计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(10): 2523−2544, 2025

音频驱动的说话人面部视频生成与鉴别综述

乐 铮 胡永婷 徐 勇
(哈尔滨工业大学(深圳)计算机科学与技术学院 广东深圳 518055)
(深圳市视觉目标检测与判识重点实验室(哈尔滨工业大学(深圳)) 广东深圳 518055)
(23s151080@stu.hit.edu.cn)

Survey of Audio-Driven Talking Face Video Generation and Identification

Le Zheng, Hu Yongting, and Xu Yong
(*School of Computer Science and Technology*, *Harbin Institute of Technology*, *Shenzhen*, *Shenzhen*, *Guangdong* 518055)
(*Shenzhen Key Laboratory of Visual Object Detection and Recognition* (*Harbin Institute of Technology*, *Shenzhen*), *Shenzhen*, *Guangdong* 518055)

Abstract With the rapid advancement of artificial intelligence generation models and deepfakes, the techniques for generating talking face videos using various methods have become increasingly mature. Among them, audio-driven talking face video generation methods have attracted significant attention due to their remarkably realistic and natural output. Such methods utilize audio as a driving source to synthesize videos where the target character's mouth movements synchronize with the audio, often combining image or video materials. Currently, these technologies are widely applied in fields such as virtual anchors, gaming animation, and film and television production, demonstrating vast prospects for development. However, the potential negative impacts of this technology are also becoming apparent. Improper or abusive use could lead to serious political and economic consequences. In this context, research on identifying various types of facial forgery videos has emerged. Our research primarily assesses the authenticity of videos by detecting the veracity of individual video frames or the spatio-temporal consistency of video sequences. Firstly, we systematically analyze the classic algorithms and latest advancements in audio-driven talking face video generation tasks based on the timeline and the development history of foundational models. Secondly, we exhaustively list the commonly used datasets and evaluation criteria for this task, conducting comprehensive comparisons across multiple dimensions. Subsequently, we meticulously analyze and summarize the forgery facial video identification task, categorizing it based on whether the discrimination technology focuses on individual video frames or multiple frames, and also summarize its commonly used datasets and evaluation criteria. Finally, we outline the challenges and future directions in this research field, aiming to provide valuable references and support for subsequent related research.

Key words deep learning; deepfake; audio-driven; talking face video generation; forgery facial video identification

摘 要 随着人工智能生成模型和深度伪造的迅速兴起,利用各种方法生成人脸说话视频的技术日益成熟,其中音频驱动的说话人面部视频生成方法因其生成效果的逼真自然而备受瞩目.该类方法利用音频作为驱动源,结合图片或视频素材,用以合成与输入音频口型同步的目标角色讲话视频.目前,相应的技术已经被广泛应用于虚拟主播、游戏动漫、影视剧制作等内容创作领域,并展现出广阔的发展前景.然而,这些技术的潜在负面影响也日益显现,若被不当利用或滥用,极有可能触发严重的政治和经济后果.在此

收稿日期: 2024-03-21; **修回日期**: 2025-01-07
基金项目: 国家自然科学基金项目(62371157)
This work was supported by the National Natural Science Foundation of China (62371157).
通信作者: 徐勇(laterfall@hit.edu.cn)

背景下,针对面部伪造视频的鉴别研究应运而生,主要通过分析单视频帧的真实性或视频帧序列的时空一致性来评估视频的真实性.首先,依据时间脉络及基础模型的发展轨迹,系统性地剖析了音频驱动面部视频生成任务的经典算法与最新研究成果.其次,详尽列举了该任务领域内常用的数据集及评估标准,并从多个维度出发,对这些数据集与标准进行了全面深入的对比分析.紧接着,针对伪造面部视频鉴别任务,对鉴别技术所针对的对象(即单帧或多帧)进行了细致的分类与归纳,同时,也对其常用的数据集及评估标准进行了系统的总结与梳理.最后,展望了该研究领域面临的挑战与未来的发展方向,旨在为后续的相关研究提供有价值的参考与坚实的支撑.

关键词 深度学习;深度伪造;音频驱动;说话人面部视频生成;伪造脸部视频鉴别

中图法分类号 TP391.41

DOI: 10.7544/issn1000-1239.202440207 **CSTR:** 32373.14.issn1000-1239.202440207

近年来,随着人工智能的快速进步,深度伪造技术也不断发展,涌现出多种可以通过篡改或替换视频中的人脸以制作伪造视频的技术.然而,在日常交流中,人物的面部表情和语音高度同步,即使是微小的不匹配也很容易被察觉[1],这使得人脸视频的伪造成为一项非常具有挑战性的任务.

相比于以文字或图片作为伪造视频的驱动源,音频驱动方法能更准确地捕捉说话者的嘴唇动作和表达方式,从而生成更自然、更逼真的面部表情和嘴部运动,使得合成的面部视频更加接近实际说话情境,因此音频驱动面部视频生成技术引起了广泛关注.该技术以音频作为驱动源,作用于目标人物的面部图像或视频,旨在生成流畅真实的说话人面部视频,在虚拟主播、游戏动漫、影视剧制作等领域展现出了诱人的应用前景.例如HeyGen成功将美国富豪马斯克的一段全英语采访转换成全中文采访,并且做到声音、声调、口型的高度对应以至以假乱真[2].

音频驱动的说话人面部视频生成方法,紧跟计算机视觉领域主流方法的步伐,其发展历程见证了多种方法论的迭代与融合.起始于矢量量化等早期方法,随后过渡至神经网络及隐马尔可夫模型的应用,这些早期探索为后续的深度学习浪潮奠定了基础.深度学习模型的兴起,尤其是卷积神经网络(convolutional neural network, CNN)和长短期记忆网络(long short-term memory, LSTM),显著提升了视频生成任务的处理能力,不仅在视频分析上表现出色,还深化了音频与视频信号之间复杂关联的理解.生成式对抗网络(generative adversarial network, GAN)与变分自编码器作为生成模型的代表,通过模拟和优化真实数据分布,为说话人面部视频的生成提供了新的视角和可能性.近年来,随着大规模音频-视频数据集的涌现,研究者得以训练更深层、更精细的模型结构,其中Transformer架构的自注意力机制尤为突出,它有效捕捉了序列间依赖性,增强了模型对时空信息的敏感度.此外,神经辐射场的引入开辟了生成3D动态场景的新纪元,其高保真的重建能力为说话人面部的立体表现力带来了质的飞跃.而扩散模型作为一种新兴的生成方法,因其生成的视频具有高清晰度和高连贯性,正在成为该领域研究的热点,预示着未来音频驱动的说话人面部视频生成方法将更加成熟且多样化.

在人脸视频生成技术为生活带来便利的同时,不法分子也可能利用该技术传播虚假信息、编造虚假证据、进行电信诈骗、煽动舆论等不法行为[3],尤其是重要政治人物的伪造视频极易产生巨大的社会影响甚至引发国家动荡.这些虚假视频的内容往往音频与嘴唇高度同步,肉眼几乎不能直接辨别,而又由于人脸和音频的隐私性较差,可以轻易被获得、利用和伪造[4],使这项技术极易被滥用.

为应对伪造视频泛滥所引发的负面效应,遏制失实信息的蔓延,人物面部伪造视频的检测方法应运而生,迅速演进并累积了丰富的研究成果.从对伪造视频评估的角度,可将其鉴别策略归纳为两大核心类别:一类聚焦于单帧图像的剖析,另一类则着重于帧间关联性的考察.伪造视频的检测手段经历了从依赖人工设计的特征向自动提取深度特征转变的不同阶段.早期的手工特征分析,虽在特定情境下展现了一定的检测能力,但其局限性逐渐凸显.随着深度学习的崛起,利用神经网络自动挖掘的深度特征展现出了更强的表征能力和泛化性能,这标志着伪造视频检测方法迈入了一个新纪元.不同的鉴别方法通过对视频中的面部特征、表情变化、语音与口型匹配等多个维度进行分析和比较来判断视频的真实性.在实际应用中,面部视频鉴别技术已经应用于

多个领域，如司法鉴定、新闻传播等.

通过对现有文献的回顾分析，本文全面总结了音频驱动说话人面部视频生成与鉴别任务的关键算法及其发展脉络，揭示了该领域从初步探索到当前创新实践的演变过程. 本文不仅分类整理了最具影响力和创新性的研究成果，还深入剖析了这些算法的原理和应用场景. 在数据集和评价指标方面，本文进行了全面梳理，明确了当前研究中的常用资源及其评估标准，为后续研究提供了参考框架. 最后，本文指出了该领域尚存的挑战与未来可能的发展方向，为研究者提供新的思路与启示.

1 音频驱动的说话人面部视频生成方法概述

在早期阶段，音频驱动的说话面部生成视频的合成痕迹较为明显，有些唇部变形夸张，有些唇部动作过于微弱，都无法达到与输入音频完全匹配的效果，很难输出一段自然流畅的人物说话视频. 此外，这些方法生成的面部往往十分模糊，分辨率较低.

然而，随着深度学习的迅速发展和广泛应用，生成的说话面部视频质量已逐渐提高. 深度学习模型能够学习并利用大量真实人脸说话视频，不仅克服了早期方法的局限，还能够在保持语音与口型一致性的同时呈现出更加自然细腻的面部细节.

图 1 展示了本文总结的部分代表性音频驱动说话人面部视频生成方法的分类.

1.1 早期方法

人物面部生成研究的演进历程，彰显了人工智能领域从传统统计模型向深度学习架构过渡的创新轨迹. 早期研究主要围绕基于矢量量化（vector quantization, VQ）、神经网络（neural network, NN）和隐马尔可夫模型（hidden Markov model, HMM）等方法展开.

基于 VQ 的方法[5]以其直观的编码机制，将语音数据简化为码字，进而映射至视觉输出，但这种方法的简朴性也带来了量化误差的不确定性及视觉输出的不连贯性. 相比之下，基于 NN 的方法[6]能生成更为平滑的视觉估计，有效规避量化误差，却同样忽视了音频信号中蕴含的上下文信息，限制了其表现力.

鉴于此，HMM 的引入为音频到视觉转换的建模提供了有力支撑，通过捕捉时间序列的内在关联性，显著提升了输出的连贯性和准确性. 此类方法中比较有代表性的研究工作如下：Yamamoto 等人[7]通过将 HMM 与上下文相关的唇部参数结合，减少了时间平均误差，体现了对序列数据的深化理解；Lee 等人[8]进一步通过分别对面部图像和音素序列进行 HMM 建模展现了多模态融合的潜力；Aleksic 等人[9]开发了一个相关性隐马尔可夫模型（coupled hidden Markov model, CHMM）系统，集成了独立训练出的声学隐马尔可夫模型（acoustic hidden Markov model, AHMM）和视觉隐马尔可夫模型（visual hidden Markov model, VHMM）系统，结合延迟集成策略，有效地减少了训练数据需求和时间对齐误差，标志着多模型协同优化的重要进步.

鉴于 HMM 对于时间序列数据关联性的建模能力以及深度神经网络（deep neural network, DNN）在非线性关系拟合方面的显著优势，研究者们提出了结合 HMM 与 DNN 的音频驱动的说话人视频生成方法.

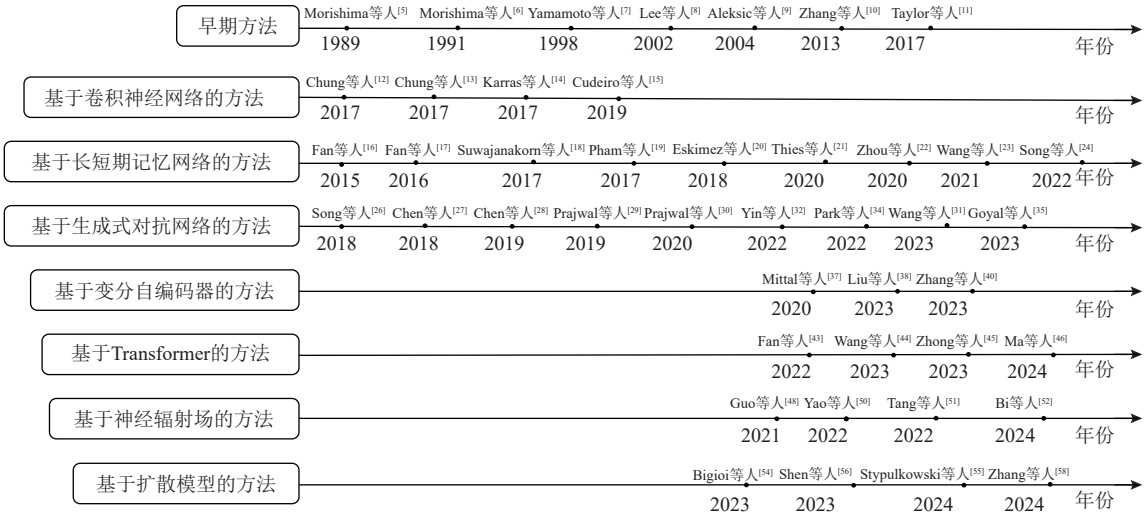


Fig. 1 Classification of audio-driven talking face video generation methods

图 1 音频驱动的说话面部视频生成方法分类

Zhang 等人[10]受上下文相关的深度神经网络隐马尔可夫模型(context-dependent deep neural network hidden Markov model, CD-DNN-HMM)启发, 提出将声学语音输入映射到 9 000 个绑定状态之中, 即对于给定的语音输入, 由 DNN 预测可能的状态, 并通过预测出的状态网络和 HMM 唇部运动合成器渲染出唇部动画, 其结构如图 2 所示. 后续的研究则逐渐转向直接运用 DNN 进行人脸视频的合成, 例如 Taylor 等人[11]使用连续深度学习的滑动窗口方法使预测器能够建立音素描述与连续语音视频表示间的复杂非线性关系, 更直接地聚焦于捕捉局部上下文和协同效应.

