doi:10.11887/j.cn.202402013 http://journal.nudt.edu.cn

深度视觉语音生成研究进展与展望

刘 丽[1],隋金坪[2],丁 丁[3],赵凌君[1],匡纲要[1],盛常冲[4*]

(1. 国防科技大学 电子科学学院,湖南 长沙 410073;2. 海军大连舰艇学院 作战软件与仿真研究所,辽宁 大连 116016;
3. 国防科技大学 教研保障中心,湖南 长沙 410073;4. 海军工程大学 电磁能技术全国重点实验室,湖北 武汉 430033)

摘 要:为了进一步推进深度学习技术驱动的视觉语音生成相关科学问题的研究进展,阐述了视觉语音生成的研究意义与基本定义,并深入剖析了该领域面临的难点与挑战;在此基础上,介绍了目前视觉语音生成研究的现状与发展水平,基于生成框架的区别对近期主流方法进行了梳理、归类和评述;最后探讨视觉语音生成研究潜在的问题和可能的研究方向。

关键词:视觉语音生成;深度学习;计算机视觉;计算机图形学

中图分类号:TP183 **文献标志码**:A **开放科学(资源服务)标识码(OSID)**:

文章编号:1001-2486(2024)02-123-16

Reasearch progress and prospects of deep learning for visual speech generation

LIU Li[1], *SUI Jinping*[2], *DING Ding*[3], *ZHAO Lingjun*[1], *KUANG Gangyao*[1], *SHENG Changchong*[4*]

(1. College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China;
2. Operational Software and Simulation Institute, Dalian Navy Academy, Dalian 116016, China;
3. Center for Teaching and Research Support, National University of Defense Technology, Changsha 410073, China;
4. National Key Laboratory of Electromagnetic Energy, Naval University of Engineering, Wuhan 430033, China)

Abstract: In order to further advance the development of visual speech learning, the task definition and research significance of visual speech generation was expounded and the difficulties and challenges were deeply analyzed in this field. Besides, the current status and development level of visual speech generation research was introduced, and the recent mainstream methods were sorted, classified and commented based on the difference of generation frameworks. At the end of the paper, the potential problems and possible research directions of visual speech generation were discussed.

Keywords: visual speech generation; deep learning; computer vision; computer graphics

人类产生语音信号以及对语音的感知过程本质上是视觉和听觉双模态的。视觉语音是指语音信号在视觉域的表现形式,即人在说话时产生的嘴唇、舌头、牙齿、下巴以及其他面部肌肉的自发运动[1],而音频语音则是指说话者所发出的声波波形。1976年,著名的“麦格克效应”[2]表明,人类的语音感知是一种感性认知现象,不仅只取决于听觉信息,还会受到嘴唇运动等视觉线索的影响。因此,不可否认对视觉语音的研究有助于提升人类对语音感知的主观舒适度,特别是对于遭受听力受损或听力障碍的人群。

作为计算机视觉、计算机图形学和多媒体领域的一个基本且具有挑战性的课题,视觉语音生成(visual speech generation, VSG)近年来受到越来越多的关注,因为它在许多新兴应用中发挥着重要作用。其典型的学术和现实应用包括说话人识别与验证[3]、医疗救助、公共安全、视频压缩、影视娱乐、人机交互、情感理解[4-5]等。例如:在公共安全领域,视觉语音可以应用于人脸伪造检测[6]和活体检测[7]。在人机交互中,视觉语音可以作为一种新型交互信息,提高交互的多样性和鲁棒性[8-9]。在影视娱乐领域,VSG技术可以在虚拟游戏中生成语音驱动的个性化3D虚拟角色[10],以及为电影后期制作(如视觉重配音)实现高保真度的视频合

收稿日期:2022-06-09

基金项目:国家自然科学基金资助项目(61872379)

