型，如图 4 所示，该模型通过渐进式训练的方式，让生成器先生成低清图像，再逐渐升高分辨率，解决了生成高分辨率图像的难题。

KARRAS 等[19]提出 StyleGAN 模型，其包含两个部分。如图 5 所示，其中映射网络针对特征纠缠问题，利用多层全连接层解耦，合成网络用于生成目标。然而，该方法生成的图像上有明显的斑点状伪影。于是，KARRAS 等[20]于 2020 年提出 StyleGAN2 模型，通过调整归一化层的使用，有效避免了斑点状伪影。然而，GAN 难以控制输出特征的精确位置。KARRAS 等[21]于 2021 年提出 StyleGAN3 模型，通过调整 StyleGAN2 的生成器网络结构，使其具有高质量等变性。

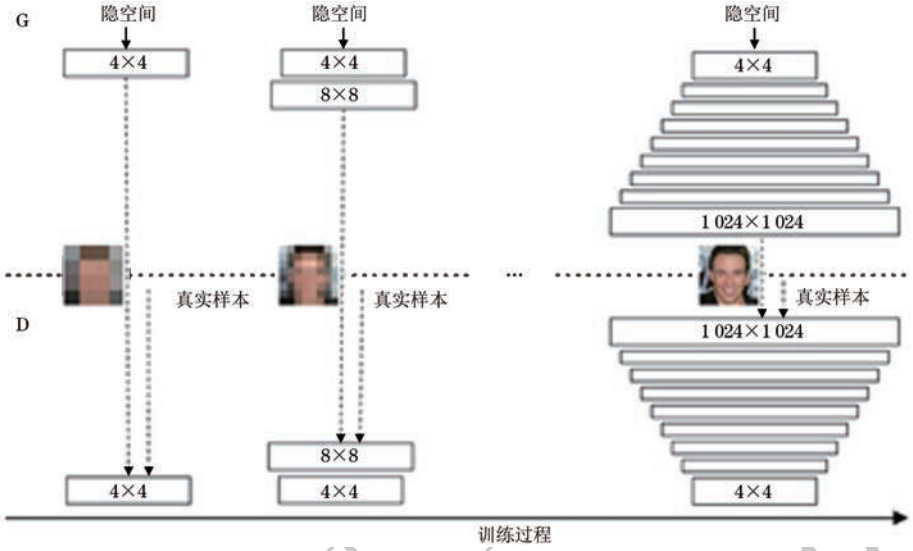

图 4 ProGAN 渐进式训练[18]

Fig. 4 ProGAN progressive training[18]

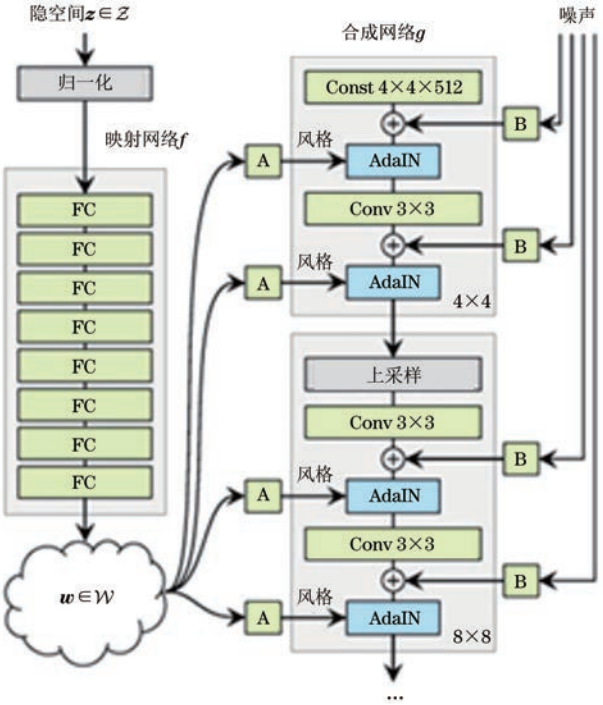

图 5 StyleGAN 架构[19]

Fig. 5 StyleGAN architecture[19]

直接使用 GAN 生成图像较难控制，使用文本信息生成人脸则更加可控。XIA 等[22]提出的 TediGAN 模型生成了文本控制的人脸图像。SUN 等[23]提出的 AnyFace 模型实现了用文本信息生成更多样化的人脸图像。

面部生成主要依靠 GAN 等生成网络生成所需面部，早期的面部生成的可控性较差，随着技术的发展，以 StyleGan 等为代表的生成模型引入各类可控机制，增强了对模型的可控性。

综上，视频伪造分类示意图如图 6 所示。

图 6 视频伪造分类

Fig. 6 Video forgery classification

2 语音伪造

随着深度学习语音处理技术的发展以及人们对于个性化语音合成的需求，语音伪造算法应运而生。语音伪造包括语音合成和语音转换。

2.1 语音合成

语音合成技术也称为文本到语音(TTS)技术，将指定文本信息转换为对应语音。

2.1.1 管道式语音合成

早期的语音合成技术大多基于管道式语音合成结构，将语音合成任务分为文本分析模块和声学系统模块，生成方法包括拼接合成法和参数合成法两种。

1)拼接合成法。

拼接合成法将语音数据库中的语音单元按照一定的规则拼接，生成的语音有着极高的自然度。但由于单词之间的协同发音效果，字级波形的串联听起来不自然[24]。为了合成的语音更自然流畅，拼接合成法需要建立大型语音数据库来提升语音片段的数量和多样性。此外，其会采用一些拼接方法，例如使用特殊技术使音频波形在拼接处尽可能平滑[25]。

虽然拼接合成法可以生成高质量、自然的语音，但需要大量的存储空间和计算资源来管理和搜索语

音数据库，并且需要复杂的算法来保证拼接的语音在声学上是连续和平滑的。此外，如果数据库中的语音数据和需要合成的文本在风格、情感和说话人特征上存在不匹配，那么合成的语音可能会听起来不自然或失真。

2）参数合成法。

参数合成法通过提取和利用语音的声学参数，借助数学模型模拟人类声音。其无需直接操作语音数据库，而是通过模型学习和泛化来适应不同的语音条件。该方法的数据需求相对较低，典型代表为基于隐马尔可夫模型（HMM）的参数式语音合成[26]。

参数合成法利用数学模型对数据进行紧凑表示，能够在有限的数据下通过数学模型的泛化生成新的语音，一定程度上解决了拼接合成法的过大数据库以及单词拼接不平滑问题。然而，该方法需要准确地建模和预测复杂的声学特征，难以在声码器阶段生成高质量的波形。

3）结合深度学习的管道式语音合成。

传统语音合成算法虽然在某些场景下能生成可理解的语音，但受制于模型建模准确性、平滑拼接、数据库等问题，传统算法在自然度、流畅性和音质方面难以达到人类语音的水平。深度学习技术的引入极大地改善了这一状况，合成的语音在多个维度上都更加接近真实人类语言水平。

百度人工智能实验室[27]结合改进后的 WaveNet 声码器提出 Deep Voice 系统，使用神经网络代替传统的文本-语音管道式合成结构实现语音合成。此后，该实验室[28]提出基于注意力机制的全卷积语音合成系统 Deep Voice 3。

深度学习在一定程度上弥补了传统语音合成方法的不足。然而，管道式语音合成系统的模块之间会产生误差积累，需要文本标注以及文本特征和声学特征的强制对齐，会限制语音合成的效果。

2.1.2 端到端语音合成

端到端架构将文本分析模块和声学系统模块连接起来，直接输入文本或者字符，输出对应语音。其极大程度地简化了语音合成系统，降低了对语言学知识的要求。此外，更少的模块有效避免了管道式系统的误差累积问题，在性能上实现了突破。按照选用的模型，端到端的语音合成系统可分为自回归模型与非自回归模型。

1）自回归模型。

自回归模型利用过去时刻的采样点生成下一时刻的采样点。WaveNet[29]是该类模型的典型，如图 7 所示，其基于前面生成的样本来生成后面的样本。Tacotron[30]采用帧水平自回归生成的方式，实现了比 WaveNet 更快的生成速度。

虽然自回归生成模型生成效果不错，但其存在以下三个缺点：第一，生成速度缓慢，在对速度与实时性有较高要求的场所下难以应用；第二，其存在重复吐词或漏词的现象，难以应用于商业环境；第三，其无法细粒度控制语速、韵律及停顿等。

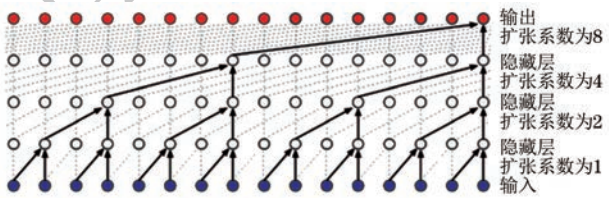

图 7 WaveNet 语音生成[29]