这些早期方法为说话人面部视频生成任务奠定了理论与实践基础, 但视觉表现准确性和逼真度较低, 与真实视频效果存在明显差距.

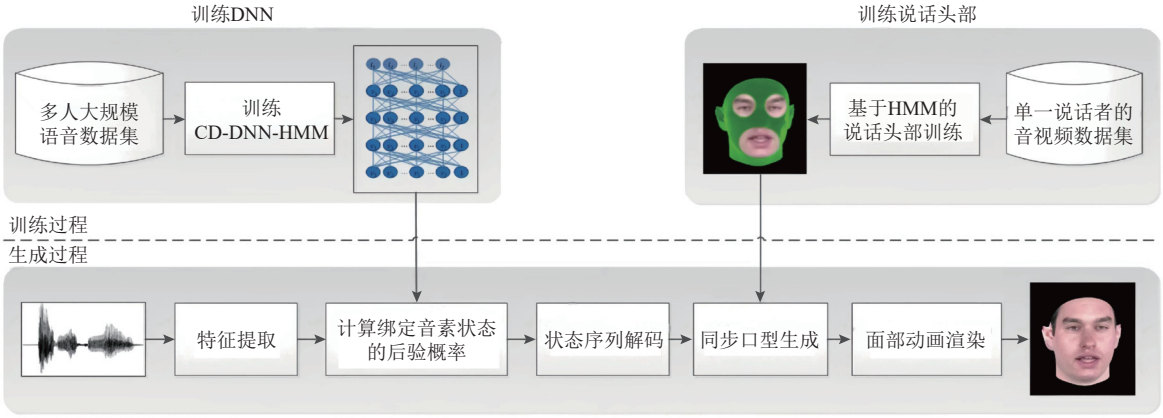


Fig. 2 Talking face generation method based on CD-DNN-HMM[10]

图 2 基于 CD-DNN-HMM 的说话面部生成方法[10]

1.2 基于卷积神经网络的方法

随着深度学习架构的进一步发展, CNN 在图像与视频这种网格结构数据的表征方面表现出了强大的能力, 其局部的信息抽取与归纳的特性不仅具有相对高的计算效率, 而且与人类视觉感知具有一定的一致性, 逐渐成为计算机视觉领域的主流模型. 这一趋势也延展到了音频驱动的说话人面部视频生成领域, 并推动了一系列创新方法的诞生与发展.

Chung 等人[12]提出了第 1 个端到端的音频-视频同步系统, 使用 ConvNet 架构开发出一种跨语言、跨说话者的通用唇同步解决方案. 随后, 他们推出的 Speech2Vid[13]借助无监督学习策略实现了仅凭音频片段和目标个体的静态图像即可合成动态的说话人面部视频. 这一方法避免了传统的音素到视频帧的直接映射, 转而专注于学习原始音频与视频数据间隐含的关联, 为后续的研究提供了新的思路, 其结构如图 3 所示.

Karras 等人[14]提出了一种创新的端到端框架, 能够从输入的音频波形直接学习到面部模型的 3D 顶点坐标, 实现了音频驱动 3D 面部视频的生成, 被视作使用 CNN 实现 3D 面部视频生成的代表性工作. Cudeiro 等人[15]提出的 VOCA 在一个专门的面部数据集上进行训练, 能够接受任何语音信号作为输入并输出多样化的真实说话风格, 是可以直接适用于未知主体而无需额外的重定向步骤的逼真 3D 面部动画模型.

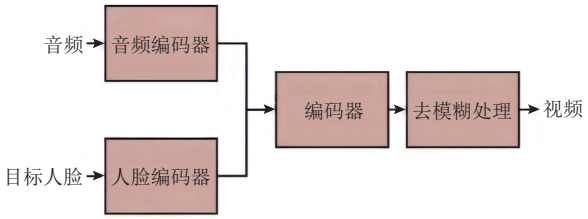


Fig. 3 Structure of Speech2Vid[13]

图 3 Speech2Vid 结构[13]

尽管 CNN 在音频驱动的说话人面部视频生成任务中取得了显著成果, 但仍存在局限性, 尤其是在处理长期依赖关系、时间动态变化以及提升音视频数据的敏感度和鲁棒性方面. 这些挑战为 LSTM 网络、注意力机制等技术的整合与优化提供了方向.

1.3 基于 LSTM 的方法

LSTM 网络是一种循环神经网络(recurrent neural network, RNN)的变体, 凭借门控机制(遗忘门、输入门和输出门)实现了对信息流的有效调控, 在处理音频信号与面部视频间存在的长时间依赖关系上展现出显著优势. 如连续的语音与对应的面部变化, LSTM 可有效缓解传统 RNN 面临的梯度消失与爆炸难题,

确保了模型在训练过程中的稳定收敛.

Fan 等人[16]通过修改与优化 LSTM 的网络结构显著提升了其性能表现，提出利用深度双向长短期记忆网络（deep bidirectional long short-term memory network, BLSTM）和长短期记忆循环神经网络（long short-term memory recurrent neural network, LSTM-RNN）[17]来强化对长期上下文的捕捉能力，实现了音频与视觉流之间更为精细的非线性映射. 在利用 LSTM 生成说话人脸视频的早期研究中，Suwajanakorn 等人[18]的研究非常具有代表性，他们使用大约 16 小时的美国总统奥巴马的演讲视频训练模型，当输入音频时，可以输出口型、动作均匹配的奥巴马说话视频，生动展现了 LSTM 在实时视频生成上的潜力，其合成过程如图 4 所示. Pham 等人[19]也采用 LSTM-RNN 的方法，首先输入音频中提取的一组声学特征，然后在视频生成环节将输出的面部运动用 3D 旋转和混合表情模型的表情权重来描述.

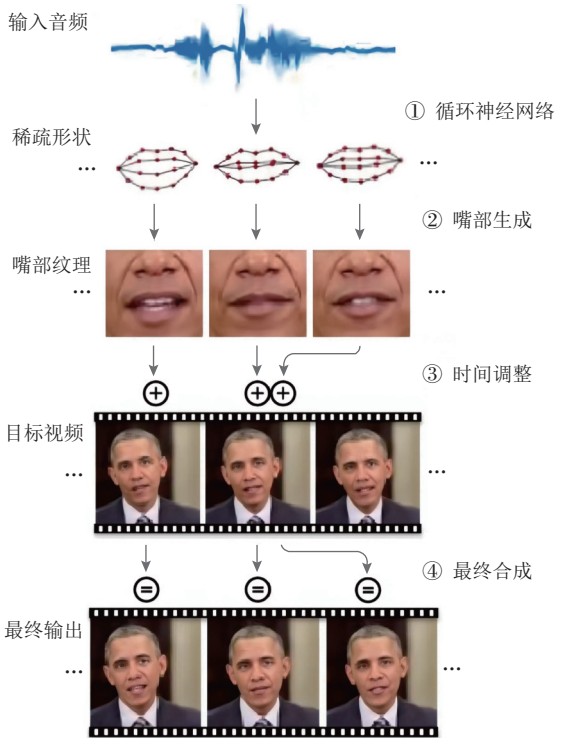


Fig. 4 Process of “synthesizing Obama”[18]

图 4 “合成奥巴马”过程[18]

随后 LSTM 被应用于各有侧重的说话人面部视频生成模型中. 在聚焦于视频帧中面部关键点的预测与转换的工作中，Eskimez 等人[20]首先对视频帧进行面部关键点的转换，然后通过将关键点转换为整个训练数据集中的平均脸形来消除身份信息，再采用 LSTM 网络进行训练，预测每帧的脸部关键点，实现了比使用形状和旋转参数更好的效果.

从个性化建模的角度出发，Thies 等人[21]将整个模型分为 2 个模块：一个是通用模块，该模块负责通过音频预测一个潜在面部向量，也就是 3D 面部模型的混合形状系数；另一个是专用模块，该模块需要目标人物的 2~3 min 长视频以捕捉其面部特质建立面部模型，最后通过一种新颖的神经渲染网络存储和重新渲染个体面部的外观.

从肖像动画生成的角度出发，Zhou 等人[22]首先利用一个语音内容动画模块，将音频映射到动画人物面部关键点的唇部、下颌及附近的面部区域，之后再通过说话者感知动画模块，将修改后的动画人物面部关键点通过帧对帧的转换泛化到真实面部视频的形成中去.

在以密集运动场为媒介的工作中，Wang 等人[23]的模型首先使用一个基于动作感知的 LSTM 的头部姿势预测器预测人物的说话状态，然后设计了一个运动场生成器，输入头部姿势、音频和参考图像生成密集运动场，用运动场和关键点表示头部和面部的运动，而运动场是由预测的关键点与参考图像的关键点之间的差异描述的. 这项工作综合了面部区域、头部和背景的运动，使得在生成视频时可以更好地控制空间和时间的一致性.

从 3D 面部重建的角度出发，Song 等人[24]首先用单目 3D 面部重建将每个目标视频帧分解为正交参数空间，包括表情、几何和姿势 3 个参数，再使用通过变换表情参数使得生成面部表情与音频更匹配和谐，而几何和姿势这 2 参数保持不变以保留原始视频素材的上下文，最后使用一种新型的神经视频渲染网络和动态规划方法以构建一个连贯真实的视频.

综上所述，LSTM 及其衍生结构的应用不仅丰富了说话人脸视频生成任务的创作工具箱，还促进了对自然语言处理与计算机视觉交叉领域的深入理解.

1.4 基于 GAN 的方法

GAN[25]由 2 个神经网络组成，分别为生成器和鉴别器，旨在利用对抗博弈的思想，通过训练生成器来产生逼真的面部视频，同时通过不断迭代优化生成器和鉴别器之间的对抗性机制来提高生成视频的逼真程度和多样性，使其更符合实际的说话场景需求，其基本结构如图 5 所示.

早期研究者在 RNN 的基础上融入了对抗学习的思想，例如 Song 等人[26]设计了一种条件循环对抗网络，在循环单元中同时整合了图像和音频，确保了视频生成过程中面部与唇部动作的连贯性与时间一致

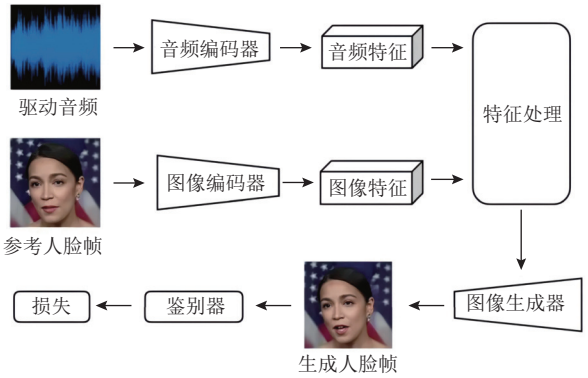


Fig. 5 Method based on GAN

图 5 基于生成式对抗网络的方法

性, 与此同时也采用对抗学习的思想, 使用唇读判别器以对抗方式训练网络, 并进一步部署空间-时间判别器优化了视频的真实感和细节表现.

Chen 等人[27]在使用 GAN 进行音频驱动的说话人脸视频生成方面做出了开创性贡献. 他们综合了音频-视觉相关性损失、特征空间损失、重构损失以及对抗学习损失, 构建了一个全面的损失函数框架, 有效提升了生成视频的质量. 此外, 他们设计了一种级联 GAN 方法[28], 通过音频变换网络(audio transformation network, AT)和视频生成(video generation, VG)网络的级联, 实现了音频特征指导下的面部关键点调整及高质量视频帧的生成.

Prajwal 等人提出的 LipGAN[29]和随后的 Wav2Lip[30]是使用 GAN 进行音频驱动的说话人脸视频生成的经典之作. LipGAN 利用 GAN 精确衡量唇部同步度, 适用于跨语言的面部视频生成, 即使面对复杂姿态也表现出色. Wav2Lip 进一步深化了这一方法, 其将面部视频作为视觉输入, 并以遮挡了下半部分的输入视频为先验知识, 除使用重建损失和 GAN 中原有的训练鉴别器之外另外设计了 2 个鉴别器: 其一为预先训练好的唇同步鉴别器, 用来准确地检测唇同步错误以增强音频和嘴唇的同步性; 其二为视觉质量鉴别器, 用来提高视觉质量和同步精度, 其结构如图 6 所示.