第一作者:刘丽(1982—),女,湖南长沙人,研究员,博士,硕士生导师,E-mail:lilyliu_nudt@163.com

*通信作者:盛常冲(1993—),男,湖北孝感人,工程师,博士,E-mail: shengcc_nudt@163.com

成[11]等。

VSG 的核心研究点在于视觉语音表征学习和序列建模。在传统机器学习方法占主导地位的时代,视觉语音的表征方法如视素[12-13]、唇部几何描述符[14]、线性变换特征[15]、统计表示[16]等,以及序列建模方法如高斯过程动力学模型[17]、隐马尔可夫模型[18](hidden Markov model, HMM)、决策树模型[19]等被广泛应用于 VSG 研究。自深度神经网络(deep neural networks, DNNs)在图像分类任务中取得重大突破[20]以来,绝大多数计算机视觉和自然语言相关问题的研究热点都聚焦在深度学习方法上,包括 VSG 问题。2016 年,基于深度学习的 VSG 方法[21]在性能上大幅度超越传统方法,引领 VSG 进入深度学习时代。同时,大规模音视数据集[22-25]的不断涌现也进一步推动了深度学习驱动的 VSG 研究。因此,本文主要关注基于深度学习的 VSG 方法。视觉语音技术从 2016 年至今的里程碑工作如图 1 所示,包括具有代表性的深度 VSG 方法以及相关的音视基准数据集。

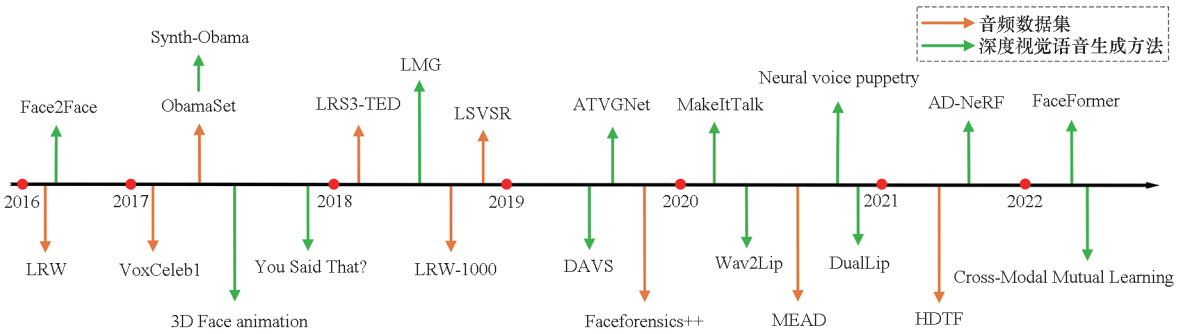


图 1 从 2016 年至今的视觉语音生成的时间里程碑

Fig. 1 Milestones on visual speech generation from 2016 to the present

尽管最近几年深度学习为 VSG 带来了可喜的进展,但不可否认 VSG 的研究仍处于早期阶段,尚未达到满足实际应用的水平,许多问题仍待解决。因此,系统性地回顾该领域的最新进展,总结阻碍其发展的主要挑战和未解决的问题,并探讨和挖掘有潜力的发展方向是非常有价值的。Mattheyses 等[26]广泛而全面地对音视语音生成进行了总结和讨论。我们建议读者参考该论文以了解视觉语音生成在 2015 年之前的详细发展历程。本文在其基础上,聚焦于深度学习驱动的 VSG 研究进展。Chen 等[27]对说话人身份独立的 VSG 方法进行了综述。以身份信息保留、音视同步以及视觉质量三个核心需求为牵引,对相关方法进行了讨论分析,并为其设计了性能评估基准。他们的核心贡献在于提出并定义了明确的评估标准,而非 VSG 方法的全面讨论和总结。总之,已有 VSG 综述调研性的工作在时效性、前瞻性、统筹性等方面尚有不足。因此,本文旨在进一步填补该领域综述调研上的空白。

本文对当前 VSG 的基本定义和研究意义、所面临的难点与挑战、主流的深度学习驱动方法进行了系统性的介绍、阐述和归纳。并于文末探讨了研究潜在的问题及未来可能的研究方向,以期进一步推动与此相关问题的研究。

1 定义与挑战

1.1 问题定义

视觉语音分析包括两个基本问题:识别和生成。这两个问题是形式对偶且相互促进的。为了更好地理解视觉语音生成问题,在此对这两个问题及其关联性进行简要介绍。

视觉语音学习的两个形式对偶基本问题如图 2 所示,视觉语音识别(visual speech recognition, VSR),也称为唇读,旨在根据说话人的嘴唇区域的视觉动态变化推断其所表达的语言内容,陈小鼎等[28]已对该问题做出较为详尽的调研。作为 VSR 的对偶任务,VSG 的目标是合成与驱动源(一段参考音频或文本)以及目标说话人相对应的逼真的高质量的唇动视频。具体来说,视觉语音生成系统首先从语音驱动源中提取语音