Fig. 7 WaveNet speech generation[29]

2）非自回归模型。

非自回归网络基于并行网络结构，可以并行生成大量语音段，极大提升语音合成速度。FastSpeech[31]基于 Transformer 架构，如图 8 所示，其实现梅尔频谱图的并行生成，加快生成速度。HiFi-GAN[32]基于 GAN 通过并行处理不同周期信号。Parallel Tacotron[33]使用变分自编码器（VAE）和迭代谱图损失提高语音合成的自然度，并且结合轻量级卷积神经网络实现自注意力机制提高生成效率。Diff-TTS[34]学习以文本为条件的梅尔谱图分布，通过扩散模型将噪声信号转换为梅尔谱图。

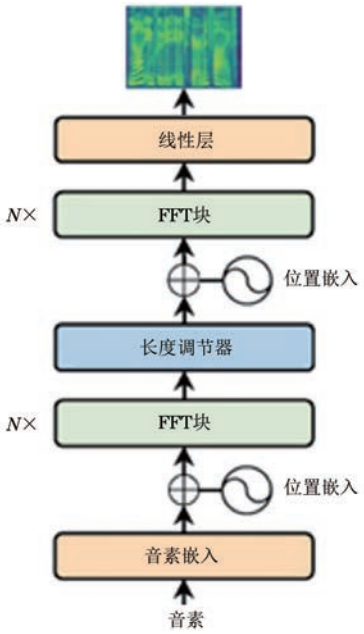

图 8 FastSpeech 架构[31]

Fig. 8 FastSpeech architecture[31]

非自回归语音合成方法有效解决了自回归模型存在的问题，其合成速度更快，有更强的可控

性,容易解决重复、漏词等问题。然而,非自回归模型构造流程复杂,一般需要借助外部工具对齐文本与声学特征,极大地增加了模型构建的复杂度与成本。

综上,语音合成伪造分类示意图如图 9 所示。

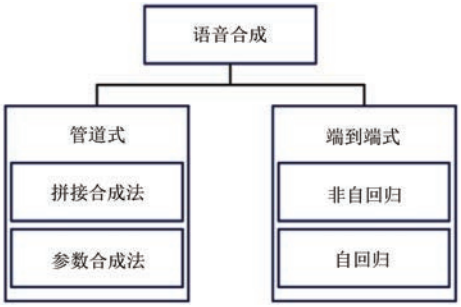

图 9 语音合成伪造分类

Fig. 9 Speech synthesis forgery classification

2.2 语音转换

语音转换通过转换源语音音色等语音特性生成目标语音,使目标语音听起来有目标人物的特性。

语音转换可以分为平行数据语音转换和非平行数据语音转换两种。平行数据表示源说话人和目标说话人的训练数据成对且内容相同。非平行数据表示源说话人和目标说话人的训练数据内容不同。

2.2.1 平行数据

传统语音转换主要使用平行训练数据,使用前一般采用动态时间规整(DTW)等对齐方法对齐语音帧。在建模时,一般采用高斯混合模型(GMM)[35]等参数模型法或矢量量化[36]等非参数模型法。

基于平行数据的语音转换有着较大的局限性:首先,数据集制作成本高昂,难以扩充;其次,它们只能用于一对一或多对一的语音转换,无法做到任意说话人到任意说话人的语音转换。

2.2.2 非平行数据

非平行数据训练需要从非平行数据中推导出说话者之间的语音片段的映射关系。一些早期的算法有 INCA 算法[37]、非负矩阵分解法[38]等。

深度学习的发展使非平行数据语音转换方法得到迅速发展。XU 等[39]提出的 Flow-VAE VC 模型采用 VAE 提高建模能力。CHOI 等[40]提出的 DDDM-VC 模型引入去噪扩散模型为生成模型中的每个属性实现有效的样式转移。ZHANG 等[41]提出的 DiffGAN-VC 模型实现了非平行多对多语音转换。

基于非平行数据的语音转换实现难度更高,然而深度学习技术使基于非平行数据的语音转换方法得到了迅速的发展。该类语音转换解决了传统平行数据语音转换的数据限制以及说话人限制问题,可以实现更好的语音转换。

综上,语音转换伪造分类示意图如图 10 所示。

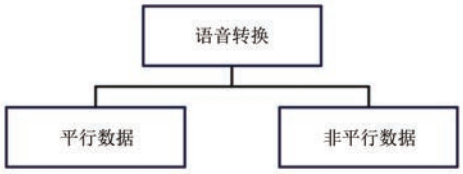

图 10 语音转换伪造分类

Fig. 10 Voice conversion forgery classification

3 伪造数据集

现有数据集根据伪造的模态的不同,可以分为视频伪造数据集、语音伪造数据集以及音视频伪造数据集。

3.1 视频伪造数据集

1) WildDeepfake[42]:该数据集视频来源于网络,内容丰富多样,视觉效果更真实,更符合真实生活场景,换脸方法多样,更难检测。

2) Deeper Forensics-1.0[43]:该数据集规模较其他数据集大了 1 个数量级,并且生成质量好,多样性高。

3) DFFD(Diverse Fake Face Dataset)[44]:该数据集伪造方法涉及面部交换、表情交换、面部属性编辑、面部生成。

4) KoDF[45]:该数据集是一个大规模的以韩国人为对象的深度伪造视频数据集,伪造方法包括面部交换和面部重建。

5) GBDF[46]:该数据集伪造方法包括面部交换和表情交换,重点针对性别分布不均衡问题,实现了更公平的数据分布。

6) DF-Platter[47]:该数据集使用 FSGAN、FaceSwap、FaceShifter 3 种方法进行面部伪造。

视频伪造数据集汇总如表 1 所示,其中“—”表示原文献中没有该信息或指标值,下同。

表 1 视频伪造数据集汇总

Table 1 Summary of video forgery datasets

名称	提出年份	真实样本数/个	伪造样本数/个	伪造方法
WildDeepfake	2020	3 805	3 509	—
Deeper Forensics-1.0	2020	50 000	10 000	FS
DFFD	2020	1 000	3 000	FS/FR/FE/FG
KoDF	2021	4 000	4 000	FS/FR
GBDF	2023	2 000	8 000	FS/FR
DF-Platter	2023	764	132 496	FS

3.2 语音伪造数据集

1)ASVspoof 挑战赛数据集:该数据集有多个版本。其中,2021 版[48]有 LA、PA、DF 3 种任务场景,其中,LA 和 DF 场景伪造涉及语音合成和语音转换。

2)HAD 数据集[49]:该数据集有 3 个类别,分别是全伪造类、部分伪造类以及全真类,构建了部分伪造数据集,伪造方法为语音合成。

3)ADD 挑战赛数据集:该数据集源自音频深度合成检测挑战赛,目前有 ADD2022[50] 以及 ADD2023[51]版本,两个版本伪造均涉及语音合成和语音转换[52],语言为中文。

4)LibriSeVoc 数据集[53]:该数据集是针对神经声码器的数据集,使用 6 个最先进的神经声码器生成语音样本。

5)DFADD 数据集[54]:该数据集语言为英语,主要针对 TTS 伪造,填补了语音 Deepfake 数据集针对扩散模型和流模型 TTS 合成系统的空缺。

6)Codecfake 数据集[55]:该数据集是针对大语言模型 Deepfake 的数据集。

7)CVoiceFake 数据集[56]:该数据集是包括英语、汉语、德语、法语和意大利语 5 种语言的大规模多语言数据集,利用了多种高性能且经典的语音合成技术。

语音伪造数据集汇总如表 2 所示。

表 2 语音伪造数据集汇总

Table 2 Summary of audio forgery datasets

名称	提出年份	真实样本数/个	伪造样本数/个	伪造方法
ASVspoof 2021(LA)	2021	14 816	133 360	TTS/VC
ASVspoof 2021 (DF)	2021	14 869	519 059	TTS/VC
HAD	2021	53 612	53 612/53 612	TTS
ADD2022	2022	3 012	24 072	TTS/VC
ADD2023	2023	3 012	24 072	TTS/VC
LibriSeVoc	2023	13 201	79 206	TTS/VC
DFADD	2024	44 455	163 500	TTS
Codecfake	2024	132 227	925 939	—
CVoiceFake	2024	—	—	TTS/VC