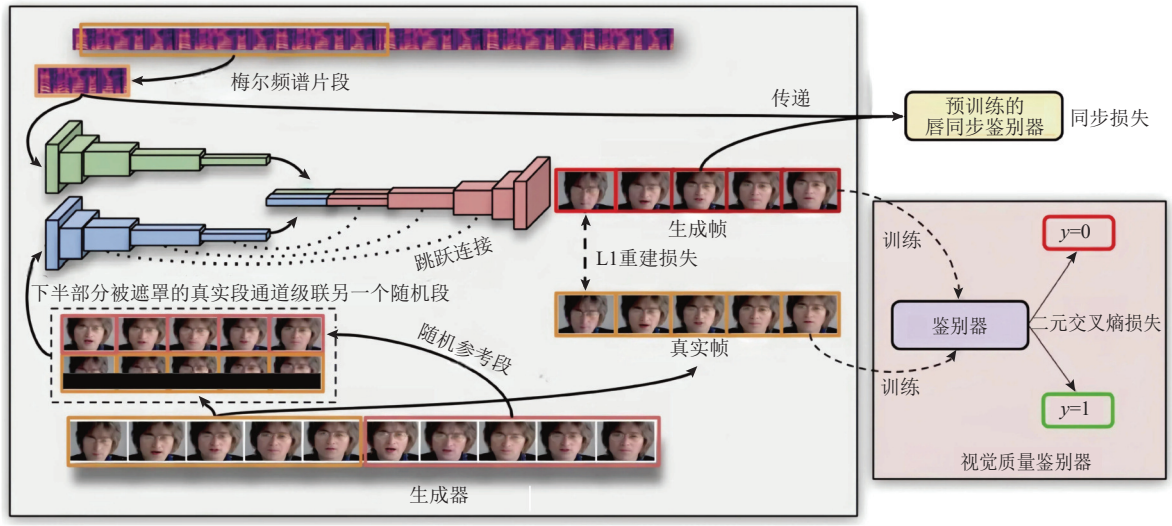


Fig. 6 Structure of Wav2Lip[30]

图 6 Wav2Lip 结构[30]

在利用 GAN 解决特定问题时, 研究者能够巧妙地定制鉴别器, 设计出专属的损失函数, 以此来满足特定应用场景的需求. 以 Wang 等人[31]的工作为例, 他们针对可被准确唇读的视频生成这一具体问题, 关注生成视频的视觉可理解性, 不同于 Wav2Lip 的专家模型更多地关注于一般性的面部运动与音频的匹配, 他们提出添加一个预训练的唇读专家模型使得形成的唇形能够传达正确的语音信息, 由此引入新的唇读损失, 通过惩罚不正确的唇读结果来提高唇部区域与音频的同步性和可读性.

在涉及利用中间媒介的视频生成领域, GAN 展现出其独特的能力, 例如 Yin 等人[32]首先通过 GAN 反演的编码器获得源图像的特征图, 再将音频和源图像输入运动生成器生成密集运动场用于扭曲特征图, 然后通过一个校准网络纠正失真的特征, 最后将特征图输入 StyleGAN[33]中生成面部视频.

基于 LipGAN 和 Wav2Lip 的成功, 后续研究在不同方面进行了拓展. 在个性化表达方面, Park 等人[34]为准确地从输入音频中过滤出与唇部运动相关的特征, 对编码后的音频特征和唇部特征进行对齐并存

储在音频唇部记忆中，查询时则将从记忆中检索到的唇部特征与音频特征融合，最后注入到解码器网络中进行视频合成. 在情绪表达方面，Goyal 等人[35]使得模型可以生成特定表情的面部视频，包括快乐、悲伤、恐惧、愤怒、厌恶和中性，他们将这 6 种情感的表情标签编码作为新的特征向量与面部特征、音频特征相融合，从而生成与输入情感匹配的唇语同步视频.

GAN 在生成逼真面部视频的应用中展现出了巨大的潜力，但其在训练稳定性和视频质量方面也有局限性，即训练过程不稳定且容易崩溃，导致生成器无法生成覆盖整个数据分布的样本，而是仅学会生成少数几个独特样本；生成的视频质量也不够，难以得到更高分辨率的视频.

1.5 基于变分自编码器的方法

变分自编码器(variational autoencoder, VAE)[36]也是一种生成模型，它通过学习一个概率模型来表示数据的生成过程，从而生成类似训练数据的新样本，其基本结构包括编码器和解码器 2 部分. 在音频驱动的说话人面部视频生成任务中，编码器可以将音频特征和面部动画数据映射到低维隐空间，而解码器则根据隐空间的表示生成对应的面部动画序列. 音频输入后，VAE 可以在隐空间中进行查询，从而生成对应的说话视频. 由于 VAE 的隐空间通常有良好的解耦性，因此可以方便地对面部表情、头部姿态等细节进行精确控制，另外其训练过程也通常比 GAN 更稳定可靠.

在音频驱动的人脸视频生成领域，VAE 得到了广泛的应用. Mittal 等人[37]开创性地利用 VAE 框架成功地将音频中的语音内容、情感色彩等要素解耦成不同的表示，用 KL 散度、负对数似然以及边际排名损失确保所学习的表示确实是解耦且有意义的，再将其与输入的面部图像一起输入基于 GAN 的视频生成器以实现面部视频的生成，其用于解耦音频的 VAE 架构如图 7 所示.

此后的研究者持续深化对 VAE 及其衍生模型的应用，聚焦于更为细分的任务领域，例如 Liu 等人[38]设计了基于概率条件的变分自编码器[39]模型，基于音频驱动和原面部生成自然的头部姿态，再据此预测无监督关键点以建模面部结构位置，最后使用流导向的考虑遮挡的生成器产生面部流来指示面部结构；Zhang 等人[40]则设计了 PoseVAE 用来建模给定姿态的残差从而学习真实且具有身份感知的风格化头部运动，结合将 3DMM(3D morphable model)的运动系数视为中间表示的隐式 3D 系数调制系统生成逼真的说话人脸视频.

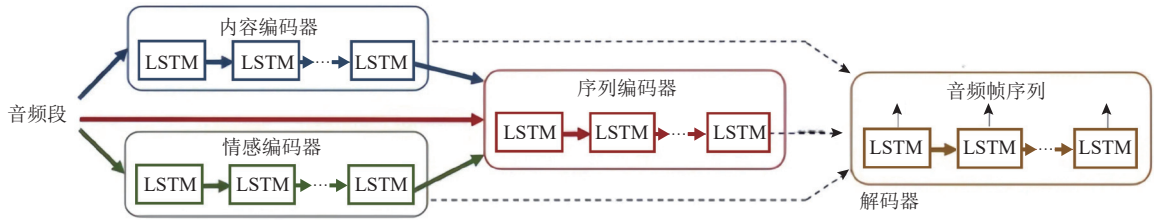


Fig. 7 VAE decouples audio[37]

图 7 VAE 解耦音频[37]

由于 VAE 在优化重构损失时可能会过度平滑数据分布，从而导致生成的样本缺乏足够的细节并影响生成视频的逼真度和生动性，许多方法均在 VAE 的基础上结合其他生成模型来改进和优化生成过程.

1.6 基于 Transformer 的方法

Transformer[41]是一种基于自注意力机制(self-attention)的深度学习架构，最初在自然语言处理领域用于处理序列到序列的任务. 音频驱动的说话人面部视频生成不仅可被看作一个序列到序列的任务，而且在该任务中，Transformer 的编码器和解码器中包含的多层自注意力模块使之可以有效处理序列中的长距离依赖关系，从而更好地分析音频的时间结构，生成精准同步的面部画面.

但传统的 Transformer 架构对数据需求量大，需要足够的数据集进行训练[42]，因此直接应用在人脸视频生成任务中时往往效果不佳，其次传统的编码器-解码器注意力机制无法处理模态对齐. 为了解决这些问题，Fan 等人[43]提出的 FaceFormer 使用自监督、预训练的语音模型增强在数据稀缺的情况下的视频生成效果，并设计了偏置跨模态多头注意力以对齐不同的模态，从而能实现高度逼真且稳定的面部 3D 面部视频生成，其结构如图 8 所示.

为了解决现有的数据集分辨率、质量较低以及身份数量有限的问题，Wang 等人[44]通过学习重建 2D 面部得到预学习面部编码，并将其作为强大的先验，从视觉保真度和可泛化性的角度促进说话人头像

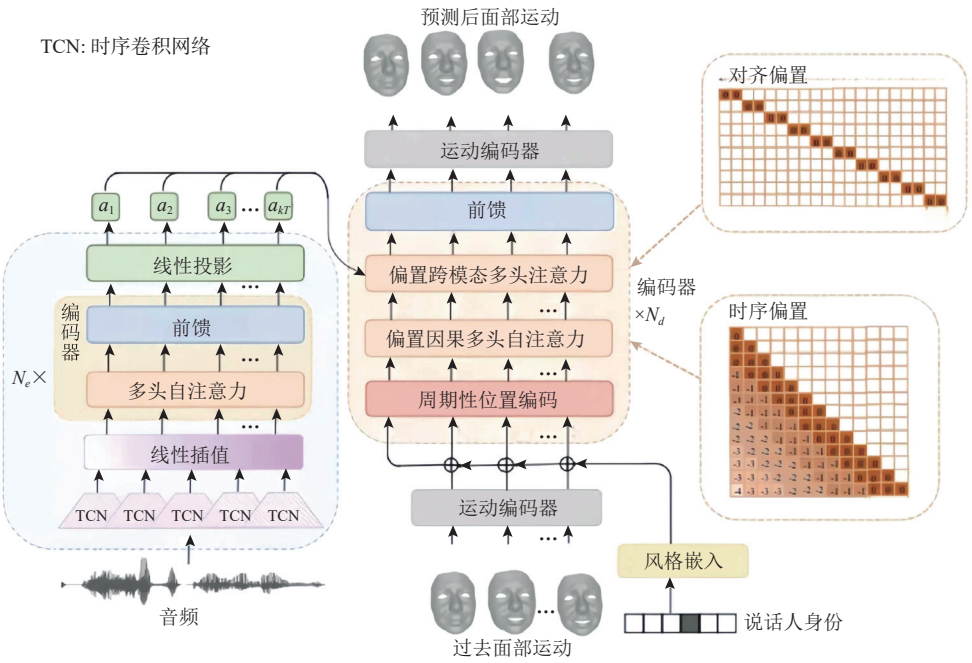


Fig. 8 Structure of FaceFormer[43]

图 8 FaceFormer 结构[43]

的合成，并由此将说话人头像生成任务简化为寻找适当的唇部编码以表征说话过程中嘴唇的变化，利用基于 Transformer 架构的 Lipformer 建模音频-视觉的一致性，凭借输入的音频特征预测唇部编码序列.

众多方法采纳 Transformer 架构作为其核心组件，以实现对音视频信号的深度特征抽取与分析，随后将这些提取到的特征与后续模型无缝对接，协同作用以达成更为卓越的综合性能. Zhong 等人[45] 关注唇部和下颌这 2 个说话面部生成的关键区域，将工作也分为 2 个模块：第 1 个部分为音频到面部关键点的生成，设计了一个基于 Transformer 的面部关键点生成器，将音频、完整面部关键点和除唇部和下颌外的关键点输入其中，得到预测的唇部关键点、下颌关键点与先验关键点相结合，从而构建完整目标草图；第 2 个部分则输入目标草图以及遮蔽下半部分的人物面部，以已知参考图像和关键点为辅助，生成最终的面部影像. Ma 等人[46] 提出的 CVTHead 结合了基于点的神经渲染和 Transformer 架构，使用预训练的 3D 面部重建网络的顶点对渲染出的面部通过形状、表情和头部姿势系数进行显式控制. 在这个过程中使用 Transfomer 学习顶点之间的非局部关系和对应关系，通过引入全局注意机制，模型可以捕获所有顶点特征中的长程依赖关系，得到更加优越的性能.

1.7 基于神经辐射场的方法

神经辐射场(neural radiance fields, NeRF)[47] 凭借其强大的 3D 表达能力成为 3D 视觉领域的主流方法，也逐渐被研究者们应用到说话人面部生成的任务中. Nerf 的基本思想是使用深度神经网络来表示场景中光线的颜色和密度，通过优化网络的权重，可以学习到如何从任何视点渲染场景，从而实现高质量的图像合成和视点变换. NeRF 能以 2D 的图片为监督，表示多视角的复杂 3D 场景，其生成的图像细节精细，可以渲染出不同视角下的场景，这对于生成连贯真实的视频帧至关重要，其基本渲染过程如图 9 所示.

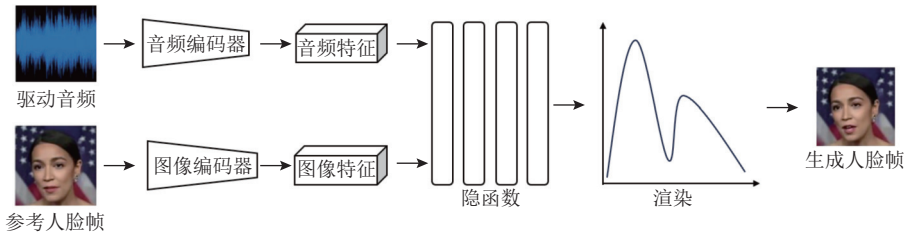


Fig. 9 Method based on NeRF

图 9 基于 NeRF 的方法

Guo 等人[48]提出了 AD-NeRF 模型，这是一个端到端的基于音频驱动的 NeRF 架构，其直接学习一个有条件的隐式函数将相应的音频特征映射为动态神经辐射场[49]. 考虑到头部的运动与上半身的运动不一致，他们将神经辐射场表示分为 2 个部分，一个用于前景面部，另一个用于前景躯干，分别建模使得生成视频更为流畅自然，其结构如图 10 所示. 不同于 Guo 等人的工作，Yao 等人[50]不直接将音频输入到神经辐射场，而是首先将面部动作视为嘴唇运动和个性化属性 2 种表示，使用一种对比学习策略预测嘴唇运动，提出一种概率模型预测眨眼、头部姿态等随机属性，随后在这些特征上优化动态神经辐射场的条件，确保视频口型同步且运动自然，为渲染提供更为精准的指导.