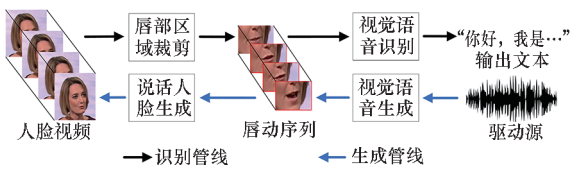


图 2 视觉语音学习的两个形式对偶基本问题

Fig. 2 Two formal-dual fundamental problems of visual speech learning

特征表示,然后将学习到的语音特征表示与目标身份特征融合并最终输出连续的目标唇动图像序列。从学习目标的角度来说,VSG 的目标比 VSR 更加主观化和多样化,这也使得 VSG 相比于 VSR 更加具有挑战性。

1.2　难点与挑战

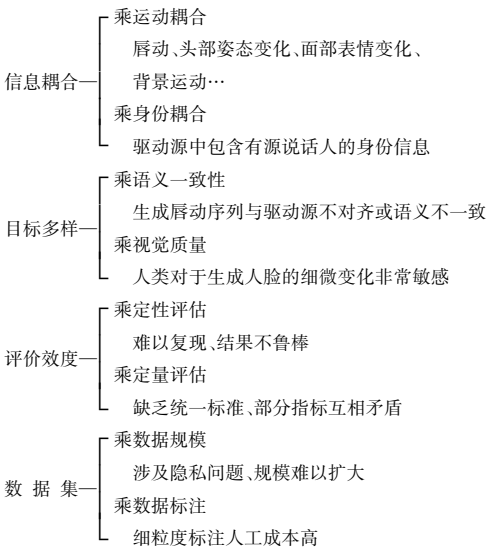
尽管深度学习极大地促进了 VSG 的发展,但由于各种现实挑战,大多数方法仍无法满足实际应用的要求。视觉语音学习中的挑战分类如图 3(a)所示,为了系统地阐述视觉语音生成所面临的挑战,本文将主要挑战从信息耦合、目标多样、评价效度和数据集四个方面进行阐述。图 3(b)和图 3(c)中提供了一些典型挑战的可视化实例以供读者直观理解。

说话人脸视频包含多种耦合信息,例如各种运动信息和说话人身份相关信息。运动信息可以分为两种类型:内在运动(头部姿势、面部表情、嘴唇运动等)和外在运动(相机运动、背景运动等),所有这些不同的运动信息都是高度耦合的。运动耦合所带来的挑战不仅源于将嘴唇运动与所有这些与语音无关的运动解耦开,还存在于将合成的嘴唇运动序列耦合到给定的目标人脸图像中。

VSG 的另一种信息耦合问题是身份信息耦合。如图 3(b)所示,由于生成的目标说话人脸耦合了部分语音驱动源人物信息,人类在观察这些图像时会感到怪异和不适。这种现象也被称为“恐怖谷效应”[29],当人们观察到一张类似但并不完美的合成人脸时就会产生这种不舒服的感性认知。因此,相应的挑战是如何从驱动源中去除源身份信息,以避免目标人脸合成过程中的身份耦合问题。此外,由于不同的说话人在外表、语速、说话习惯等方面存在显著差异,大多数现有方法仅适用于特定范围内的目标说话人脸生成。因此,缺乏说话人泛化能力也是一个重要的挑战。

语义一致性和视觉质量是 VSG 方法最重要的属性。语义一致性表示合成的唇形序列应与语音驱动源同步且语音一致。如图 3(c)所示,语义一致性主要涉及两个方面:时域对齐和语音匹配。然而,由于语音信号在不同模态的时间分辨率和内在特性的差异,难以有效保证驱动源和生成的谈话视频之间的同步语音映射。

至于视觉质量,存在两个难点:①缺乏明确的训练目标,这是因为生成的唇部运动序列的保真度和视觉质量难以进行定量定义;②因为人类对细微的伪影很敏感,所以将生成的唇形序列整合到整个面部而不会出现面向人的感知错误是一个复杂的问题。



(a) 视觉语音学习中的挑战分类

(a) Taxonomy of challenges in VSG

(b) 身份耦合、低视觉质量

(b) Identity coupling and low visual quality

(c) 音视语义不匹配

(c) Audio-visual semantic mismatch

图 3　视觉语音生成的主要挑战

Fig. 3　Main challenges of visual speech generation

除上述困难之外,有效评估 VSG 方法是另一大挑战。VSG 现有评估指标,包括定性和定量指标都存在明显的局限性。例如:像用户主观评价这样的定性指标是不可复现且不稳定的。虽然目前存在十多个定量评价指标,但其中一些并不合适该任务,甚至部分指标在评估结论上存在相互矛盾。