3.3 音视频伪造数据集

1)FakeAVCeleb[57]:该数据集包括 4 种伪造类型:真视频真语音,真视频假语音,假视频真语音,假视频假语音。

2)Lav-DF[58]:该数据集伪造涉及面部重建以及语音合成。

3)PolyGlotFake[59]:该数据集伪造涉及视频和语音两个方面,包括 7 种语言,生成技术新颖。

4)AV-Deepfake1M[60]:该数据集伪造涉及面部重建以及语音合成。

5)DfakeAVMiT[61]:该数据集伪造涉及面部交换、面部重建以及语音合成。

音视频伪造数据集汇总如表 3 所示。

表 3 音视频伪造数据集汇总

Table 3 Summary of audio and video forgery datasets

名称	提出年份	真实样本数/个	伪造样本数/个	方法
FakeAVCeleb	2021	500	19 500	FS/FR/TTS
Lav-DF	2022	36 431	99 873	FR/TTS
PolyGlotFake	2023	766	14 472	FR/TTS
AV-Deepfake1M	2023	286 721	860 039	FR/TTS
DfakeAVMiT	2023	540	6 480	FS/FR/TTS

4 视频伪造鉴别

传统视频鉴伪主要基于图像取证方法,通过传统信号处理方法提取图像特征,利用图像频域和统计特征鉴伪,例如局部噪声估计[62]等。然而,这些方法多数依赖特定的伪造特征。视频伪造技术的发展使伪造视频可以克服这些缺陷,导致检测效果不佳。

合成视频通常存在一些与真实世界不一致的信息,可以基于生理信息一致性、图像篡改痕迹和时域一致性鉴伪。

4.1 基于生理信息一致性的鉴伪

合成视频往往会忽视人的真实生理信息,因此,可以基于生理信息鉴伪。CIFTCI 等[63]利用面部的 3 个区域提取光电容积脉搏信号鉴伪。HALIASSOS 等[64]根据嘴部运动检测假视频。WANG 等[65]设计分析整个面部和嘴部区域动态信息的两个分支进行鉴伪。NGUYEN 等[66]通过眉毛匹配进行鉴伪。

基于生理信息一致性的鉴伪方法有明确的检测对象,如脉搏、心率等,可以使用一些预训练的提取脉搏、心率等生理信息的模型来提取特征,简化前端训练过程。然而,其严重依赖于可检测的生理信号,而随着深度生成技术的进步,伪造内容逐渐克服了各类生理特征的缺陷,导致基于生理信息一致性的鉴伪模型失去了可依托的检测依据。

4.2 基于图像篡改痕迹的鉴伪

合成视频通常会存在一些伪造痕迹,可根据伪造痕迹鉴伪。NIRKIN 等[67]检测人脸与头发、耳朵、脖子等差异鉴伪。KHORMAIL 等[68]利用

Transformer 模型从局部图像特征和不同伪造尺度下像素的全局关系中学习隐藏的扰动痕迹。

基于图像伪造痕迹的鉴伪方法通过提取合成过程中不可避免且共通的伪造痕迹鉴伪，具有更高的适应能力。然而，其严重依赖明显的图像伪造痕迹，随着伪造技术的提升，通过扩散模型等新型生成技术已经可以合成极其逼真的图像，伪造痕迹的隐蔽性大幅提升，只关注图像的伪造痕迹显然已经不够了。

4.3 基于时域一致性的鉴伪

伪造视频是通过逐帧操作生成的，帧与帧之间存在不自然的图像过渡或时间不一致性，如图 11 所示，其可以利用视频的时域一致性进行鉴伪。GU 等[69]使用空间、时间模块提取特征并整合信息，捕捉视频中的时空不一致性来鉴伪。ZHANG 等[70]检测视频的面部局部分布的时空不一致来鉴伪。XU 等[71]提取视频的连续 4 个帧鉴伪。

基于时域一致性的鉴伪方法比前两种方法更加稳健，充分利用了视频的单模态信息，但其对帧与帧之间的变化很敏感，并且没有解决单帧图像的鉴伪问题，而是转换到时域角度解决鉴伪问题。

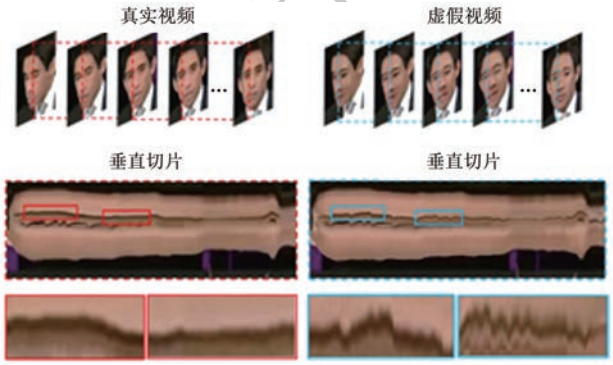

图 11 时域差异性[69]

Fig. 11 Differences in time domains[69]

4.4 基于数据驱动的鉴伪

除了基于一致性的鉴伪方法以外，还有不少研究者关注基于数据驱动的方式来鉴伪，这类鉴伪方式使用大量的数据训练模型来鉴伪。ZHAO 等[72]提出多注意力深度伪造检测网络来鉴伪。LI 等[73]对伪象与不相关信息解耦来鉴伪。HUA 等[74]使用补丁通道将面部潜在特征转换为多通道可解释特征，促进可解释的人脸伪造检测。张怡雯等[75]提取视频图像局部和全局特征，利用稀疏交叉注意力融合鉴伪。QIAO 等[76]设计了完全无监督的 Deepfake 探测器，使用伪标签生成器来标记训练样本，并通过对比学习完善判别特征。LU 等[77]提出使用远距离注意力分别建模时间和空间特征来进行细粒度鉴伪。ZHOU 等[78]提出一种新的测试时域泛化框架 FAS，该框架利用测试数据来提高模型的泛化性。

基于数据驱动的鉴伪方法依靠大量数据训练模型来鉴伪，通过数据驱动模型寻找伪造特征。该类方法没有特别关注视频的某一类特征，如生理信息、伪造痕迹等特定特征，因此可学习的信息更多，准确度可能会更高。然而，该类方法对数据分布依赖性较高，要求测试对象与训练数据具有相近的数据分布，这也导致其存在泛化性的问题。

综上，视频鉴伪分类示意图如图 12 所示。

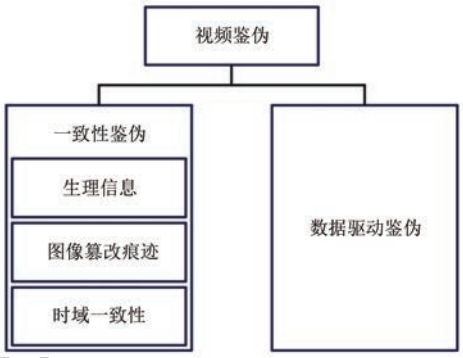

图 12 视频鉴伪分类

Fig. 12 Video counterfeit detection classification

5 语音伪造鉴别

语音鉴伪按照功能结构可分为前端和后端。前端分析语音信号并提取具有区分性的特征，后端根据提取的特征鉴伪。检测器可以将前端与后端结合成管道式鉴别器来鉴伪，也可以使用端到端鉴伪系统以取得更好的效果。

5.1 管道式系统

5.1.1 前端系统

传统前端系统主要使用信号处理提取特征。其中，梅尔频率倒谱系数[79]与线性频率倒谱系数[80]经常作为语音鉴伪的特征。

深度神经网络可以比传统特征更好地提取伪造痕迹。MARTÍN-DOÑAS 等[81]使用预训练的 Wav2Vec2 作为特征提取器来鉴伪。GUO 等[82]使用 WavLM 获得帧级特征鉴伪。XIE 等[83]基于 Wav2Vec2 前端在掩码特征编码器上求解对比任务来训练，可以表示语音情境化信息。

5.1.2 后端系统

后端系统利用前端系统提取的特征进行分析并分类。传统的分类器大多基于机器学习方法，其中最广泛的是 GMM 分类器[84]和支持向量机[85]。

近年来，伪造语音检测系统的后端大多基于神经网络分类器，这类分类器在性能方面超越了

传统分类器。例如，SENet[86]通过挤压-激励操作显示建模通道之间的相互依赖性，自适应地重新校准通道特征响应；Transformer[87]捕捉输入序列的全局和局部关系，在伪造音频定位任务中表现出色。

管道式系统将鉴伪系统分为前端和后端两个模块，增强了可解释性，降低了训练压力。模型可以针对不同任务替换相应模块，减轻了替换成本，增加了灵活性。然而，管道式系统的模块之间存在误差累积，需要合理地设计两个模块和它们的关系来降低模块间的误差传递。因此，管道式系统难以实现多模块且高复杂度的模型。

5.2 端到端系统

端到端系统将特征提取器和分类器结合起来联合优化，十分有竞争力。GE 等[88]提出的 RawPC 自动学习语音深度伪造和欺骗检测解决方案，同时优化其他网络组件和参数。TAK 等[89]引入频谱和时间子图以及图池化策略训练光谱-时间图注意力网络。