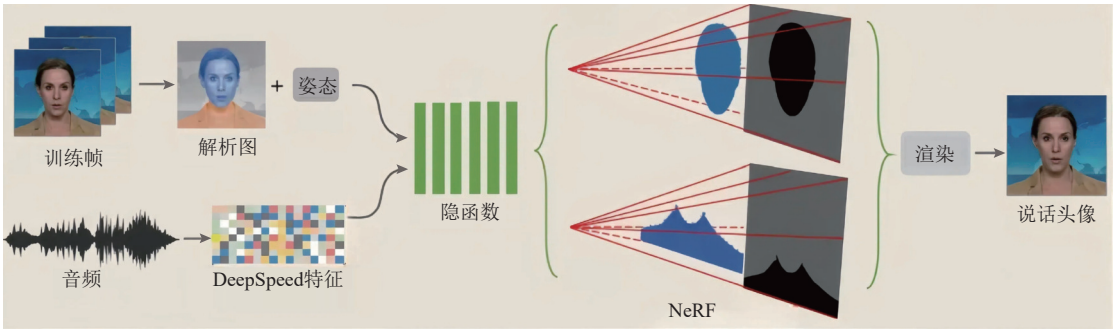


Fig. 10 Structure of AD-NeRF[30]

图 10 AD-NeRF 结构[30]

尽管 NeRF 展现出了诸多优势，但其计算密集型的特性导致渲染速度缓慢. 为了缓解这一问题，Tang 等人[51]提出了一种更为高效的框架，实现了实时合成说话肖像和更快的收敛速度. 他们将高维说话面部表示分解为 3 个网格：3D 空间、2D 音频和 2D 视频. 为了大幅减少了多层感知机在所有采样点的计算负担，显著提升了处理效率，他们利用编码模块处理前 2 个网格以建模动态头部，利用伪 3D 可变性模块处理后一个网格以生成躯干.

针对 NeRF 在唇形生成上的局限，Bi 等人[52]提出了一种基于注意力解耦的条件化 NeRF 方法，利用与语音相关的面部动作单元(action units, AU)将面部特征分离为音频-面部与身份-面部，前者聚焦于语音运动关联区域，后者关注说话者个体特征. 为精确匹配嘴唇运动，仅音频-面部部分与音频特征结合，进行精细化渲染，这种方法通过任务分解和差异化监督，使 NeRF 更加专注于关键学习目标，从而显著改善了整体生成质量.

1.8 基于扩散模型的方法

近年来，扩散模型(diffusion model, DM)[53]被广泛地应用于各类图像生成任务中，取得了很好的效果，该模型通过前向扩散过程将数据逐渐添加噪声直到变成高斯白噪声，然后通过反向扩散过程逐步去除噪声，最终恢复出清晰的数据样本，可以提供稳定的生成过程和高保真度的视频，其基本架构如图 11 所示.

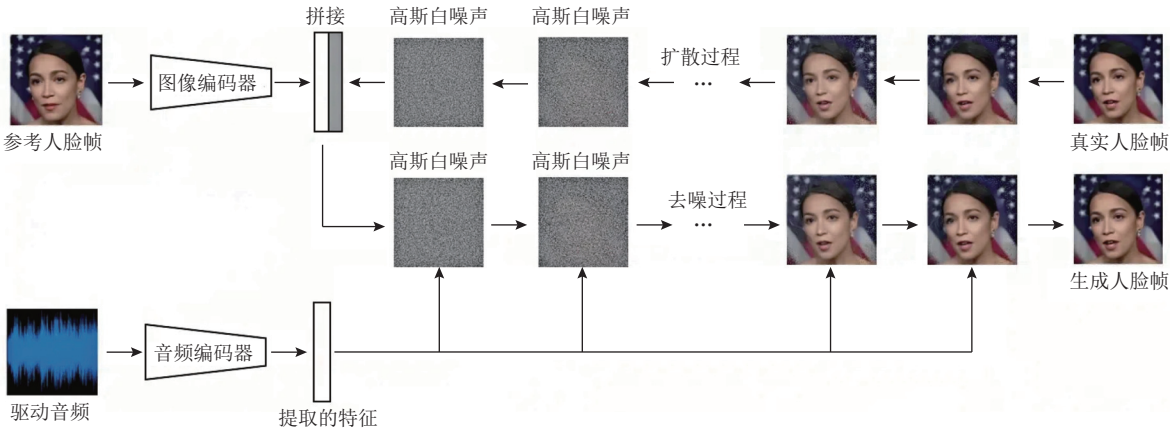


Fig. 11 Method based on DM

图 11 基于扩散模型的方法

受到扩散模型启发, Bigioi 等人[54]提出了一种新颖的非结构化端到端方法, 通过在帧与帧之间基于音频频谱特征进行条件化, 可以实现在去噪扩散模型上生成同步的面部运动. Stypułkowski 等人[55]提出的 Diffused Heads 是一种基于帧的扩散模型, 使用一个固定的身份帧和不同的音频段来不断生成新的视频帧, 并通过运动帧和音频嵌入注入了有关过去动作和未来表情的额外信息, 其结构如图 12 所示. 这 2 项工作引领了将扩散模型引入音频驱动的视频生成任务的潮流.

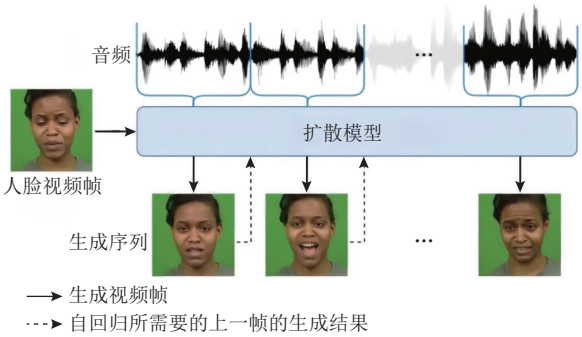


Fig. 12 Structure of Diffused Heads[54]

图 12 Diffused Heads 结构[54]

虽然扩散模型能够生成清晰度较高的视频, 但由于其对图像进行像素级的处理, 若用来生成视频则需要大量的计算资源和时间. 针对这个问题, Shen 等人[56]设计的 DiffTalk 应用了新兴的潜在扩散模型(latent diffusion model, LDM)[57]将 DM 的训练和推断过程转移到一个压缩的低维潜在空间, 从而成功地将说话人面部生成建模为一个音频驱动的时间连续去噪过程, 最后生成高分辨率的视频.

在 DiffTalk 的基础上, Zhang 等人[58]设计了一种可以编辑人脸情感的方法 EmoTalker. 该方法可以通过利用包含复杂情绪和强度的文本提示, 引导去噪过程生成期望的面部表情. 为了克服在处理具有挑战性的人脸姿态时泛化能力的局限性, EmoTalker 在推断阶段精进了其去噪算法. 这一改进确保了所生成的每一帧都能够更精准地与人物面部特征对齐, 在复杂的人脸角度下, 系统也能保持相对稳定的表现.

2 音频驱动的说话面部视频生成数据集和评估指标

数据集对人工智能的发展起到了至关重要的作用, 对于音频驱动说话面部视频生成方法来说也是如此. 在早期, 由于技术和数据限制, 音频驱动说话面部生成数据集相对较小且质量不高. 这些数据集通常包含少量面部视频, 缺乏多样性和真实性, 导致生成结果的逼真度和多样性受到限制. 随着越来越多研究者投入到这项工作之中, 说话面部生成数据集逐渐变得丰富和多样化, 涵盖了不同年龄、种族、性别和表情等多个方面, 也包含了拍摄视角、人物表情等元素. 通过大量真实的数据集, 模型可以学习不同人脸的特征、表情和姿势, 从而更好地生成逼真的面部视频.

音频驱动说话面部视频生成的评价指标对于评估生成结果质量、指导模型优化、比较不同模型以及制定研究方向都至关重要. 这些指标帮助衡量生成的视频序列与真实视频之间的相似度和质量, 发现模型存在的问题和不足, 并提供客观的比较和评估依据, 同时也反映了该领域的研究热点和难点, 推动着这一领域的不断发展进步.

本节将介绍广泛应用于模型训练和评估的常用数据集以及常用的评价指标.

2.1 音频驱动的说话面部视频生成数据集

2.1.1 从实验室收集的数据集

实验室数据集是指在实验室或研究机构内, 为了研究或测试说话人面部视频生成问题, 特别雇佣演员或志愿者进行录制而得到的数据集. 由于构造数据集时能够人为可控地模拟出实验所需的各种情境, 视频的质量普遍较高, 视频人物可以有各类情绪、录制角度等不同方面的呈现, 但是录制成本较为高昂. 实验室数据集规模一般不大, 主要有以下 5 个:

1)GRID[59]. 由美国谢菲尔德大学团队提出, 包括 34 位说话人(18 位男性和 16 位女性), 每位说话人都会说出 1 000 个短句子, 每个句子包含 6 个词.

2)CREMA-D[60]. 可用于情感识别研究的音频-视觉数据集, 包含了来自不同族裔背景的 91 位说话人(48 位男性和 43 位女性)的 7 442 个包含脸部的视频片段. 说话人在 12 个句子中进行选择, 使用 6 种情绪(快乐、悲伤、恐惧、愤怒、厌恶、中性)之一呈现.

3)MEAD[61]. 该数据集记录了大约 40 h 的室内情感视频, 其中情绪中性且正面视角的视频可以作为 MEAD-Neutral 数据集.

4)Multiface[62]. 在实验室环境下收集了 13 个人物的多视角、高分辨率人脸数据集, 每个角色有超过 100 种面部表情.

5)SingingHead[63]. 该数据集是一个大规模唱歌头部数据集, 包括超过 27 h 的唱歌视频、3D 面部动作、

歌唱音频以及来自76个人物和8种音乐类型的背景音乐.

2.1.2 从互联网收集的数据集

互联网数据集是指从互联网收集而来，为了研究说话面部视频生成问题而特别筛选出的视频组成的数据集. 此类数据集视频的来源一般是一些主流的视频新闻网站，收集较为方便，规模较大，但是由于视频本身不是为了实验目的而录制，难免会有部分视频有质量不高、噪声较多等问题. 互联网数据集主要有以下10类：

1）LRW[64]. 来源于BBC广播电视节目，其中有1 000条包括500个单词的语句，包括29帧（约1 s）的视频、同步的音频以及视频中出现的目标单词. 整个数据集中出现了上千个说话者.

2）Voxceleb1[65]. 该数据集是一个经典的大规模开源音视频数据集，素材来源于Youtube网站，包含各种族、口音的英语视频，场景丰富，但分辨率较低、噪声较多.

3）LRS2[66]. 包含来自BBC电视节目的48 164个视频剪辑，每个视频都附带一个与句子相对应的音频，句子长度最多为100个字符，包含超过140 000条语句.

4）LRS3[67]. 包含来自TED或TEDx室内演讲的151 819个视频. LRS2和LRS3之间没有重叠的部分，这2个数据集的视频在拍摄场景、光照条件、动作等方面都是不同的.

5）Voxceleb2[68]. 该数据集与Voxceleb1数据集没有重复样本，在规模上前者是后者的5倍，包含来自145个不同国籍以及超过6 000名说话者的超过100万个样本的语音片段.

6）LRW-1000[69]. 该数据集是一个大规模的汉语唇读数据集，拥有超过2 000名不同发言人的718 018个样本，涵盖超过100万个汉字的实例，覆盖了286个汉语音节，包括1 000个类别，其中每个类别对应一个汉语词的音节，由1个或多个汉字组成.

7）HDTF[70]. 视频来源于Youtube视频网站，包含的人物身份有300多个，总共大约430个视频，分辨率为720 P或1 080 P，时长共约16 h.

8）CelebV-HQ[71]. 该数据集在维基百科数据集的基础上进行筛选处理而来，是一个高质量的名人视频数据集，包含35 666个视频片段，分辨率至少为512×512，涉及15 653个人物，并且所有片段都手动标注了83个面部属性，包括40个外貌属性、35个动作属性和8个情绪属性.

9）VFHQ[72]. 包括来自7 228个视频的16 827个多样的采访场景高保真片段，是由Youtube网站的高清视频经过多角度评估过滤得到的. 与Voxceleb1相比，VFHQ的视频分辨率要高得多.

10）EFHQ[73]. 该数据集是一个多用途极端姿态人脸高质量数据集，包括来自5 000个视频片段的约450 000帧的极端姿态人脸图像，其样本从VFHQ和CelebV-HQ筛选和采样而来，同时结合了同一主体的正面图像用作图像对.

表1列出了各个数据集，并对每个数据集从人物数量、数据规模、视频尺寸以及样本来源这4个方面进行了全面的比较.