音视数据集相关问题也显著影响着视觉语音生成的发展。由于当前大多数深度学习方法都是数据驱动的,因此数据集的重要性不言而喻。然而,由于隐私保护及降低人力成本的需求,现有的音视数据集存在规模小、弱标注的问题。现有部分工作[30-32]尝试研究基于无标签的音视数据实现跨模态自监督视觉语音学习。尽管如此,数据集规模限制的问题仍有待解决。

2 数据集与评估指标

数据集在视觉语音生成的整个研究历史中都发挥了重要作用,尤其是在大数据时代。首先,基准数据集可以作为衡量和比较 VSG 方法性能的通用平台;其次,作为典型的数据驱动学习策略,深度学习技术所取得的重大进展很大程度上依赖于大规模标注数据集;最后,数据集还进一步推动该领域解决日益复杂和更具挑战性的问题。因此,本节回顾现有常用的音视数据集,进而介绍针对 VSG 方法现有的评估指标。表 1 总结了数据集统计信息。

表 1 常用音视数据集统计信息

Tab. 1 Statistics of commonly used audio-visual datasets

名称	时长/h	词汇量	话语数量	说话人数	尺寸/像素	帧率/(帧/s)	录制环境	发布年份
GRID[33]	≈28	51	33×10^3	33	720×576	25	实验室	2006
MODALITY[34]	≈31	182	5 880	35	$1\,920\times1\,080$	100	实验室	2015
OuluVS2[35]	≈2	—	2 120	53	$1\,920\times1\,080$	30	实验室	2015
LRW[22]	≈111	500	$\approx539\times10^3$	$>1\times10^3$	256×256	25	BBC 节目	2016
LRS2-BBC[36]	≈225	$\approx62.8\times10^3$	$\approx144.5\times10^3$	$>1\times10^3$	160×160	25	BBC 节目	2017
ObamaSet[37]	≈14	—	—	1	—	—	Youtube 视频	2017
LRS3-TED[38]	≈475	$\approx71.7\times10^3$	$\approx151.8\times10^3$	$>5\times10^3$	224×224	25	TED 视频	2018
VoxCeleb2[24]	$\approx2.4\times10^3$	—	$\approx1.1\times10^6$	$>6\times10^3$	—	—	Youtube 视频	2018
LRW-1000[25]	≈57	1×10^3	$\approx718\times10^3$	$>2\times10^3$	—	—	中国电视节目	2019
VOCASET[39]	—	—	255	12	5 023 顶点	60	实验室	2019
MEAD[40]	≈39	—	—	60	$1\,920\times1\,080$	30	实验室	2020
HDTF[41]	≈15.8	—	$>10\times10^3$	>300	—	—	互联网视频	2021

2.1 数据集

现有的音视数据集大体可以分为两类:面向受控环境和面向非受控环境。

2.1.1 受控环境下的音视数据集

在 2015 年之前,视觉语音研究主要面向受控环境下的音视数据。控制因素包括录制条件、录制设备、数据类型、语料内容等。这些数据集为视觉语音研究提供了良好的基础。本小节回顾了在受控环境下收集的一些具有代表性的音视数据集。

GRID[33]由 34 位志愿者参与,记录他们说的 1 000个句法相同短语的高质量音频和视频,专为全面的音视感知分析和微观建模而构建。它可支撑多项音视任务,包括视觉语音生成、音视语音识别、音视语音分离等。

MODALITY[34]包含总共约 31 h 的高帧率立体视频流数据,它与其他数据集的区别在于其录制设备为高分辨率 RGB-D 相机。

OuluVS2[35]是为视觉语音分析而构建的多视图音视数据集。它记录了 53 位志愿者发出的三种特定语料类型的话语。此外,由于用户在说话时可能不会一直面对摄像机,为模拟真实环境,该数据集从正面和多个侧面共五个不同视角同步录制。

VOCASET[39]是一个 4D 说话人脸数据集,包含大约 29 min 的 4D 人脸扫描,以及来自 12 位说话者(6 位女性和 6 位男性)的同步音频,以 60 帧/s的帧率录制。VOCASET 作为具有代表性的高质量 4D 人脸视听数据集,极大地推动了 4D 视觉语音生成的研究。