端到端系统采用整体训练的方式，可以在训练过程中自动优化模块间关系，一定程度上降低了设计难度，因此可以实现更复杂的系统，在鉴伪性能上更强。然而，端到端系统的整体训练加大了训练压力。此外，该系统的可解释性差，更像一个黑盒，设计者难以透过模型看到内部数据的处理过程。

综上，语音鉴伪分类示意图如图 13 所示。

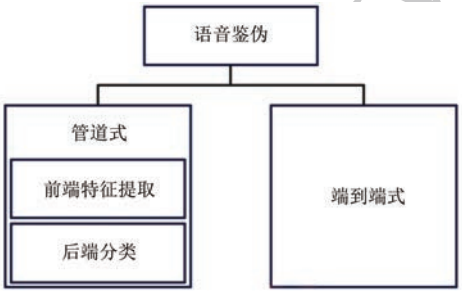

图 13 语音鉴伪分类

Fig. 13 Audio counterfeit detection classification

6 音视频多模态鉴伪

单模态鉴伪方法受到输入模态的限制，难以应对多模态的灵活伪造，而多模态鉴别伪造方法将音视频特征信息融合鉴伪，可以获得更好的鉴伪性能。按照其训练方式可以分为端到端训练方式和预训练方式。

6.1 端到端训练方式

端到端训练的检测方法较为常见，通过融合视听特征，对比视听特征不一致等方式得到鉴伪结果。

6.1.1 集成学习

集成学习使用两个模态的信息在决策层上投票来鉴伪。ZHOU 等[90]提出一种基线模型，如图 14 所示，该模型对视频流和音频流独立训练并进行预测，最后将预测的概率结合进行鉴伪。AVFakeNet[91]使用 Dense Swin Transformer 网络组成音频和视频分支，通过比较两个分支的预测标签得到最终标签。

集成学习利用音频和视频分支的预测结果投票来得到最终预测结果。这种方法虽然很直观，但并没有利用音视频的内在联系，多模态鉴伪模型的性能将极大取决于单个分支的性能。

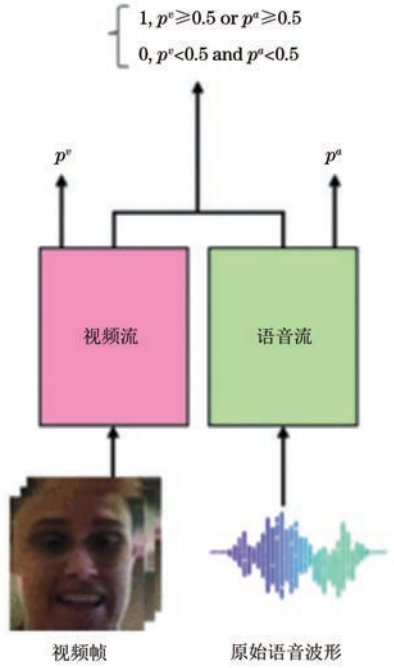

图 14 集成学习架构[90]

Fig. 14 Ensemble learning architecture[90]

6.1.2 联合学习

基于联合学习的鉴伪方法利用模态信息融合鉴伪，可以更充分地利用音视频的内在联系。ZHOU 等[90]应用交叉注意力聚合音频和视觉特征。RAZA 等[92]提出 Multimodaltrace 模型，结构如图 15 所示，其融合了音视频的特征进行鉴伪。此外，其使用多标签多类的最终分类，充分考虑了音频和视频的分类。WANG 等[93]提出 AVT2-DWF 模型，该模型通过带有 n 帧标记化策略编码器和 Transformer 结构来放大模态内伪造线索，使用基于跨模态注意力的 DWF 模块融合音视频模态信息放大跨模态伪造线索。

基于联合学习的鉴伪方法更好地融合了语音和视频模态的信息。然而，其忽略了模态间的相关性等关系，仅在最后分类部分使用损失函数对

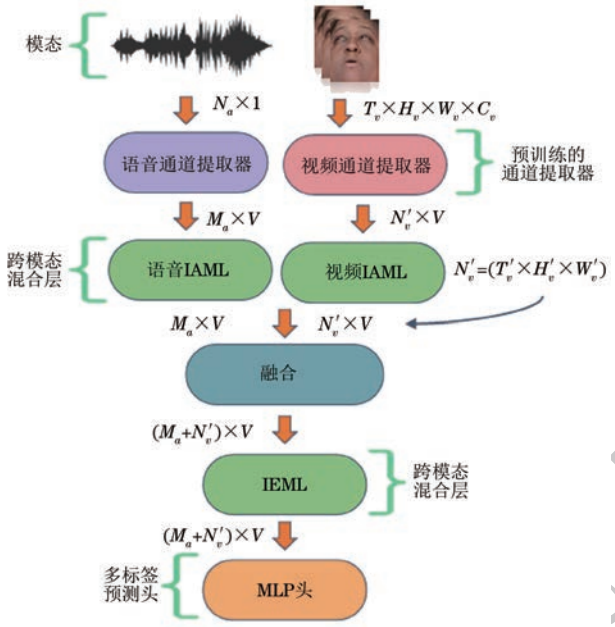

图 15 Multimodaltrace 架构[92]

Fig. 15 Multimodaltrace architecture[92]

整体进行优化，这样的结构对每个模态分支的优化欠佳。

6.1.3 视听一致性

基于视听一致性的鉴伪方法关注音视频之间的各类不一致问题，包括时间不同步、情感不一致等问题。MITTAL 等[94]研究了音视频的情绪不匹配问题，并以此作为多模态鉴伪线索。然而，该方法严重依赖情绪识别模型，如果情绪识别模型性能不佳，则最终的鉴别性能也会受到影响。CHUGH 等[95]使用 MDS 来衡量音视频差异。CHENG 等[96]通过面部和音频的内在相关性，从语音-人脸匹配视角提出 VFD 模型。

上述多模态鉴伪方法仅考虑了视频的伪造，将真实的音频当作辅助监控信号，没有考虑音频模态的伪造问题。YANG 等[61]提出的 Avoid-DF 模型使用双向交叉注意力模块对多模态交互建模。KATAMNENI 等[97]提出 MIS-AVoiDD 模型，其包含两个解码器，一个提取音视频同质特征，另一个提取音视频异质特征，缓解了模态间分布差异。LIU 等[98]提出 MCL 模型，该模型利用跨模态对比学习探索模态内和跨模态伪造线索，减少模态差距并探索多模态伪造痕迹。LIU 等[99]针对伪造时间定位问题，通过具有自注意力的嵌入级融合机制进行伪造定位，并且利用包括视听不一致和时间不一致的多维对比损失来优化模型，如图 16 所示。

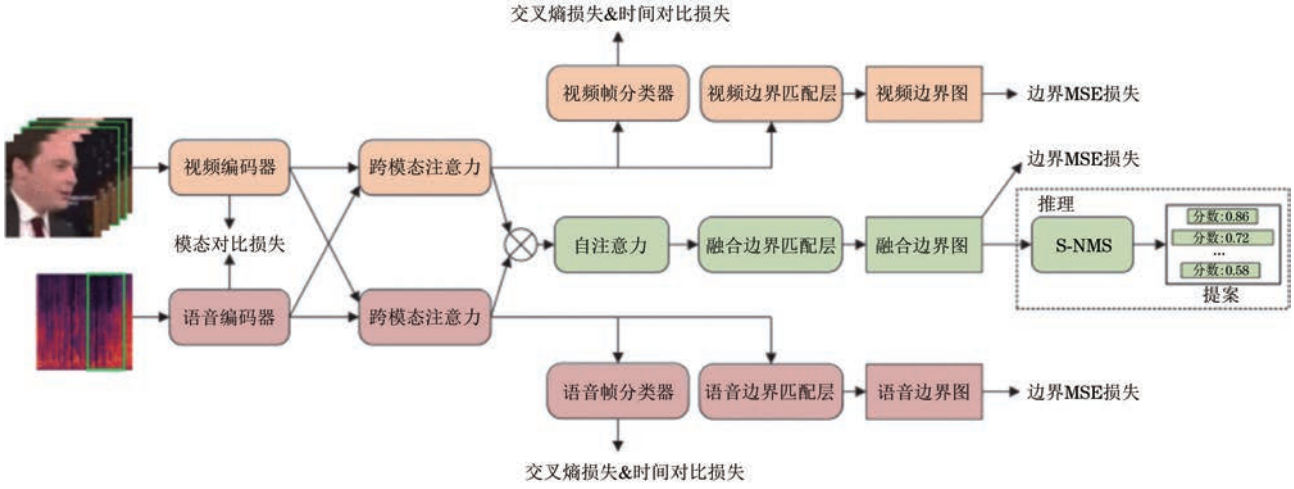

图 16 基于嵌入级融合和多维对比损失的多模态鉴伪[99]