Table 1 Audio-Driven Talking Face Video Generation Datasets

表1 音频驱动的说话面部视频生成数据集

数据集	年份	人物数量	数据规模	视频尺寸	样本来源
GRID[59]	2006	34	共34 000句	360×288，720×576	实验室环境录制
CREMA-D[60]	2014	91	每人12句	960×720	实验室环境录制
MEAD[61]	2020	60	每人20句		实验室环境录制
Multiface[62]	2022	13	每人50句	2 048×1 334，1 024×1 024	实验室环境录制
SingingHead[63]	2023	76	共447首歌曲	3 840×2 160	实验室环境录制
LRW[64]	2016	1 000多	共1 000句	224×224	BBC
Voxceleb1[65]	2017	1 200	共153 500句	224×224	Youtube
LRS2[66]	2018	500多	超过140 000句	224×224	BBC
LRS3[67]	2018	5 000多	超过152 000句	224×224	TED
Voxceleb2[68]	2018	6 100	共1 100 000句	224×224	Youtube
LRW-1000[69]	2019	2 000多	共71 800句	1 024×576，1 920×1 080	有线新闻节目
HDTF[70]	2021	300多	超过10 000句	1 280×720，1 920×1 080	Youtube
CelebV-HQ[71]	2022	15 653		512×512以上	维基百科
VFHQ[72]	2022		共16 827个片段	700×700~1 000×1 000	Youtube
EFHQ[73]	2024		约450 000帧	1 024×1 024	其他数据集

2.2 音频驱动的说话面部视频生成评价指标

1）主观评价指标是评估生成结果的质量和真实感的重要标准，主要是通过人类主观感受和主观评分来进行评价. 可以请受试者或专家对生成的面部视频进行主观评分，通常使用1~5或1~10的评分等级来评价生成结果的质量，包括逼真度、清晰度、自然度等方面，也可以将生成的面部视频与真实视频或其他生成结果进行比较，让受试者选择哪个更逼真、更自然，从而评估生成结果的优劣. 主观评价指

标能够充分考虑人类主观感受,因为最终用户对生成结果的满意度是最重要的.通过收集受试者的主观评分、比较和意见,可以了解他们对生成结果的真实感知和情感反应,但是主观评价指标容易受到个体主观偏好和主观判断的影响,这可能降低评价指标的可靠性和客观性,并且进行主观评价需要招募和培训受试者进行实验设计和数据收集,涉及较高的时间、人力和经济成本.

2)峰值信噪比(peak signal-to-noise ratio, *PSNR*)[74]是一种常用于衡量图像或视频质量的指标,它通过比较原始图像(或视频)与经过压缩或处理后的图像(或视频)之间的相对误差来评估质量.*PSNR* 的计算利用了原始图像与处理后图像之间每个对应像素值的差的平方的平均值,即均方误差(mean square error, *MSE*),其计算公式为

$$MSE=\frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}(I(i,j)-K(i,j))^2, \tag{1}$$

$$PSNR=10\times\lg\left(\frac{MAX_I^2}{MSE}\right), \tag{2}$$

其中 I 表示原始图像,K 表示处理后图像,像素大小均为 $m\times n$,MAX_I 表示图片可能的最大像素值.

PSNR 表示原始图像与处理后图像之间的信号与噪声比,其单位为分贝(dB),数值越高,表示图像质量相对较好.*PSNR* 的优点在于简单易懂、计算方便,但是它对人类视觉系统的感知不敏感,无法完全反映人眼对图像质量的主观感受,并且只能提供整体图像质量的信息,不能提供局部细节的质量评估.

3)结构相似性(sructural similarity, *SSIM*)[75]相比于 *PSNR*,考虑了亮度、对比度和结构 3 个方面,从而更全面地描述了图像之间的相似性.其计算公式为

$$SSIM=l(x,y)^{\alpha}\times c(x,y)^{\beta}\times s(x,y)^{\gamma}, \tag{3}$$

其中 $l(x,y)$,$c(x,y)$,$s(x,y)$ 分别表示亮度、对比度和结构的对比函数,α,β,γ 则代表不同特征在 *SSIM* 衡量中的占比.

SSIM 指标的优点在于它能够比较全面地评估图像的相似性,同时考虑了亮度、对比度和结构等多个方面,更符合人眼的直观感受.

4)CPBD(cumulative probability of blur detection)[76]在恰可察觉失真(just noticeable distortion, JND)模型基础上发展而来,是一种基于目标感知的无参考图像锐度评价方法,能够预测图像不同区域的锐度,且与人类视觉的主观感知有较好的相关性.

5)FID(Fréchet inception distance)[77]是一种用于评估 GAN 生成图像质量的指标,其利用 Inception 模型提取生成图像和真实图像的特征向量,再通过比较生成图像的特征向量分布与真实图像的特征向量分布之间的 Fréchet 距离得到.

FID 结合了生成图像的多样性和真实图像的分布信息,能够比较全面地评估生成图像的质量,其值越低,意味着真实图像的分布和生成图像的分布越接近,生成图像质量越高,但计算复杂度较高,需要使用预训练的 Inception 模型进行特征提取,同时对数据集的大小和特征的分布敏感.

6)学习感知图像块相似度(learned perceptual image patch similarity, LPIPS)[78]与传统的像素级别或统计特征级别的评价指标不同,LPIPS 基于深度学习模型学习到的感知相似性来评估图像之间的差异,其使用预训练的深度卷积神经网络(通常是 VGG 网络)提取图像的特征表示,并通过映射网络计算感知相似度.

LPIPS 通过使用深度学习模型学习特征表示,可以考虑图像的结构、纹理和语义信息,能够更好地反映人类视觉系统对图像质量的感知.

7)关键点距离(landmark distance, LMD)[26]是用于评估唇部动作准确性的一种指标,通过关键点来标注真实视频和生成视频的唇部坐标并比较 2 个视频中对应的关键点之间的欧氏距离来衡量生成视频中人物口型的准确度,关键点距离越小表示生成视频越准确,但由于关键点数量有限,LMD 难以检测面部细节的微小变化.

8)SyncNet 得分[12]包括 LSE-C(lip sync error-confidence)和 LSE-D(lip-sync error-distance),首先预先训练的 SyncNet 通过双通道网络提取视频和音频的特征,接着计算 2 类特征之间的距离,得到口型同步误差,从而量化口型语音同步.

3 说话面部视频鉴别方法概述

图 13 为音频驱动的说话面部视频鉴别方法的分类图.由于说话面部视频本质上是由一系列连续的伪造视频帧构成的,这些帧在时间和空间上紧密相连,共同构建了一个看似真实的面部动态表达.因此,生成的视频不仅需要保证视频帧纹理的真实感,而且要保证视频帧间的连续性[79],所以可以将现有的说话面部视频鉴别方法划分为两大类别:单帧鉴别方法和多帧鉴别方法.单帧鉴别方法主要侧重于对单个视频帧的细致分析,而多帧鉴别方法则更强调对连续视频帧之间关联性的深入探究.单帧及多帧

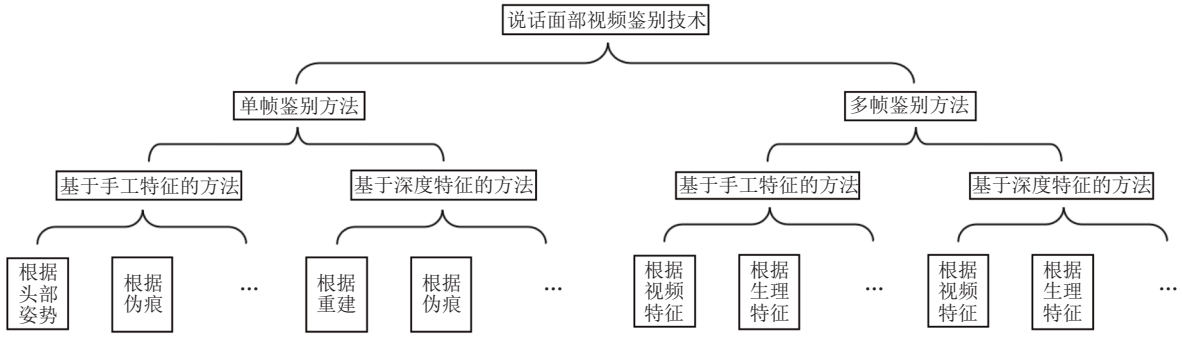


Fig. 13 Classification of audio-driven talking face video identification methods

图 13 音频驱动的说话面部视频鉴别方法分类

鉴别方法又都可以进一步分为基于手工特征的方法和基于深度特征的方法，基于手工特征的方法通常需要专业知识来进行特征设计和选择，而基于深度特征的方法则能够自动从大量数据中学习到有效复杂的特征表示.

3.1 单帧鉴别方法

单帧鉴别方法通常依赖于对单个视频帧进行特征提取和分析. 例如，可以利用传统的图像处理算法或基于深度学习的方法提取面部的纹理、形状、光照等特征，鉴别是否存在异常、矛盾或伪造痕迹. 这些方法的优点是对每个视频帧进行独立分析，不需要依赖于其他帧的信息，因此可以快速准确地鉴别出伪造.

3.1.1 基于手工特征的单帧鉴别方法

基于手工特征的单帧鉴别方法通常基于传统的计算机视觉和信号处理方法，这些方法主要依赖于手工设计的特征提取器，利用伪造内容的某些固有特征与真实面部图片特征不相符来判断人脸的真假.

3D 头部姿势特征常被用来检测图像是否被篡改，比如 Yang 等人[80] 认为许多修改面部图像的方法都是将合成的区域嵌入到原始图像中，而在这个过程中会引入错误，虽然这些错误人眼难以观察，但对这些图像进行 3D 头部姿势估计时，原始图像和生成图像中标记点位置不匹配会导致头部姿势之间的差异过大.

修改图像时产生的伪痕也可以作为鉴别的标准，例如 Matern 等人[81] 发现深度伪造方法在伪造新的面部图像时，利用数据插值会产生规律性的视觉伪痕，例如眼睛、鼻子或牙齿在细节上会有缺失，因此利用光流估计得到图像中的伪造痕迹，并以此判别图像真伪. Li 等人[82] 发现修改后的混合面部边界是可以被检测到的，提出使用面部 X 射线显示混合面部的边界伪痕，从而分析出视频帧伪造与否. Tan 等人[83] 则发现在 GAN 或 Diffusion 生成的合成图像中，上采样运算符会造成伪痕. 他们提出了一种简单而有效的表示方法，称为邻近像素关系（neighboring pixel relationships, NPR），旨在捕获由上采样操作产生的广义结构伪痕. NPR 可以展现对各类伪造方法的泛化处理能力，从而实现广义伪造图像鉴别[83].

手工特征种类非常丰富，不一而足，但总的来说，手工特征鉴别方法的适用条件较为单一，面对目前越来越复杂的深度学习方法生成的图像，其难以捕捉到复杂的深度特征，导致检测性能下降[84]，一些压缩或添加噪声的后期处理手段也可能会导致这些鉴别方法的失效.

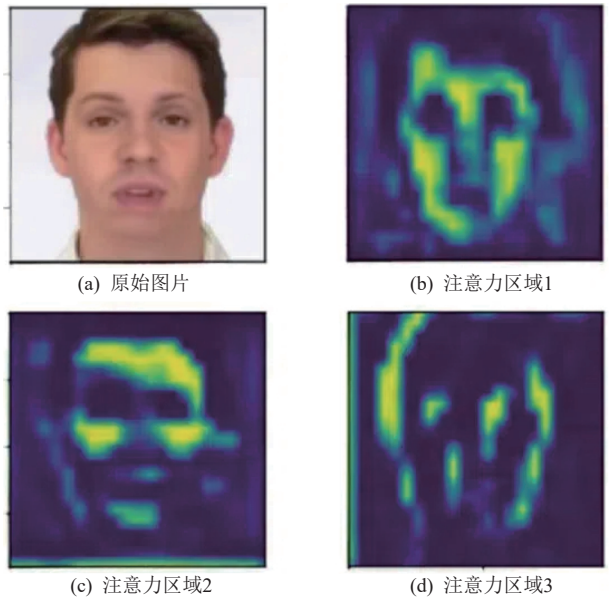
3.1.2 基于深度特征的单帧鉴别方法

随着近年来深度学习的快速发展，目前有许多鉴别手段基于深度特征. 深度学习以数据作为驱动源，通过神经网络提取图像中的抽象特征向量，更容易学习到可靠的深度特征，能够应对的场景也更加丰富. 基于深度特征的单帧鉴别方法通过训练深度神经网络模型，将输入的图像作为单个帧进行分析，以识别图像中的目标物体或特征.

Yang 等人[85] 利用 CNN 代替传统手工特征提取来执行面部单帧鉴别任务. 部分研究直接采用在图像分类领域性能十分优异的 CNN，如 Xception[86]、EfficientNet[87] 实现单帧鉴别的功能，这种做法在一些特定伪造方法中能够取得一定效果，但是泛化能力不强，面对训练集中未出现过的伪造帧时往往不能得到较好的鉴别效果.

生成图片过程中出现的伪痕也可以通过深度学习方法来获取，例如 Li 等人[88] 认为当前的面部生成方法只能生成有限分辨率的图像，并且需要进一步变形以匹配源视频中的原始面部，而这样的变换会在生成视频中留下独特的痕迹，而这些伪痕可以使用 CNN 捕捉. Zhao 等人[89] 关注真实面部和伪造面部间的局部细节伪痕，提出了一种基于多注意力机制

的纹理增强检测模型来更多关注面部细节并分析细微纹理伪痕间的不同,将图像中的伪痕特征进行细粒度分类,同时引入区域独立性损失捕获不同的局部伪痕特征,通过这种方式获取的多注意力区域如图 14 所示.