MEAD[40],即 Multi-view Emotional Audio-visual Dataset,是一个大规模、高质量、多视图的情感音视数据集。与以前的数据集不同,它专注于带有自发情绪的说话人脸生成,并定义和引入了多种情绪状态(三种强度级别的八种不同情绪)。

2.1.2 非受控环境下的音视数据集

为进一步贴近现实应用,研究人员逐渐将注意力转移到非受控环境下的视觉语音学习上。因此,近期出现许多大规模的非受控环境下音视数据集。

LRW[22]是由多阶段数据自动采集管道构建的单词级音视数据集。它基于 BBC 节目的丰富数据量,革命性地扩大了数据集规模和说话人数

量。它包含超过 1 000 人所说的超过 100 万个单词实例。LRW 的主要目标是测试说话人独立的单词级唇读方法,同样可用于视觉语音生成研究。

LRS2-BBC[36]是一个语句级音视数据集,具有与 LRW 数据集相似的数据收集管道和数据源。LRS2-BBC 中的所有视频均来自 BBC 新闻节目,其中包含超过 144.5×10^3 个的话语实例,词汇量约为 62.8×10^3。

ObamaSet[37]是一个特殊的音视数据集,专注于美国前总统奥巴马的视觉语音分析。所有视频样本都是从他每周的演讲视频中收集的。与以前的数据集不同,它只关注奥巴马,且不提供任何人工标注。因此,它仅用于面向奥巴马的视觉语音生成研究。

LRS3-TED[38]是一个大规模的语句级音视数据集。与 LRS2-TED 相比,它在时长、词汇量和说话人数量方面具有更大的规模。它包括来自 400 多个小时的 TED 和 TEDx 的说话人脸视频、相应的字幕和单词对齐边界。此外,它是现有公共可用的带标注的英语视听数据集中规模最大的。

VoxCeleb[24]是一个从开源 YouTube 媒体收集的与文本无关的大规模音视数据集。VoxCeleb1 包含来自 1 251 位名人的超过 10 万条话语实例。它主要用于说话人的识别,但也可用于视觉语音生成。VoxCeleb2 是 VoxCeleb1 的拓展版本,主要扩展了说话人种族多样性。在 VoxCeleb2 中,VoxCeleb1 数据集被重新用作说话人验证的测试集。此外,它是目前最大的公共可用音视数据集。

LRW-1000[25]是最大的词语级汉语普通话音视数据集。它包含 1 000 个单词类,包含来自 2 000多个说话者的 718 018 个音视样本。每一类对应一个由一个或几个汉字组成的普通话词语。所有视频均来自中国电视节目。

HDTF[41]是一个专为说话人脸生成而构建的大规模室外音视数据集。它由在线收集的大约 362 个不同的高分辨率视频组成。由于原始视频高质量,裁剪后的以人脸为中心的视频也具有比以前的数据集(如 LRW)更高的视觉质量。

2.2　评估指标

截至目前,对 VSG 方法性能进行有效评估仍然是一个悬而未决的问题,最近的许多工作从各种角度尝试探索合适的评估指标。本小节总结已有评估指标,根据身份保留、视觉质量、语义一致这三个学习目标对这些指标进行分类。

身份保留。VSG 最重要的目标之一是在视频生成过程中尽可能地保持目标人脸完整性,因为人类对合成视频中的细微外观变化非常敏感。但身份是一个语义上的感性概念,直接量化评估是不可行的。因此,为了评估生成的视频对目标身份的保留程度,现有工作通常使用生成的视频帧和真实目标人脸在特征空间的距离度量来衡量身份保持性能。例如,Vougioukas 等[42]基于 OpenFace[43]获得的人脸特征向量,采用平均内容距离(average content distance, ACD)来测量生成图像与真实图像的平均欧氏距离。此外,余弦相似度[44]常用来作为特征向量的距离测度。

视觉质量。为评估合成视频帧的视觉质量,重建误差度量(如均方误差)是一种最为直观的评估方式。然而,重建误差只关注像素对齐,而忽略了全局视觉质量。因此,现有的工作通常采用峰值信噪比(peak signal-to-noise ratio, PSNR)和结构相似性指数(structural similarity index measure, SSIM)来评估生成图像的全局视觉质量。Prajwal 等[45]引入了弗雷歇起始距离(Frechet inception distance, FID)来评估合成数据和真实数据分布之间的距离,因为 FID 更符合人类感知评估。此外,累积概率模糊检测(cumulative probability blur detection, CPBD)作为一种非参考图像质量度量指标,也被广泛用于评估生成视频的清晰度损失。