Fig. 16 Multimodal counterfeit detection based on embedding-level fusion and multi-dimensional contrastive loss[99]

近年来，人们发现伪造视频的独立模态伪造会导致学习表示的不一致和不确定性，仅用最终的分类结果来指导每个模态的优化容易导致混乱。针对此问题，ZOU 等[100]提出模态内正则化和跨模态正则化方法指导多模态模型优化，其结构如图 17 所示，拉近真视频的音视频特征距离，拉远合成视频的特征距离。此外，通过拉近或拉远不同视频的单模态特征来细化单模态表示。

基于视听一致性的鉴伪方法不仅融合了模态信息，还充分利用了模态间的关系。该方法不仅会在最终分类中加入损失函数，一般还会在模态融合区域使用对比学习构建更全面的损失函数来优化整个网络结构。这种方法对每个模态分支的优化都有了更精细的把控，可以获得更好的性能。

6.2 预训练方式

预训练方式先在真实视频上训练，学习真实视频的特点，再在真伪数据集上进行微调训练来提升泛化性能，通过对真实视频的准确建模来鉴伪。

OORLOFF 等[101]提出一种两阶段跨模态学习方法 AVFF，如图18所示，第一阶段使用掩码和特

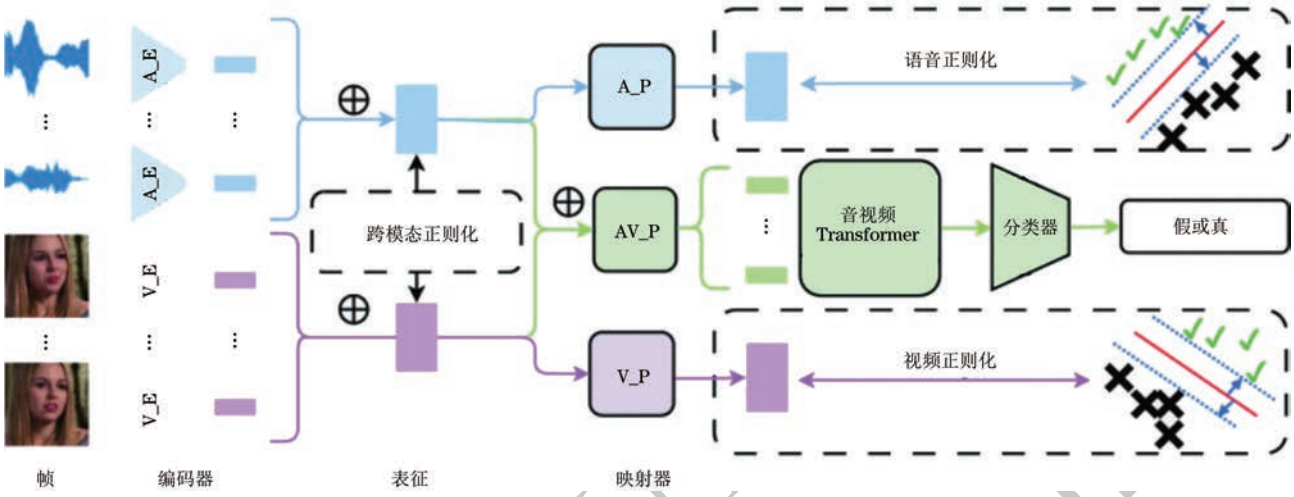

图 17 正则化的多模态架构[100]

Fig. 17 Regularized multimodal architecture[100]

征融合等手段捕捉真实的视听对应关系，第二阶段对学习到的表示进行微调。YU 等[102]提出 PVASS-MDD 架构，在 PVASS 辅助阶段将两个增强的视觉视图与相应的音频线索相关联来探索真实音视频对应关系，在 MDD 阶段，使用冻结的 PVASS 网络来鉴伪。LI 等[103]在真实数据上预训练音频识别模型和视觉识别模型提取特征，再计算两个序列之间的距离来鉴伪。FENG 等[104]使用自回归 Transformer 模型为视听同步特征分配概率来鉴伪。

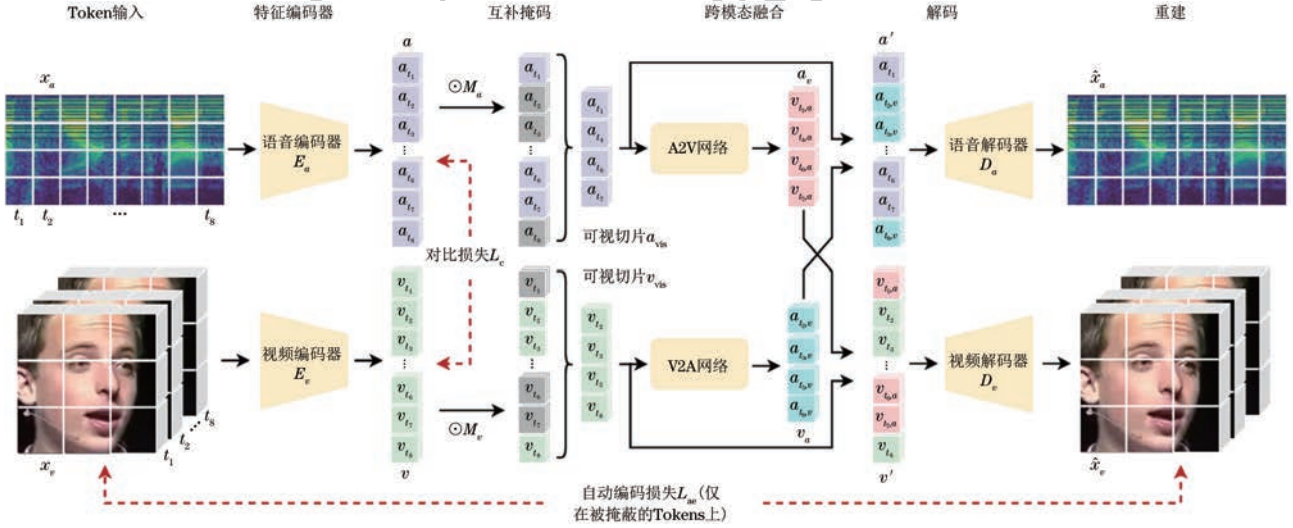

图 18 AVFF 架构[101]

Fig. 18 AVFF architecture[101]

预训练方式虽然泛化能力更好，但其第一阶段仅在真实视频上学习，对数据集要求高，否则容易错过关键伪造特征而产生误判。此外，其模型的精度要求高，否则难以对任意类型的伪造产生响应。

音视频多模态鉴伪汇总如表 4 所示。多模态鉴伪分类示意图如图 19 所示。

7 未来展望

7.1 鉴伪技术问题

1）泛化性能问题：目前的各式鉴伪算法对已知伪造方法取得了不错的效果。然而，不同的生成算法侧重点不同、特征不同，导致泛化性问题。现有的多数应对方法是将新的生成算法加入到伪造数据集中做增强训练[105]。这种方法时效性差，在面对新的生成手段时仍有可能乏力。另一些学者提出了其他增强泛化性能的方法，但这些方法可能会造成其他方面的问题，例如，音视频多模态鉴伪时使用预训练方式[101-104]虽然可以一定程度上增强泛化性，但其对模型和数据要求严格。因此，鉴伪泛化性问题仍旧是目前的一大难点，并且亟待解决。

表 4 音视频多模态鉴伪汇总

Table 4 Audio and video multimodal counterfeit detection summary

年份	文献来源	类别
2021	文献[90]	集成学习
2023	文献[91]	集成学习
2021	文献[90]	联合学习
2023	文献[92]	联合学习
2024	文献[93]	联合学习
2020	文献[94]	视听一致性学习
2020	文献[95]	视听一致性学习
2023	文献[96]	视听一致性学习
2023	文献[97]	视听一致性学习
2023	文献[98]	视听一致性学习
2023	文献[61]	视听一致性学习
2024	文献[99]	视听一致性学习
2024	文献[100]	视听一致性学习
2023	文献[102]	预训练方式
2023	文献[104]	预训练方式
2024	文献[101]	预训练方式
2024	文献[103]	预训练方式