(a) 原始图片 (b) 注意力区域1 (c) 注意力区域2 (d) 注意力区域3

Fig. 14 Multi-attention regions[89]

图 14 多注意力区域[89]

利用检测特定伪痕的方法处理其他的伪造方式制作的图片难免局限于训练样本,缺少泛化性,一些工作直接针对图片的重建本身来进行鉴别. Cao 等人[90]提出了一个包含 3 部分的端到端的鉴别框架:其一为重建学习,包括编码器和解码器,旨在仅对真实面部图像的分布进行建模,学习各种可能的伪造模式;其二为多尺度图推理,旨在以多尺度方式聚合重建网络的编码器和解码器中关于真实和伪造面部之间的差异信息;其三为重建引导注意力,旨在引导分类网络更多地关注伪造痕迹,提高鉴别效果,其结构如图 15 所示. Wodajo 等人[91]使用的重建结构更为复杂,他们利用自编码器(auto-encoder, AE)和 VAE 从潜在数据分布中进行学习并重建图像,又结合了 ConvNeXt 和 Swin Transformer 模型提取原始图像和生成图像的视觉特征,以此来鉴别视频帧是否伪造. Yan 等人[92]则在重建过程中试图找到所有伪造方法共同的伪造特征,他们使用解耦框架将图像信息分解为与伪造无关的特征、与伪造方法相关的特征和共有伪造特征 3 个不同的组成部分,最后只利用共有伪造特征来实现可推广的深度伪造鉴别.

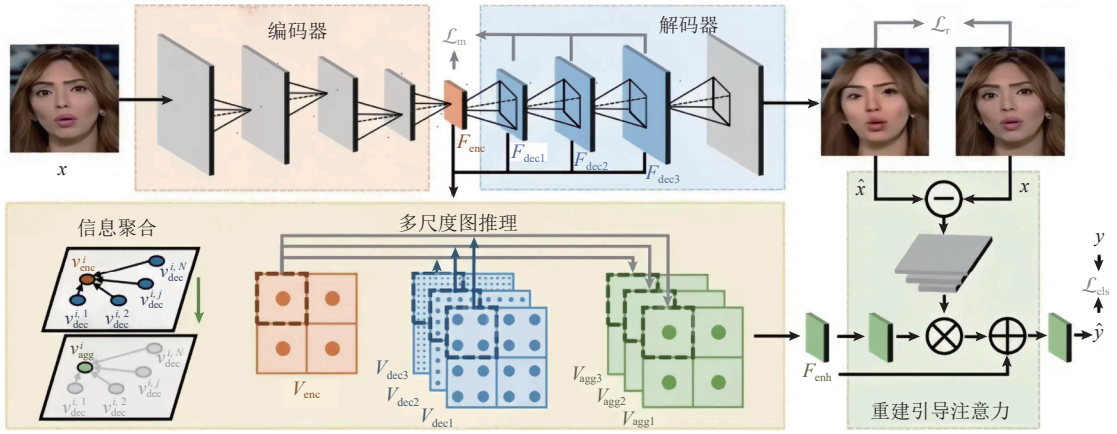


Fig. 15 Deep learning-based single-frame discrimination method[90]

图 15 基于深度学习的单帧鉴别方法[90]

3.2 多帧鉴别方法

多帧鉴别方法则更注重于视频序列中的时空一致性. 通过比较连续视频帧之间的变化、运动模式、面部位置和大小等信息,可以鉴别出伪造视频中的不一致之处. 这种方法的优点是能够捕捉到伪造视频中的动态特征和运动痕迹,对于一些精细的伪造可能具有更好的鲁棒性.

3.2.1 基于手工特征的多帧鉴别方法

基于手工特征的多帧鉴别方法首先使用传统的特征提取算法来捕获视频序列中的运动信息和空间关系,之后结合机器学习算法对目标进行鉴别和跟踪. 尽管这些方法在一定程度上能够实现目标鉴别的任务,但是由于手工设计特征难以捕获复杂场景下的高级语义信息,导致在处理复杂视频数据时性能有限.

一些方法针对视频本身的特征进行真假的鉴别. 例如 Koopman 等人[93]利用光响应不均匀性(photo response non-uniformity, PRNU)这一视频的固有物理

特性，从视频中提取帧并裁剪出待检查的面部均分为 8 组，通过比对各组视频帧的 PRNU 返回一个平均标准化互相关系数（normalized cross correlation score, NCCS），结果表明原视频的 NCCS 与伪造视频的明显不同，由此实现了对伪造视频的鉴别. 然而，此方法仅在自建的小型数据集上进行实验，说服力不够.

除了视频本身的特征，也可以利用人物的生理特征进行真假鉴别. Fernandes 等人[94] 利用心率信号鉴别伪造的视频，鉴别视频中因为血液流动导致的面部皮肤颜色变化、前额欧拉图像及其平均光密度变化，再依据神经常微分方程（neural ordinary differential equations, Neural ODE）来测算视频中人的心跳频率，由于伪造人脸时这个特征会遭到破坏，因此可以判断视频是否为伪造. 基于这项工作，Qi 等人[95] 根据血液流过面部会引起的皮肤颜色微小周期性变化，以及伪造视频中人物心跳节律会被破坏甚至完全消失的原理，设计并整合了心跳节律运动放大模块以及可学习的空间-时间注意力机制应用于网络模型的各个阶段.

3.2.2 基于深度特征的多帧鉴别方法

由于深度学习能够有效捕获和处理视频帧间的时空关系，越来越多的多帧鉴别方法开始采用深度特征. 这些方法利用视频序列中连续帧之间的时空信息，通过 DNN 来对目标进行跟踪和鉴别，从而在视频分析和行为识别等领域取得了显著的进展. 深度学习的引入使得多帧鉴别方法能够更好地理解视频中的动态变化和上下文信息，提高了对复杂场景下目标的准确性和鲁棒性.

以下方法致力于挖掘伪造视频本身的深度特征. Amerini 等人[96] 根据真实视频和伪造视频形成的光流大小、方向等方面都存在差异的特点，将光流特征送入神经网络中进行伪造视频的判别，在常见数据集上取得了不错的效果. Knafo 等人[97] 提出了一种依赖于多模态数据的新方法，首先以文字、视频、音频的多模态信息为输入，用自监督方式预训练多模态骨干网络，再微调骨干网络到视频深度伪造鉴别任务中. 伪造的视频一般会留下噪声痕迹，Wang 等人[98] 提出一种基于噪声的深度伪造鉴别模型，首先从连续视频帧中裁剪出的面部和背景正方形，再使其通过 Siamese 架构进行噪声特征提取，然后利用这 2 组噪音特征获取多头相对交互程度，以此进行深度伪造鉴别，其架构如图 16 所示.

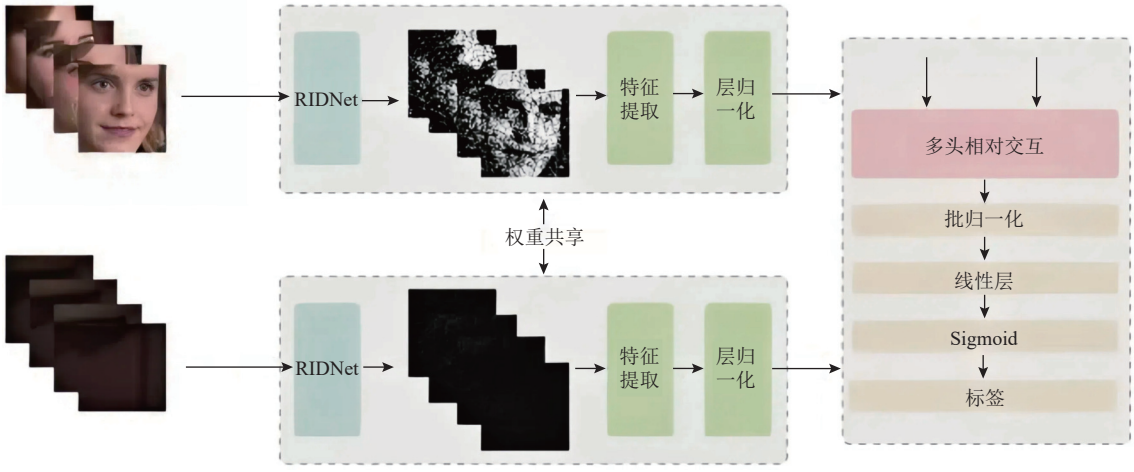


Fig. 16 Multi-frame discrimination method using video noise and deep learning [98]

图 16 利用视频噪声和深度学习的多帧鉴别方法[98]

视频中人物的生理特征也可以由深度学习方法挖掘. Li 等人[99] 以视频中人物的眨眼行为是否正常为标准，提出了一种基于眨眼频率的生理特征鉴别方法. 对于许多伪造面部视频来说，它们是基于大量的人脸图像进行训练得到的模型生成的，而大多数在网上获取的人物照片不会显示他们闭着眼睛，所以缺乏眨眼是伪造视频的一个显著特征. 这项工作利用长期循环卷积网络模型以考虑之前的时间知识来区分眼睛的睁开和闭合状态，检测视频的眨眼频率，发现许多伪造视频都低于正常标准，其架构如图 17 所示. Liu 等人[100] 提出了一种专门用于唇形伪造识别的新方法，通过自注意力模型捕捉不规则嘴唇运动，利用唇部运动和音频信号之间的不一致性进行口型同步伪造检测，还通过捕捉唇部和头部区域之间微妙的生物联系来模拟人类的自然认知来提高鉴别的准确性.

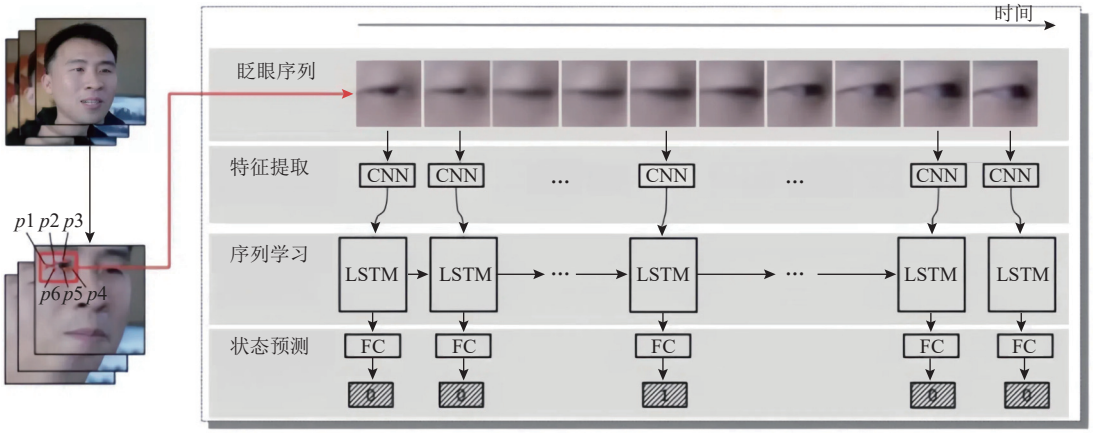


Fig. 17 Multi-frame discrimination method using physiological features and deep learning[99]

图 17 利用生理特征和深度学习的多帧鉴别方法[99]

4 说话面部视频鉴别数据集

说话面部视频鉴别数据集一般同时包含多个不同人物说话的原始视频以及利用不同伪造手段处理这些视频后得到的伪造视频，有些数据集会提供伪造算法，而有些不会提供. 随着鉴别方法的不断发展，针对不同的具体任务，新的数据集也在不断出现.

本节将介绍 12 个被广泛应用的常用的数据集.

1）UADFV[99] 是早期针对面部视频鉴别的数据集，包括 49 个真实视频和 49 个伪造视频，数据规模较小，视频被处理方式较为单一且效果不佳，视频整体质量较低，因此被鉴别出的难度也较低.

2）Deepfake-TIMIT[101] 是一个使用 GAN 创建的深度伪造视频数据集，其真实样本是从 VidTIMIT 数据库[102] 中提取的 640 个视频，分为高质量（128×128, HQ）和低质量（64×64, LQ）版本各 320 个，包含具有各种视觉瑕疵和修改的视频，整体清晰度不高.

3）FaceForensics++[86] 被广泛地使用，其中包括 1 000 个真实视频，并使用 deepfakes, face2face, faceswap, neural textures, faceshifter 这 5 种伪造方法得到 5 000 个伪造视频，且提供原视频（c0 或 Raw）、高质量（c23 或 HQ）以及低质量（c40 或 LQ）三种不同的压缩率. FaceForensics++数据集规模较大，但视频质量不够高.

4）DFDC[103] 是 Facebook 公司与 Kaggle 举办的伪造鉴别挑战赛中使用的数据集，共有约 12 万个真人拍摄视频，其中真实视频约占 1/6，采用了多种视频伪造方法. 所有均为现实生活场景，人脸占整个场景的比例较小，伪造痕迹也较为明显.

5）DeeperForensics-1.0[104] 包括 60 000 个真人拍摄视频，共计 1 760 万帧伪造视频则是通过一种新提出的端到端面部交换框架 DF-VAE 生成的，并且用光流法对生成的面部视频进行平滑处理，视频质量较好，规模庞大.

6）FFIW-10k[105] 包含 10 000 个伪造视频，采用 deepfacelab, FSGAN, faceswap 作为伪造手段. 其特殊之处在于每帧中都会出现多张人脸，平均每帧都会出现 3 张人脸，其中被伪造的人脸数量不定，这样更能模拟现实生活的实际状况，视频环境更加复杂.

7）ForgeryNet[106] 是商汤科技公司发布的大规模伪造视频数据集，共包括 17 种面部合成方法，且涵盖了图像级和视频级数据的 4 个任务：图像伪造分类、空间定位分类、视频伪造分类和时间伪造定位.

8）DF-Mobio[107] 有 46 000 个视频，其中包括 15 000 个伪造视频和 31 000 个真实视频，其真实样本来自 Mobio 数据集[108]，均为视频会议时使用手机或笔记本电脑时人脸直接面对摄像头的场景.