语义一致。生成的视频与驱动源在语义层次上的一致性主要涉及时域对齐和语音匹配。对于时间同步的性能评价,比较常见的评估指标是嘴唇区域标志点距离[46](landmark distance, LMD),LMD 表示合成视频和真实视频之间的唇部区域标志点的平均欧氏距离。另一种常用的同步评估指标是使用训练好的音视频同步网络[32]来预测生成视频与驱动源音频的时间偏移量。而对于语音一致性,Chen 等[27]提出了一种唇形同步评估度量,即唇读相似度距离(lip-reading similarity distance, LRSD)。它通过训练好的唇读网络获得语义级语音表征,再计算生成视频与真实视频表征的欧氏距离来进行评估。

除了上述客观量化的指标外,用户主观打分等主观指标也广泛用于 VSG 的性能评估。

3　视觉语音生成方法研究进展

视觉语音生成,也称为唇形序列生成,旨在合成与语音驱动源(一段音频或一段文本)对应的唇形序列。传统的 VSG 方法面临着严峻的实际挑战[26],例如复杂的生成管道、受限的适用环境、

过度依赖细粒度的视素(音素)标注等。为了实现将驱动源映射到唇部动态,代表性的传统 VSG 方法主要采用跨模态检索方法[47-48]和基于 HMM[49-50]的方法。例如,Thies 等[51]引入了一种典型的基于图像检索的唇动合成方法,该方法通过从离线样本中检索最匹配的嘴型来生成逼真的唇部动态。然而,基于检索的方法是静态的文本-音素-视素映射,并未真正考虑语音的上下文信息。同时,基于检索的方法对头部姿势的变化非常敏感。基于 HMM 的方法也存在一些缺点,例如先验假设的限制(如高斯混合模型及其对角协方差)。近期,深度学习技术广泛推动了 VSG 的发展,本节重点回顾基于深度学习的 VSG 方法。为了让读者更明确 VSG 问题的研究边界,在此先解释一下 VSG 与另一个热门研究问题,即说话人脸生成(talking face generation, TFG),的关系和区别[52]。TFG 旨在合成与驱动源和目标身份相对应的逼真、高质量的说话人脸视频。根据驱动源的类型区分,TFG 可分为音频驱动、文本驱动和视频驱动。其中,视频驱动的 TFG 主要侧重于面向视频的面部表情转换,而不是视觉语音生成。因此,视频驱动的 TFG 方法不在本文调研范围内。传统上,VSG 可以被视为音频驱动(文本驱动)TFG 的一个关键组件。另一个组件是视频编辑,遵循特定的编辑管道,根据生成的唇形序列输出最终合成的说话人脸视频。最近,为了减少人工干预并简化整个管道的复杂度,越来越多的研究人员尝试以端到端的方式合成说话人的完整面部,而不是唇部序列。因此,VSG 和音频驱动(文本驱动)TFG 之间的定义边界变得越来越模糊,这意味着一些音频驱动(文本驱动)TFG 方法也在我们的调研范围内。因此,为了对 VSG 进行全面综述,本文还回顾了一些由文本和音频驱动的 TFG 方法,因为这些工作隐含或显式涉及 VSG 模块。

3.1　总体流程

视觉语音生成总体流程如图 4 所示,给定目标人物的人脸参考(目标人脸图像或 3D 面部模型)和语音驱动源(一段音频或文本),VSG 的总体目标是生成唇(脸)动视频。由于现有的 VSG 方法的基本属性差异,例如驱动模式(文本驱动或音频驱动)、合成策略(基于计算机图形学或基于图像重构)、说话人泛化(说话人独立或说话人依赖)、学习范式(监督学习或无监督学习),对这些方法进行统筹综述并非易事。经过详细调研,本文根据生成框架的区别,将现有 VSG 方法区分成两个框架:①双阶段框架,包括两个映射步骤,即驱动源到面部参数和面部参数到视频;②单阶段框架,此类方法无须对中间面部参数建模,直接将驱动源映射到视频。本文分别在 3.2 节和 3.3 节中详细回顾和分析了当前的双阶段和单阶段 VSG 方法以及它们的优缺点。