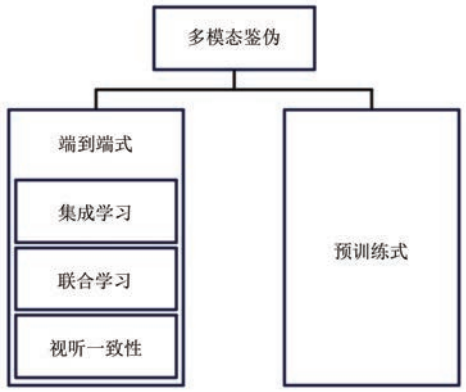

图 19 多模态鉴伪分类

Fig. 19 Multimodal counterfeit detection classification

2)抗干扰问题:目前多数鉴伪算法在实验环境中训练以及测试,但现实世界远比实验环境复杂。例如,物理空间中环境光照、噪声的干扰,数字空间中各种压缩算法或音视频后处理的干扰,人为加入的对抗样本等干扰均会影响模型性能,导致模型在实验室环境中的高识别率在实际应用中明显下降。

3)实时性问题:目前的鉴伪算法大多侧重检测准确率、召回率等指标。然而,现实环境中存在视频会议、网络直播等场景,如何在有限算力下处理实时传输的视频仍是一个问题。

4)数据集问题:一方面,目前部分公开数据集分辨率较低,随着网络带宽的增加,高分辨率视频逐渐占据主流,低分辨率下训练的模型可能难以适应现实环境中的高分辨率视频,导致模型实用性不高。另一方面,数据集的标注信息不够全面,如未标注伪造方法、伪造区域、其他音视频编辑处理方法,部分数据集来源不可靠,特别是其中所谓真实视频来自互联网,不能确保其未经过其他音视频处理软件加工。

5)可解释性问题:目前的大多数鉴伪方法注重分类问题,只是将视频、语音正确分类,却没有过多关注判别原因,解释性差。一方面,鉴伪模型难以做到完全准确的分类,需要用户做出一定程度的判断,因此,在鉴伪模型投入使用时,如果无法向用户合理指出伪造区域或解释判伪原因,可能难以博得用户的信任。另一方面,鉴伪模型类似黑盒,如果可以加强其可解释性,有助于推动鉴伪模型的进步。

7.2 未来研究方向

1)研究泛化性强的算法:目前的算法在泛化性能上表现不佳,难以对训练集数据中未出现的伪造方法准确分类。为了提升模型的泛化性,可从以下 2 个方面进行研究。第 1 种研究方案以数据为中心,首先可以探索更多的伪造类型,通过更多样化的数据来增强泛化性,其次可以对数据预处理,例如平移、旋转、加噪等方式,最后可以使用零样本学习等训练方式。第 2 种研究方案以模型为中心,首先可以通过调整模型结构,增强模型的适应能力,其次可以调整目标函数或采用一些正则化方法,最后可以通过调整模型的超参数,寻找最优配置。

2)研究适应复杂环境的鉴伪算法:目前鉴伪算法在面对复杂环境,例如噪声环境或数据缺失等异常情况时,表现不佳,鲁棒性差,可以在训练时通过加噪、缩放、压缩等数据增强手段提升模型的鲁棒性。

3)主动防御算法:目前鉴伪算法都是被动防御,可以通过主动向网络视频中注入一些标志来辅助鉴伪,也可以加入一些噪声扰动来弱化生成模型生成视频的能力[106]。

4)提升检测算法的实时性:在实际应用中,鉴伪算法经常面对实时传输的视频,目前这方面研究还不是很深入,因此需要研究更好的实时检测算法。

5)制作更好的数据集:数据集作为模型训练与学习的基础,其重要性不言而喻。然而,如何处理好音视频分辨率和数据可信性仍是问题。因此,需要研究制作高分辨率、细粒度标注、来源可信的数据集。

6)研究可解释性强的鉴伪模型:鉴伪模型的可解释性在分类任务中至关重要。然而,目前的大多

数工作并没有重视这方面的研究。一些可行的研究方向存在以下现状：首先，伪造视频、语音并不一定是全部伪造的，有时仅有部分区域是伪造的，因此，对其伪造区域的定位问题是值得深入研究的一个方向[99]；其次，对于可解释性语言生成，鉴伪模型虽然是分类模型，但也需要面对用户，可以为鉴伪模型加入一些生成类算法，通过合成解释性文字等方式向用户解释[107]。

8 结束语

近年来，深度学习的飞速发展带动了合成技术的发展，出现了一系列深度伪造音视频技术，可以模仿受害者的面部、语音或操纵受害者的面部、语音特征等，其滥用严重侵犯公民的合法权益，损害企业商业信誉，威胁国家安全和公共安全。本综述归纳了单模态伪造方法及单模态、多模态伪造鉴别方法，进行了科学的分类与分析，指出了目前鉴伪方法存在的问题，并探讨了未来的研究方向，以期推动深度伪造鉴别技术的进一步发展。

参考文献

[1] 李旭嵘, 纪守领, 吴春明, 等. 深度伪造与检测技术综述[J]. 软件学报, 2021, 32(2): 496-518.
LI X R, JI S L, WU C M, et al. Survey on deepfakes and detection techniques[J]. Journal of Software, 2021, 32(2): 496-518. (in Chinese)

[2] 任延珍, 刘晨雨, 刘武洋, 等. 语音伪造及检测技术研究综述[J]. 信号处理, 2021, 37(12): 2412-2439.
REN Y Z, LIU C Y, LIU W Y, et al. A survey on speech forgery and detection[J]. Journal of Signal Processing, 2021, 37(12): 2412-2439. (in Chinese)

[3] 梁瑞刚, 吕培卓, 赵月, 等. 视听觉深度伪造检测技术研究综述[J]. 信息安全学报, 2020, 5(2): 1-17.
LIANG R G, LÜ P Z, ZHAO Y, et al. A survey of audiovisual deepfake detection techniques[J]. Journal of Cyber Security, 2020, 5(2): 1-17. (in Chinese)

[4] 李泽宇, 张旭鸿, 蒲誉文, 等. 多模态深度伪造及检测技术综述[J]. 计算机研究与发展, 2023, 60(6): 1396-1416.
LI Z Y, ZHANG X H, PU Y W, et al. A survey on multimodal deepfake and detection techniques[J]. Journal of Computer Research and Development, 2023, 60(6): 1396-1416. (in Chinese)

[5] FaceSwap. FaceSwap Github[EB/OL]. [2024-11-14]. https://github.com/MarekKowalski/FaceSwap.

[6] Deepfakes. Deepfakes Github[EB/OL]. [2024-11-14]. https://github.com/Deepfakes/faceswap.

[7] LI L Z, BAO J M, YANG H, et al. FaceShifter: towards high fidelity and occlusion aware face swapping[EB/OL]. [2024-11-14]. https://arxiv.org/abs/1912.13457.

[8] XU C, ZHANG J N, HUA M, et al. Region-aware face swapping[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2022: 7622-7631.

[9] LIU Z A, LI M M, ZHANG Y, et al. Fine-grained face swapping via regional GAN inversion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2023: 8578-8587.

[10] ZHAO W L, RAO Y M, SHI W K, et al. DiffSwap: high-fidelity and controllable face swapping via 3D-aware masked diffusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2023: 8568-8577.

[11] BALIAH S, LIN Q L, LIAO S C, et al. Realistic and efficient face swapping: a unified approach with diffusion models[EB/OL]. [2024-11-14]. https://arxiv.org/abs/2409.07269.

[12] THIES J, ZOLLHÖFER M, THEOBALT C, et al. Headon: real-time reenactment of human portrait videos[J]. ACM Transactions on Graphics, 2018, 37(4): 1-13.

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

[14] LIANG B R, PAN Y, GUO Z Z, et al. Expressive talking head generation with granular audio-visual control[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2022: 3377-3386.

[15] CHOI Y, CHOI M, KIM M, et al. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D.C., USA: IEEE Press, 2018: 8789-8797.

[16] GAO Y, WEI F Y, BAO J M, et al. High-fidelity and arbitrary face editing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2021: 16115-16124.

[17] XU Y B, YIN Y Q, JIANG L M, et al. TransEditor: Transformer-based dual-space GAN for highly controllable facial editing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2022: 7673-7682.

[18] KARRAS T, AILA T, LAINE S, et al. Progressive growing of GANs for improved quality, stability, and variation[EB/OL]. [2024-11-14]. https://arxiv.org/abs/1710.10196.

[19] KARRAS T, LAINE S, AILA T M. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2019: 4396-4405.

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

[21] KARRAS T, AITTALA M, LAINE S, et al. Alias-free generative adversarial networks[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc., 2021: 1-12.

[22] XIA W H, YANG Y J, XUE J H, et al. TediGAN: text-guided diverse face image generation and manipulation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2021: 2256-2265.