9）FMFCC-V[109] 包括 38 102 个伪造视频和 44 290 个真实视频，涵盖了 83 个亚洲人每人 40 min 的讲话内容，摄影环境包括室内和室外. 为了模拟真实情景，在真伪视频中均加入了 12 种类型的扰动，包括增亮、变暗、添加噪声等.

10）Div-Df[110] 由 150 个不同职业的名人视频和 250 个深度伪造视频组成，包括 100 个面部交换视频、100 个面部重新演绎视频以及 50 个唇同步视频，其中面部交换和唇同步的视频样本质量较高，面部重新演绎视频的样本质量一般.

11）DF-Platter[111] 从 Youtube 采集了 764 个印度族裔人物说话视频，通过多种伪造方法生成了 132 496

个高分辨率(HR)和低分辨率(LR)的伪造视频，包括单个主体和多个主体的面部伪造.

12)AV-Deepfake1M[112]是一个大规模音视频数据集，包括来自2 068名角色在多样化背景环境中捕获的1 886 h音视频数据，由视频伪造、音频伪造、音视频伪造3部分内容组成.

表2列出了各个数据集，并对每个数据集从数据规模、真伪样本比例以及真实样本来源这3个方面进行了全面比较.

Table 2 Talking Face Video Discrimination Datasets

表2 说话面部视频鉴别数据集

数据集	年份	规模	真伪样本比例	真实样本来源
UADFV[99]	2018	98个视频	1∶1	Youtube
Deepfake-TIMIT[101]	2018	1 280个视频	1∶1	VidTIMIT数据库[102]
FaceForensics++[86]	2019	6 000个视频	1∶5	Youtube
DFDC[103]	2020	119 154个视频	1∶5.26	演员录制
DeeperForensics-1.0[104]	2020	60 000个视频	1∶5	演员录制
FFIW-10k[105]	2021	20 000个视频	1∶1	Youtube
ForgeryNet[106]	2021	221 247个视频	1∶1.22	其他数据集
DF-Mobio[107]	2022	46 000个视频	1∶0.5	Mobio数据集[108]
FMFCC-V[109]	2022	82 392个视频	1∶0.86	演员录制
Div-Df[110]	2023	400个视频	1∶1.67	Youtube
DF-Platter[111]	2023	133 260个视频	1∶173.4	Youtube
AV-Deepfake1M[112]	2023	1 146 760个视频	1∶3	Youtube

5 研究难点与展望

5.1 音频驱动的说话面部视频生成

尽管音频驱动的说话面部生成技术近年来已经取得了显著的进展，但该领域仍蕴藏着诸多亟待深化与优化的空间，预示着未来广阔的发展空间，具体分为如下5个方面：

1)多语言视频合成. 鉴于现有音视频数据集多侧重于英语素材，视频生成模型往往经过特定优化以适应英语语境，然而，不同语言固有的韵律、声调及发音方式的差异性，使得此类模型在处理非英语口语面部动画时遭遇显著挑战，表现为生成视频中口型动作的自然度显著低于英语情境. 为攻克这一难题，目前亟需构建包含丰富语言种类的高品质数据集，并致力于研发具备更强语言通用性的模型架构，确保跨语言环境下的口语视频合成质量与表现力.

2)数据偏见问题. 在利用大规模视听数据集训练模型时，存在加剧已有说话者身份偏见的风险，包括但不限于种族、性别、年龄等特征的不当再现，这可能导致生成视频在应用于未充分代表的角色时，其逼真程度与自然度显著下降. 可以通过运用平衡数据集构成、增强模型适应性、实施偏见检测与矫正等方法缓解这一问题，促进生成视频内容的广泛适用性和公正性.

3)极端视角挑战. 在现有的训练数据集中，人物往往倾向于正对镜头的姿态，鲜少包含极端视角下的表现形式. 这一局限性导致在处理头部大幅度转动或非正面角度的图像及视频素材时，生成效果显著下降，主要因为面部某些部分因遮挡而缺失，影响了整体的连贯性和真实感. 为了克服这一难题，需要主动搜集并整合涵盖广泛视角的人物图像与视频资料以弥补现有数据集的不足，同时也需要设计并实现能够基于2D输入构建3D面部模型的生成算法，即使在面对遮挡或非典型视角的情况下，模型也能预测并合成缺失的面部细节.

4)精微调控的实现. 追求卓越的视频生成质量，意味着我们必须深化对输出内容的精细调节能力. 具体而言，除了确保嘴唇形态能够精准响应音频信号的变化之外，人物面部的其他特征，如眼神的灵动、表情的微妙转换，乃至整个面部区域的光影效果，都应当呈现出自然且和谐的演变. 这一目标的达成，要求我们的模型具备对视频内容进行细粒度、可控制编辑的强大功能.

5)高昂的计算与标注开销. 构建全面而详尽的音视频数据集需要人力密集型的手动标注工作. 当前主流的技术途径对标注数据的重度依赖，加之模型训练与执行过程中对计算资源的庞大需求，构成了技术发展的瓶颈. 为突破这一限制，可以探索自监督学习等方法以降低对标注的依赖，同时致力于研发轻量级模型.

5.2 说话面部视频鉴别

随着视频生成技术的进步，制作伪造视频变得更加容易，其传播也更加便捷. 为了减少伪造视频可能带来的负面后果，需要更先进的视频鉴别方法. 然而，目前存在以下3个需要解决的问题：

1)泛化能力. 现今，诸多面部检测方法侧重于精进其在特定数据库内的识别准确率，而这些方法往往未能展现出对各类伪造手法的全面适应性与高检测效能[113]. 特别是在遭遇复杂多变的面部篡改或新兴的生成方法时，现有模型的表现会出现明显退步. 鉴于此，增强算法的泛化能力，使其能够稳健地应对多样化的伪造手段至关重要.

2)对干扰的稳健性. 微小的输入视频变动可能会显著地影响检测模型的表现,导致其行为偏离预期轨迹. 这一问题在面对旨在误导鉴别器的对抗性方法时尤为突出,这些方法进一步加剧了任务的复杂度[114]. 因此,探索能够抵抗此类对抗性攻击的有效防御策略,对于提升伪造视频鉴别的可靠性十分重要.

3)实时检测. 尽管高精度的深度伪造检测方法能够提供准确的识别结果,但如果其推理时间过长,则可能限制其实用价值和广泛应用. 鉴于社交媒体用户基数的不断膨胀,以及深度伪造方法的创作与应用日益泛滥,高效能的实时检测方案变得更加关键.

6 总 结

随着深度学习的飞速进步,音频驱动的说话人面部视频生成技术经历了显著的迭代与优化. 利用前沿的深度学习模型,生成的视频质量日益提升,展现出前所未有的清晰度、生动性与逼真度. 这一技术革新不仅开辟了广阔的商业应用前景,同时也因其潜在的欺骗性,对社会安全、隐私保护及信息真实性构成了严峻挑战. 因此,对伪造视频的鉴别提出了更为严苛的要求,促使该领域研究不断深入. 本文从音频驱动说话人脸视频生成方法与视频防伪鉴别方法的核心概念出发,系统分类并总结这2项任务的演进历程,回顾并分析具有里程碑意义的经典算法与最新研究成果. 同时,本文还深入探讨了支撑这2项任务的关键数据集与评估方法,以期全面把握当前研究现状. 最后,本文深刻剖析了当前面临的挑战,并展望了未来可能的发展方向,为相关领域的研究人员与从业者提供有价值的参考与启示.

作者贡献声明: 乐铮负责文献调研并撰写论文;胡永婷对论文提出修改意见;徐勇提出指导意见并修改论文.

参 考 文 献

[1] Song Yifei, Zhang Wei, Chen Zhineng, et al. A survey on talking head generation[J]. Journal of Computer-Aided Design & Computer Graphics, 2023, 35(10): 1457−1468 (in Chinese)
(宋一飞, 张炜, 陈智能, 等. 数字说话人视频生成综述[J]. 计算机辅助设计与图形学学报, 2023, 35(10): 1457−1468)

[2] Bainey K. AI-Driven Project Management: Harnessing the Power of Artificial Intelligence and ChatGPT to Achieve Peak Productivity and Success[M]. Hoboken, NJ: John Wiley & Sons, 2024

[3] Zhang Yiwen, Cai Manchun, Chen Yonghao, et al. Multi-scale deepfake detection menthod with fusion of spatial features[J/OL]. Computer Engineering, [2024-07-06]. https://doi.org/10.19678/j.issn.1000-3428.0067789 (in Chinese)
(张溢文, 蔡满春, 陈咏豪, 等. 融合空间特征的多尺度深度伪造检测方法[J/OL]. 计算机工程, [2024-07-06]. https://doi.org/10.19678/j.issn.1000-3428.0067789)

[4] Sheng Wenjun, Cao Lin, Zhang Fan. Forged facial video detection based on supervised attention network[J]. Computer Engineering and Design, 2023, 44(2): 504−510 (in Chinese)
(盛文俊, 曹林, 张帆. 基于有监督注意力网络的伪造人脸视频检测[J]. 计算机工程与设计, 2023, 44(2): 504−510)

[5] Morishima S, Aizawa K, Harashima H. An intelligent facial image coding driven by speech and phoneme[C]//Proc of the 13th Int Conf on Acoustics, Speech, and Signal Processing. Piscataway, NJ: IEEE, 1989: 1795−1798

[6] Morishima S, Harashima H. A media conversion from speech to facial image for intelligent man-machine interface[J]. IEEE Journal on Selected Areas in Communications, 1991, 9(4): 594−600

[7] Yamamoto E, Nakamura S, Shikano K. Lip movement synthesis from speech based on hidden Markov models[J]. Speech Communication, 1998, 26(1/2): 105−115

[8] Lee S, Yook D S. Audio-to-visual conversion using hidden Markov models[C]//Proc of the 7th Pacific Rim Int Conf on Artificial Intelligence. Berlin: Springer, 2002: 563−570

[9] Aleksic P S, Katsaggelos A K. Speech-to-video synthesis using MPEG-4 compliant visual features[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2004, 14(5): 682−692

[10] Zhang Xinjian, Wang Lijuan, Li Gang, et al. A new language independent, photo-realistic talking head driven by voice only[C]//Proc of the 14th Annual Conf of the Int Speech Communication Association. New York: ISCA, 2013: 2743−2747

[11] Taylor S, Kim T, Yue Y, et al. A deep learning approach for generalized speech animation[J]. ACM Transactions on Graphics, 2017, 36(4): 93:1−93:11

[12] Chung J S, Zisserman A. Out of time: Automated lip sync in the wild[C]//Proc of the 13th Asian Conf on Computer Vision. Berlin: Springer, 2017: 251−263

[13] Chung J S, Jamaludin A, Zisserman A. You said that?[J]. arXiv preprint, arXiv: 1705.02966, 2017

[14] Karras T, Aila T, Laine S, et al. Audio-driven facial animation by joint end-to-end learning of pose and emotion[J]. ACM Transactions on Graphics, 2017, 36(4): 94:1−94:12

[15] Cudeiro D, Bolkart T, Laidlaw C, et al. Capture, learning, and synthesis of 3D speaking styles[C]//Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 10101−10111

[16] Fan Bo, Wang Lijuan, Soong F K, et al. Photo-real talking head with deep bidirectional LSTM[C]//Proc of the 40th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE,

2015: 4884−4888

[17] Fan Bo, Xie Lei, Yang Shan, et al. A deep bidirectional LSTM approach for video-realistic talking head[J]. Multimedia Tools and Applications, 2016, 75(9): 5287−5309

[18] 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

[19] Pham H X, Cheung S, Pavlovic V. Speech-driven 3D facial animation with implicit emotional awareness: A deep learning approach[C]//Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2017: 80−88

[20] Eskimez S E, Maddox R K, Xu Chenliang, et al. Generating talking face landmarks from speech[C]//Proc of the 14th Int Conf on Latent Variable Analysis and Signal Separation. Berlin: Springer, 2018: 372−381

[21] 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. Berlin: Springer, 2020: 716−731

[22] Zhou Yang, Han Xintong, Shechtman E, et al. Makelttalk: Speaker-aware talking-head animation[J]. ACM Transactions On Graphics, 2020, 39(6): 1−15

[23] Wang Suzhen, Li Lincheng, Ding Yu, et al. Audio2head: Audio-driven one-shot talking-head generation with natural head motion[J]. arXiv preprint, arXiv: 2107.09293, 2021

[24] Song Linsen, Wu W, Qian Chen, et al. Everybody's talkin': Let me talk as you want[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 585−598

[25] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139−144

[26] Song Yang, Zhu Jingwen, Li Dawei, et al. Talking face generation by conditional recurrent adversarial network[J]. arXiv preprint, arXiv: 1804.04786, 2018

[27] Chen Lele, Li Zhiheng, Maddox R K, et al. Lip movements generation at a glance[C]//Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 520−535

[28] 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

[29] Prajwal K R, Mukhopadhyay R, Philip J, et al. Towards automatic face-to-face translation[C]//Proc of the 27th ACM Int Conf on Multimedia. New York: ACM, 2019: 1428−1436

[30] 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]// Proc of the 28th ACM Int Conf on Multimedia. New York: ACM, 2020: 484−492

[31] Wang Jiadong, Qian Xinyuan, Zhang Malu, et al. Seeing what you said: Talking face generation guided by a lip reading expert[C]//Proc of the 36th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 14653−14662

[32] Yin Fei, Zhang Yong, Cun Xiaodong, et al. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan[C]//Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2022: 85−101

[33] Karras T, Laine S, Aittala 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