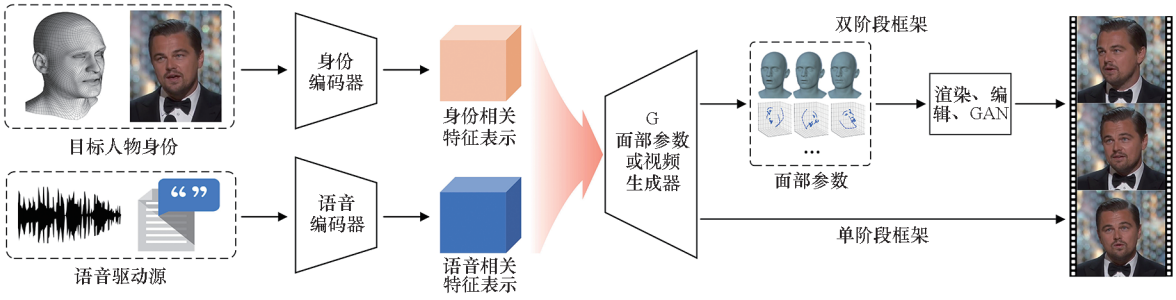


图 4　视觉语音生成总体流程

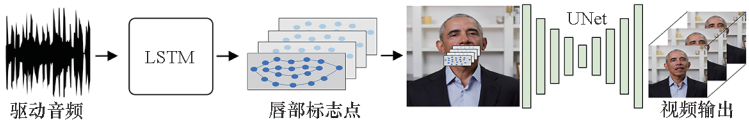
Fig. 4　Overall flowchart of visual speech generation

3.2　双阶段 VSG

基于深度学习的双阶段 VSG 框架主要包括两个步骤:①使用 DNNs 将驱动源映射到面部参数;②基于 GPU 渲染、视频编辑或生成对抗网络[53](generative adversarial networks, GANs)将学习到的面部参数转换为输出视频。根据面部参数类型的区别,现有的双阶段 VSG 方法可以分为基于面部标志点、基于面部系数以及基于面部轮廓顶点。图 5 为几种代表性的双阶段 VSG 方法。

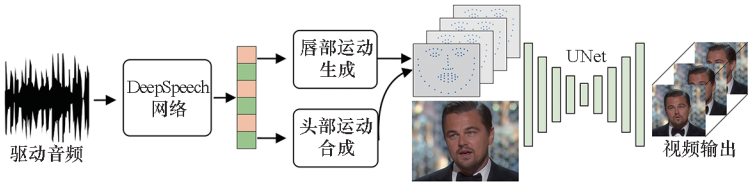
3.2.1　基于面部标志点的 VSG 方法

面部标志点(facial landmark points)是指由于头部运动和面部表情变化引起的刚性和非刚性面部形变[54]。面部标志点广泛用于各种面部分析任务,包括 VSG。作为深度 VSG 开创性的工作之一,Suwajanakorn 等[37]利用一个简单的单层长短期记忆(long short-term memory, LSTM)网络与时间延迟(time delay)机制来学习从音频梅尔频率倒谱系数(Mel frequency cepstral coefficient, MFCC)特征到嘴唇标志点的非线性映射。此后,该模型



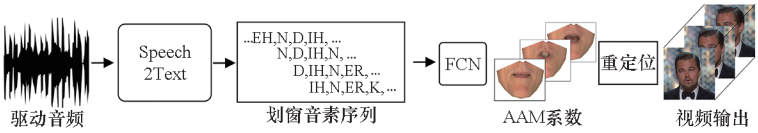
(a) 奥巴马网络[55]

(a) Obama Net[55]



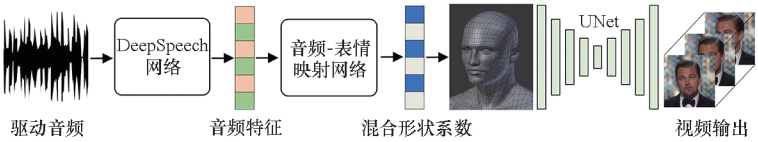
(b) 实时语音肖像[65]

(b) Live speech portraits[65]