[23] SUN J X, DENG Q Y, LI Q, et al. AnyFace: free-style text-to-face synthesis and manipulation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE

Press, 2022: 18666-18675.
[24] RABINER L R, SCHAFER R W. Introduction to digital speech processing[J]. Foundations and Trends® in Signal Processing, 2007, 1(1/2): 1-194.
[25] BALESTRI M, PACCHIOTTI A, QUAZZA S, et al. Choose the best to modify the least: a new generation concatenative synthesis system[C]//Proceedings of the 6th European Conference on Speech Communication and Technology. Budapest, Hungary: ISCA, 1999: 2291-2294.
[26] YURI N, SHINNOSUKE T, TOMOKI T, et al. HMM-based speech synthesis system with prosody modification based on speech input[J]. IEICE Technical Report Speech, 2014, 114: 81-86.
[27] ARIK S O, CHRZANOWSKI M, COATES A, et al. Deep voice: real-time neural text-to-speech[EB/OL]. [2024-11-14]. https://arxiv.org/abs/1702.07825.
[28] PING W, PENG K, GIBIANSKY A, et al. Deep Voice 3: 2000-speaker neural text-to-speech[C]// Proceedings of International Conference on Learning Representations. Vancouver, Canada: [s. n.], 2018: 1094-1099.
[29] VAN DEN OORD A, DIELEMAN S, ZEN H G, et al. WaveNet: a generative model for raw audio[C]// Proceedings of the Speech Synthesis Workshop. Berlin, Germany: Springer, 2016: 13-15.
[30] WANG Y X, SKERRY-RYAN R, STANTON D, et al. Tacotron: towards end-to-end speech synthesis[EB/OL]. [2024-11-14]. https://arxiv.org/abs/1703.10135.
[31] REN Y, RUAN Y J, TAN X, et al. FastSpeech: fast, robust and controllable text to speech[EB/OL]. [2024-11-14]. https://arxiv.org/abs/1905.09263.
[32] KONG J, KIM J, BAE J. HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis[EB/OL]. [2024-11-14]. https://arxiv.org/abs/2010.05646.
[33] ELIAS I, ZEN H G, SHEN J, et al. Parallel Tacotron: non-autoregressive and controllable TTS[C]//Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D. C., USA: IEEE Press, 2021: 5709-5713.
[34] JEONG M, KIM H, CHEON S J, et al. Diff-TTS: a denoising diffusion model for text-to-speech[C]// Proceedings of the Interspeech 2021. Brno, Czechia: ISCA, 2021: 3605-3609.
[35] KAWANAMI H, IWAMI Y, TODA T, et al. GMM-based voice conversion applied to emotional speech synthesis[C]// Proceedings of the 8th European Conference on Speech Communication and Technology. Brno, Czechia: ISCA, 2003: 2401-2404.
[36] WU D Y, LEE H Y. One-shot voice conversion by vector quantization[C]// Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D. C., USA: IEEE Press, 2020: 7734-7738.
[37] SUDA H, KOTANI G, SAITO D. Nonparallel training of exemplar-based voice conversion system using INCA-based alignment technique[C]// Proceedings of the INTERSPEECH 2020. Shanghai, China: ISCA, 2020: 4681-4685.
[38] SUDA H, KOTANI G, SAITO D. INmfCA algorithm for training of nonparallel voice conversion systems based on non-negative matrix factorization[J]. IEICE Transactions on Information and Systems, 2022, 105(6): 1196-1210.
[39] XU L, ZHONG R X, LIU Y, et al. Flow-VAE VC: end-to-end flow framework with contrastive loss for zero-shot voice conversion[C]//Proceedings of the INTERSPEECH 2023. Dublin, Ireland: ISCA, 2023: 2293-2297.
[40] CHOI H Y, LEE S H, LEE S W. DDDM-VC: decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2024: 17862-17870.
[41] ZHANG X L, WANG J Z, CHENG N, et al. Voice conversion with denoising diffusion probabilistic GAN models[C]// Proceedings of Advanced Data Mining and Applications. Berlin, Germany: Springer, 2023: 154-167.
[42] ZI B J, CHANG M H, CHEN J J, et al. WildDeepfake: a challenging real-world dataset for deepfake detection[C]// Proceedings of the 28th ACM International Conference on Multimedia. New York, USA: ACM Press, 2020: 2382-2390.
[43] JIANG L M, LI R, WU W, et al. DeeperForensics-1.0: a large-scale dataset for real-world face forgery detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2020: 2886-2895.
[44] DANG H, LIU F, STEHOUWER J, et al. On the detection of digital face manipulation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2020: 5780-5789.
[45] KWON P, YOU J, NAM G, et al. KoDF: a large-scale Korean DeepFake detection dataset[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Washington D. C., USA: IEEE Press, 2022: 10724-10733.
[46] NADIMPALLI A V, RATTANI A. GBDF: gender balanced DeepFake dataset towards fair DeepFake detection[C]// Proceedings of International Conference on Pattern Recognition, Computer Vision, and Image Processing. Berlin, Germany: Springer, 2023: 320-337.
[47] NARAYAN K, AGARWAL H, THAKRAL K, et al. DF-Platter: multi-face heterogeneous deepfake dataset[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2023: 9739-9748.
[48] LIU X C, WANG X, SAHIDULLAH M, et al. ASVspoof 2021: towards spoofed and deepfake speech detection in the wild[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 2507-2522.
[49] YI J Y, BAI Y, TAO J H, et al. Half-truth: a partially fake audio detection dataset[C]// Proceedings of the INTERSPEECH 2021. Brno, Czechia: ISCA, 2021: 1654-1658.
[50] YI J Y, FU R B, TAO J H, et al. ADD2022: the first audio deep synthesis detection challenge[C]//Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D. C., USA: IEEE Press, 2022: 9216-9220.
[51] YI J, TAO J, FU R, et al. ADD2023: the second audio deepfake detection challenge[C]//Proceedings of Workshop on Deepfake Audio Detection and Analysis. [S. l.]: SAR Press, 2023: 125-130.
[52] SANDERSON C. The VidTIMIT database[EB/OL]. [2024-11-14]. https://www.researchgate.net/publication/2914105_The_VidTIMIT_Database.
[53] SUN C Z, JIA S, HOU S W, et al. AI-synthesized voice detection using neural vocoder artifacts[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Washington D. C., USA: IEEE Press, 2023: 904-912.
[54] DU J W, LIN I M, CHIU I H, et al. DFADD: the diffusion and flow-matching based audio deepfake dataset[EB/OL]. [2024-11-14]. https://arxiv.org/abs/2409.