[34] Park S J, Kim M, Hong J, et al. SyncTalkFace: Talking face generation with precise lip-syncing via audio-lip memory[C]//Proc of the 36th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022: 2062−2070

[35] Goyal S, Bhagat S, Uppal S, et al. Emotionally enhanced talking face generation[C]//Proc of the 1st Int Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice. New York: ACM, 2023: 81−90

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

[37] Mittal G, Wang Baoyuan. Animating face using disentangled audio representations[C]//Proc of the 2020 IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2020: 3290−3298

[38] Liu Jin, Wang Xi, Fu Xiaomeng, et al. Font: Flow-guided one-shot talking head generation with natural head motions[C]//Proc of the 24th IEEE Int Conf on Multimedia and Expo. Piscataway, NJ: IEEE, 2023: 2099−2104

[39] Doersch C. Tutorial on variational autoencoders[J]. arXiv preprint, arXiv: 1606.05908, 2016

[40] Zhang Wenxuan, Cun Xiaodong, Wang Xuan, et al. SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation[C]//Proc of the 36th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 8652−8661

[41] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint, arXiv. 1706.03762, 2017

[42] Khan S, Naseer M, Hayat M, et al. Transformers in vision: A survey[J]. ACM Computing Surveys, 2022, 54(10): 1−41

[43] Fan Yingruo, Lin Zhaojiang, Saito J, et al. FaceFormer: Speech-driven 3D facial animation with transformers[C]//Proc of the 35th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE , 2022: 18770−18780

[44] Wang Jiayu, Zhao Kang, Zhang Shiwei, et al. LipFormer: High-fidelity and generalizable talking face generation with a pre-learned facial codebook[C]//Proc of the 36th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE , 2023: 13844−13853

[45] Zhong Weizhi, Fang Chaowei, Cai Yinqi, et al. Identity-preserving talking face generation with landmark and appearance priors[C]// Proc of the 36th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE , 2023: 9729−9738

[46] Ma Haoyu, Zhang Tong, Sun Shanlin, et al. CVTHead: One-shot controllable head avatar with vertex-feature transformer[C]//Proc of the 2024 IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2024: 6131−6141

[47] Mildenhall B, Srinivasan P P, Tancik M, et al. NeRF: Representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99−106

[48] Guo Yudong, Chen Keyu, Liang Sen, et al. AD-NeRF: Audio driven neural radiance fields for talking head synthesis[C]//Proc of the 18th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 5784−5794

[49] Gafni G, Thies J, Zollhofer M, et al. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction[C]//Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 8649−8658

[50] Yao Shunyu, Zhong Ruizhe, Yan Yichao, et al. DFA-NeRF: Personalized talking head generation via disentangled face attributes neural rendering[J]. arXiv preprint, arXiv: 2201.00791, 2022

[51] Tang Jiaxiang, Wang Kaisiyuan, Zhou Hang, et al. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition[J]. arXiv preprint, arXiv: 2211.12368, 2022

[52] Bi Chongke, Liu Xiaoxing, Liu Zhilei. NeRF-AD: Neural radiance field with attention-based disentanglement for talking face synthesis[J]. arXiv preprint, arXiv: 2401.12568, 2024

[53] Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]// Proc of the 32nd Int Conf on Machine Learning. New York: ACM, 2015: 2256−2265

[54] Bigioi D, Basak S, Stypułkowski M, et al. Speech driven video editing via an audio-conditioned diffusion model[J]. arXiv preprint, arXiv: 2301.04474, 2023

[55] Stypułkowski M, Vougioukas K, He Sen, et al. Diffused Heads: Diffusion models beat gans on talking-face generation[C]//Proc of the 2024 IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2024: 5091−5100

[56] Shen Shuai, Zhao Wenliang, Meng Zibin, et al. DiffTalk: Crafting diffusion models for generalized audio-driven portraits animation[C]//Proc of the 36th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE , 2023: 1982−1991

[57] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proc of the 35th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 10684−10695

[58] Zhang Bingyuan, Zhang Xulong, Cheng Ning, et al. EmoTalker: Emotionally editable talking face generation via diffusion model[C]//Proc of the 49th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2024: 8276−8280

[59] Cooke M, Barker J, Cunningham S, et al. An audio-visual corpus for speech perception and automatic speech recognition[J]. The Journal of the Acoustical Society of America, 2006, 120(5): 2421−2424

[60] Cao Houwei, Cooper D G, Keutmann M K, et al. CREMA-D: Crowd-sourced emotional multimodal actors dataset[J]. IEEE Transactions on Affective Computing, 2014, 5(4): 377−390

[61] Wang Kaisiyuan, Wu Qianyi, Song Linsen, et al. MEAD: A large-scale audio-visual dataset for emotional talking-face generation[C]// Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 700−717

[62] Wuu C, Zheng Ningyuan, Ardisson S, et al. Multiface: A dataset for neural face rendering[J]. arXiv preprint, arXiv: 2207.11243, 2022

[63] Wu Sijing, Li Yunhao, Zhang Weitian, et al. SingingHead: A large-scale 4D dataset for singing head animation[J]. arXiv preprint, arXiv: 2312.04369, 2023

[64] Chung J S, Zisserman A. Lip reading in the wild[C]//Proc of the 13th Asian Conf on Computer Vision. Berlin: Springer, 2017: 87−103

[65] Nagrani A, Chung J S, Zisserman A. Voxceleb: A large-scale speaker identification dataset[J]. arXiv preprint, arXiv: 1706.08612, 2017

[66] Afouras T, Chung J S, Senior A, et al. Deep audio-visual speech recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 44(12): 8717−8727

[67] Afouras T, Chung J S, Zisserman A. LRS3-TED: A large-scale dataset for visual speech recognition[J]. arXiv preprint, arXiv: 1809.00496, 2018

[68] Chung J S, Nagrani A, Zisserman A. Voxceleb2: Deep speaker recognition[J]. arXiv preprint, arXiv: 1806.05622, 2018

[69] Yang Shuang, Zhang Yuanhang, Feng Dalu, et al. LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild[J]. arXiv preprint, arXiv: 1810.06990, 2018

[70] Zhang Zhimeng, Li Lincheng, Ding Yu, et al. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset[C]//Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 3661−3670

[71] Zhu Hao, Wu W, Zhu Wentao, et al. CelebV-HQ: A large-scale video facial attributes dataset[C]//Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2022: 650−667

[72] Xie Liangbin, Wang Xintao, Zhang Honglun, et al. VFHQ: A high-quality dataset and benchmark for video face super-resolution[C]// Proc of the 35th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 657−666

[73] Dao T T, Vu D H, Pham C, et al. EFHQ: Multi-purpose ExtremePose-Face-HQ dataset[C]//Proc of the 37th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 22605−22615

[74] Hore A, Ziou D. Image quality metrics: PSNR vs SSIM[C]//Proc of the 20th Int Conf on Pattern Recognition. Piscataway, NJ: IEEE, 2010: 2366−2369

[75] Wang Zhou, Bovik A C, Sheikh H R, et al. Image quality assessment: From error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600−612

[76] Narvekar N D, Karam L J. A no-reference image blur metric based on the cumulative probability of blur detection (CPBD)[J]. IEEE Transactions on Image Processing, 2011, 20(9): 2678−2683

[77] Heusel M, Ramsauer H, Unterthiner T, et al. GANs trained by a two time-scale update rule converge to a local Nash equilibrium[J]. arXiv preprint, arXiv: 1706.08500, 2017

[78] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proc of the 31st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 586−595

[79] Sun Yu, Zhu Xinjuan. Text audio driven facial animation generation based on improved Wav2Lip[J]. Computer Systems & Application, 2024, 33(2): 276−283 (in Chinese)
(孙瑜, 朱欣娟. 改进 Wav2Lip 的文本音频驱动人脸动画生成[J].

计算机系统应用, 2024, 33(2): 276−283)

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

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

[82] 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

[83] Tan Chuangchuang, Liu Huan, Zhao Yao, et al. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection[J]. arXiv preprint, arXiv: 2312.10461, 2023

[84] Wei Zhengzheng. Deepfake detection based on local texture difference feature enhancement[J]. Journal of Chongqing Technology and Business University: Natural Sciences Edition, 2025, 42(2): 78−85 (In Chinese)
(韦争争. 基于局部纹理差异特征增强的 Deepfake 检测方法[J]. 重庆工商大学学报: 自然科学版, 2025, 42(2): 78−85)

[85] Yang Jianwei, Lei Zhen, Li S Z. Learn convolutional neural network for face anti-spoofing[J]. arXiv preprint, arXiv: 1408.5601, 2014

[86] 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

[87] Tan Mingxing, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//Proc of the 36th Int Conf on Machine Learning. New York: ACM, 2019: 6105−6114

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

[89] 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

[90] Cao Junyi, Ma Chao, Yao Taiping, et al. End-to-end reconstruction-classification learning for face forgery detection[C]//Proc of the 35th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 4113−4122

[91] Wodajo D, Atnafu S, Akhtar Z. Deepfake video detection using generative convolutional vision transformer[J]. arXiv preprint, arXiv: 2307.07036, 2023

[92] Yan Zhiyuan, Zhang Yong, Fan Yanbo, et al. UCF: Uncovering common features for generalizable deepfake detection[J]. arXiv preprint, arXiv: 2304.13949, 2023

[93] Koopman M, Rodriguez A M, Geradts Z. Detection of deepfake video manipulation[C]//Proc of the 20th Irish Machine Vision and Image Processing Conf. Dublin, Ireland: IPRCS, 2018: 133−136

[94] Fernandes S, Raj S, Ortiz E, et al. Predicting heart rate variations of deepfake videos using neural ODE[C]//Proc of the 17th IEEE/CVF Int Conf on Computer Vision Workshops. Piscataway, NJ: IEEE, 2019: 1721−1729

[95] Qi Hua, Guo Qing, Xu J F, et al. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms[C]//Proc of the 28th ACM Int Conf on Multimedia. New York: ACM, 2020: 4318−4327

[96] Amerini I, Galteri L, Caldelli R, et al. Deepfake video detection through optical flow based CNN[C]//Proc of the 17th IEEE/CVF Int Conf on Computer Vision Workshops. Piscataway, NJ: IEEE, 2019: 1205−1207

[97] Knafo G, Fried O. FakeOut: Leveraging out-of-domain self-supervision for multi-modal video deepfake detection[J]. arXiv preprint, arXiv: 2212.00773, 2022

[98] Wang Tianyi, Chow K P. Noise based deepfake detection via multi-head relative-interaction[C]//Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 14548−14556

[99] Li Yuezun, Chang M C, Lyu S. In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking[C/OL]//Proc of the 10th IEEE Int Workshop on Information Forensics and Security. Piscataway, NJ: IEEE, 2018[2024-3-15]. https://ieeexplore.ieee.org/document/8630787

[100] Liu Weifeng, She Tianyi, Liu Jiawei, et al. Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes[J]. arXiv preprint, arXiv: 2401.15668, 2024

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

[102] Sanderson C. The VidTIMIT Database[DB/OL].[2024-07-08]. http://conradsanderson.id.au/vidtimit/

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

[104] 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

[105] 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

[106] 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

[107] Korshunov P, Marcel S. Improving generalization of deepfake detection with data farming and few-shot learning[J]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2022, 4(3): 386−397

[108] McCool C, Marcel S, Hadid A, et al. Bi-modal person recognition on a mobile phone: Using mobile phone data[C]///Proc of the 2012 IEEE Int Conf on Multimedia and Expo Workshops. Piscataway, NJ: IEEE, 2012: 635−640

[109] Li Gen, Zhao Xianfeng, Cao Yun, et al. FMFCC-V: An Asian large-scale challenging dataset for deepfake detection[C]//Proc of the 10th ACM Workshop on Information Hiding and Multimedia Security.

New York: ACM, 2022: 7−18

[110] Dagar D, Vishwakarma D K. Div-Df: A diverse manipulation deepfake video dataset[C/OL]//Proc of the 2023 Global Conf on Information Technologies and Communications. Piscataway, NJ: IEEE, 2023[2024-03-15]. https://ieeexplore.ieee.org/document/10426446

[111] Narayan K, Agarwal H, Thakral K, et al. Df-Platter: Multi-face heterogeneous deepfake dataset[C]//Proc of the 36th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 9739−9748

[112] Cai Zhixi, Ghosh S, Adatia A P, et al. AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset[J]. arXiv preprint, arXiv: 2311.15308, 2023

[113] Dong Lin, Huang Liqing, Ye Feng, et al. Survey on generalization methods of face forgery detection[J]. Computer Science, 2022, 49(2): 12−30 (in Chinese)
(董琳, 黄丽清, 叶锋, 等. 人脸伪造检测泛化性方法综述[J]. 计算机科学, 2022, 49(2): 12−30)

[114] 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 Workships. Piscataway, NJ: IEEE, 2020: 658−659



Le Zheng, born in 2002. Master candidate. His main research interests include computer vision and video generation.

乐　铮, 2002 年生. 硕士研究生. 主要研究方向为计算机视觉、视频生成.



Hu Yongting, born in 1999. PhD candidate. Her main research interest includes unsupervised video anomaly detection.

胡永婷, 1999 年生. 博士研究生. 主要研究方向为无监督视频异常检测.



Xu Yong, born in 1972. PhD, professor. Member of CCF. His research interests include pattern recognition, computer vision, deep learning, and video analysis.

徐　勇, 1972 年生. 博士, 教授. CCF 会员. 主要研究方向为模式识别、计算机视觉、深度学习、视频分析.