08731.
[55] LU Y, XIE Y K, FU R B, et al. Codecfake: an initial dataset for detecting LLM-based deepfake audio [C]// Proceedings of the INTERSPEECH 2024. [S. l.]: ISCA, 2024: 1390-1394.
[56] LI X F, LI K, ZHENG Y F, et al. SafeEar: content privacy-preserving audio deepfake detection[C]//Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. New York, USA: ACM Press, 2024: 3585-3599.
[57] KHALID H, TARIQ S, KIM M, et al. FakeAVCeleb: a novel audio-video multimodal deepfake dataset [C]// Proceedings of the Conference on Neural Information Processing Systems Track on Datasets and Benchmarks. Seattle, USA: MIT Press, 2021: 1-12.
[58] CAI Z X, STEFANOV K, DHALL A, et al. Do you really mean that? Content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization[C]// Proceedings of the International Conference on Digital Image Computing: Techniques and Applications (DICTA). Washington D.C., USA: IEEE Press, 2023: 1-10.
[59] HOU Y, FU H T, CHEN C K, et al. PolyGlotFake: a novel multilingual and multimodal DeepFake dataset[EB/OL]. [2024-11-14]. https://arxiv.org/abs/2405.08838.
[60] CAI Z X, GHOSH S, ADATIA A P, et al. AV-Deepfake1M: a large-scale LLM-driven audio-visual deepfake dataset [C]// Proceedings of the 32nd ACM International Conference on Multimedia. New York, USA: ACM Press, 2024: 7414-7423.
[61] YANG W Y, ZHOU X Y, CHEN Z K, et al. AVoiD-DF: audio-visual joint learning for detecting deepfake[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 2015-2029.
[62] PAN X Y, ZHANG X, LYU S W. Exposing image splicing with inconsistent local noise variances[C]//Proceedings of the IEEE International Conference on Computational Photography (ICCP). Washington D. C., USA: IEEE Press, 2012: 1-10.
[63] CIFTCI U A, DEMIR I, YIN L J. FakeCatcher: detection of synthetic portrait videos using biological signals[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(12): 2993-3005.
[64] HALIASSOS A, VOUGIOUKAS K, PETRIDIS S, et al. Lips don't lie: a generalisable and robust approach to face forgery detection [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2021: 5037-5047.
[65] WANG H Y, LIU Z H, WANG S L. Exploiting complementary dynamic incoherence for DeepFake video detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8): 4027-4040.
[66] NGUYEN H M, DERAKHSHANI R. Eyebrow recognition for identifying deepfake videos [C]// Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG). Washington D. C., USA: IEEE Press, 2020: 1-5.
[67] NIRKIN Y, WOLF L, KELLER Y, et al. DeepFake detection based on discrepancies between faces and their context[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 6111-6121.
[68] KHORMALI A, YUAN J S. DFDT: an end-to-end DeepFake detection framework using vision Transformer [J]. Applied Sciences, 2022, 12(6): 2953.
[69] GU Z H, CHEN Y, YAO T P, et al. Spatiotemporal inconsistency learning for DeepFake video detection[C]// Proceedings of the 29th ACM International Conference on Multimedia. New York, USA: ACM Press, 2021: 3473-3481.
[70] ZHANG D C, LIN F Z, HUA Y Y, et al. Deepfake video detection with spatiotemporal dropout Transformer[C]// Proceedings of the 30th ACM International Conference on Multimedia. New York, USA: ACM Press, 2022: 5833-5841.
[71] XU Y T, LIANG J, JIA G Y, et al. TALL: thumbnail layout for deepfake video detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Washington D. C., USA: IEEE Press, 2024: 22601-22611.
[72] ZHAO H Q, WEI T Y, ZHOU W B, et al. Multi-attentional deepfake detection [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2021: 2185-2194.
[73] LI X, NI R R, YANG P P, et al. Artifacts-disentangled adversarial learning for deepfake detection [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(4): 1658-1670.
[74] HUA Y Y, SHI R X, WANG P J, et al. Learning patch-channel correspondence for interpretable face forgery detection [J]. IEEE Transactions on Image Processing, 2023, 32: 1668-1680.
[75] 张怡雯, 蔡满春, 陈永豪, 等. 融合空间特征的多尺度 Deepfake 检测方法[J]. 计算机工程, 2024, 50(7): 240-250.
ZHANG Y W, CAI M C, CHEN Y H, et al. Multi-scale deepfake detection method with fusion of spatial features [J]. Computer Engineering, 2024, 50(7): 240-250. (in Chinese)
[76] QIAO T, XIE S C, CHEN Y L, et al. Fully unsupervised deepfake video detection via enhanced contrastive learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(7): 4654-4668.
[77] LU W, LIU L Y, ZHANG B L, et al. Detection of deepfake videos using long-distance attention [J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(7): 9366-9379.
[78] ZHOU Q Y, ZHANG K Y, YAO T P, et al. Test-time domain generalization for face anti-spoofing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2024: 175-187.
[79] WU Z Z, CHNG E S, LI H Z. Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition[C]//Proceedings of the INTERSPEECH 2012. [S. l.]: ISCA, 2012: 1700-1703.
[80] CHITALE M, DHAWALE A, DUBEY M, et al. A hybrid CNN-LSTM approach for deepfake audio detection [C]// Proceedings of the 3rd International Conference on Artificial Intelligence for Internet of Things (AIIoT). Washington D.C., USA: IEEE Press, 2024: 1-6.
[81] MARTÍN-DOÑAS J M, ÁLVAREZ A. The vicomtech audio deepfake detection system based on Wav2Vec2 for the 2022 ADD challenge[C]//Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D. C., USA: IEEE Press, 2022: 9241-9245.
[82] GUO Y L, HUANG H F, CHEN X, et al. Audio deepfake detection with self-supervised wavlm and multi-fusion attentive classifier [C]// Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D. C., USA: IEEE Press, 2024: 12702-12706.
[83] XIE Y K, CHENG H N, WANG Y T, et al. Learning a

self-supervised domain-invariant feature representation for generalized audio deepfake detection[C]//Proceedings of the INTERSPEECH 2023. Dublin, Ireland: ISCA, 2023: 2808-2812.
[84] LEI Z C, YANG Y G, LIU C H, et al. Siamese convolutional neural network using Gaussian probability feature for spoofing speech detection[C]//Proceedings of the INTERSPEECH 2020. Shanghai, China: ISCA, 2020: 1116-1120.
[85] HAMZA A, JAVED A R R, IQBAL F, et al. Deepfake audio detection via MFCC features using machine learning [J]. IEEE Access, 2022, 10: 134018-134028.
[86] LI X, LI N, WENG C, et al. Replay and synthetic speech detection with Res2Net architecture[C]//Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D.C., USA: IEEE Press, 2021: 6354-6358.
[87] LIU X H, LIU M, WANG L B, et al. Leveraging positional-related local-global dependency for synthetic speech detection [C]// Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D. C., USA: IEEE Press, 2023: 1-5.
[88] GE W Y, PATINO J, TODISCO M, et al. Raw differentiable architecture search for speech deepfake and spoofing detection[C]//Proceedings of the 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. [S. l.]: ISCA, 2021: 22-28.
[89] TAK H, JUNG J W, PATINO J, et al. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection[C]// Proceedings of the ASVspoof 2021 Workshop. [S. l.]: ISCA, 2021: 1-8.
[90] ZHOU Y P, LIM S N. Joint audio-visual deepfake detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Washington D. C., USA: IEEE Press, 2021: 14780-14789.
[91] ILYAS H, JAVED A, MALIK K M. AVFakeNet: a unified end-to-end Dense Swin Transformer deep learning model for audio-visual deepfakes detection[J]. Applied Soft Computing, 2023, 136: 110124.
[92] RAZA M A, MALIK K M. Multimodaltrace: deepfake detection using audiovisual representation learning [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Washington D.C., USA: IEEE Press, 2023: 993-1000.
[93] WANG R, YE D P, TANG L, et al. AVT2-DWF: improving deepfake detection with audio-visual fusion and dynamic weighting strategies [EB/OL]. [2024-11-14]. https://arxiv.org/abs/2403.14974.
[94] MITTAL T, BHATTACHARYA U, CHANDRA R, et al. Emotions don't lie: an audio-visual deepfake detection method using affective cues [C]//Proceedings of the 28th ACM International Conference on Multimedia. New York, USA: ACM Press, 2020: 2823-2832.
[95] CHUGH K, GUPTA P, DHALL A, et al. Not made for each other- audio-visual dissonance-based deepfake detection and localization [C]// Proceedings of the 28th ACM International Conference on Multimedia. New York, USA: ACM Press, 2020: 439-447.
[96] CHENG H, GUO Y Y, WANG T Y, et al. Voice-face homogeneity tells deepfake [J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2023, 20(3): 1-22.
[97] KATAMNENI V S, RATTANI A. MIS-AVoiDD: modality invariant and specific representation for audio-visual deepfake detection [C] // Proceedings of the International Conference on Machine Learning and Applications (ICMLA). Washington D. C., USA: IEEE Press, 2023: 1371-1378.
[98] LIU X L, YU Y, LI X L, et al. MCL: multimodal contrastive learning for deepfake detection [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(4): 2803-2813.
[99] LIU M, WANG J, QIAN X Y, et al. Audio-visual temporal forgery detection using embedding-level fusion and multi-dimensional contrastive loss [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(8): 6937-6948.
[100] ZOU H Q, SHEN M, HU Y C, et al. Cross-modality and within-modality regularization for audio-visual deepfake detection[C]//Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D.C., USA: IEEE Press, 2024: 4900-4904.
[101] OORLOFF T, KOPPISETTI S, BONETTINI N, et al. AVFF: audio-visual feature fusion for video deepfake detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2024: 27092-27102.
[102] YU Y, LIU X L, NI R R, et al. PVASS-MDD: predictive visual-audio alignment self-supervision for multimodal deepfake detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(8): 6926-6936.
[103] LI X L, LIU Z H, CHEN C, et al. Zero-shot fake video detection by audio-visual consistency[C]//Proceedings of the INTERSPEECH 2024. [S. l.]: ISCA, 2024: 2935-2939.
[104] FENG C, CHEN Z Y, OWENS A. Self-supervised video forensics by audio-visual anomaly detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2023: 10491-10503.
[105] LI X R, YU K, JI S L, et al. Fighting against deepfake: Patch&Pair Convolutional Neural Networks (PPCNN)[C]// Proceedings of the Companion Proceedings of the Web Conference 2020. New York, USA: ACM Press, 2020: 88-89.
[106] 戴磊, 曹林, 郭亚男, 等. 基于生成对抗网络的深度伪造跨模型防御方法[J]. 计算机工程, 2024, 50(10): 100-109.
DAI L, CAO L, GUO Y N, et al. Deepfake cross-model defense method based on generative adversarial network[J]. Computer Engineering, 2024, 50(10): 100-109. (in Chinese)
[107] HAQ I U, MALIK K M, MUHAMMAD K. Deep learning for deepfakes creation and detection: a survey [J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2024, 20(11): 341.

文字编辑 陆燕菲
栏目编辑 宋 圆