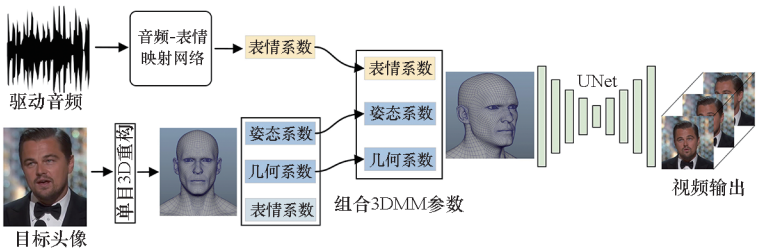
(c) 基于滑窗回归的视觉语音生成[67]

(c) Sliding window regression for VSG[67]



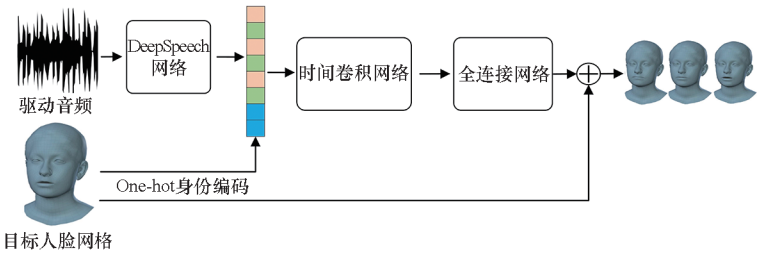
(d) 神经语音驱动视觉语音生成[68]

(d) Neural voice puppetry for VSG[68]



(e) 说话人独立视觉语音生成[69]

(e) Everybody′s talkin′ for VSG[69]



(f) 声控角色动画生成[39]

(f) Voice operated character animation generation[39]

图 5　代表性双阶段 VSG 方法

Fig. 5　Representative two-stage VSG method

经过面部纹理合成、视频重定时、目标视频合成等流程,生成美国前总统奥巴马的说话人脸视频。由于上述模型存在生成管线过于复杂、需要人工干预的问题,如图5(a)所示,Kumar 等[55]提出了LSTM + UNet 架构,将复杂的视频合成管线用pix2pix[56]框架来代替。这样,就不需要涉及具体面部细节的合成问题(如合成逼真的牙齿)。但是,由于上述两种方法仅针对奥巴马单一说话人进行训练,无法实现说话人的泛化。基于LSTM + UNet 的 VSG 主干架构在很多后续工作[10,40,57-58]中被广泛采用。与之前使用音频MFCC 特征作为输入的方法不同,Sinha 等[57-58]引入了 DeepSpeech[59]作为语音特征编码器,因为DeepSpeech 特征相较于 MFCC 特征来说对说话人的变化更具稳健性。

2018 年,Jalalifar 等[60]提出了 LSTM + C-GAN[61]的 VSG 骨干架构。由于 LSTM 网络和 C-GAN 网络是相互独立的,因此该模型可以用来自其他说话人的音频来驱动激活目标人物视觉语音生成。2019 年,Chen 等[62]提出了 LSTM + Convolutional-RNN 结构,进一步考虑了生成过程中相邻视频帧之间的相关性。此外,他们还提出了一种动态像素损失来解决视觉语音相关区域中的像素抖动问题。Wang 等[63]提出了一个三阶段的 VSG 框架。首先,他们使用 3D 沙漏网络作为运动场生成器,根据输入音频、头部运动和参考图像来预测标志点位置。然后将预测的标志点序列转换为密集的运动场模型。最后,使用一阶运动模型获得合成的说话视频。他们进一步更新了运动场发生器,将 3D 沙漏网络替换为自注意力架构[64]。

除了基于 2D 标志点的方法外,将语音驱动源映射到 3D 标志点也受到广泛关注和探索。音频信号包含大量语义级信息,包括语音内容、说话者的说话风格、情感等。Zhou 等[10]利用语音转换神经网络来学习解耦的语音内容和身份特征;引入基于 LSTM 的网络来根据语音内容特征预测 3D 标志点;使用 UNet 风格的生成器网络合成说话人脸视频。其主旨思想是从解耦合的音频内容特征和说话人感知特征中预测 3D 标志点,以便它们捕获可控的嘴唇同步和头部运动动态。如图5(b)所示,Lu 等[65]介绍了使用自回归预测编码[66](autoregressive predictive coding, APC)模型提取高级语音信息和流形投影以获得更好的泛化能力;设计了一个音频到嘴唇相关的运动模块来预测 3D 嘴唇标志点;引入了图像到图像转换 U 型网络(UNet)来合成视频帧。

3.2.2 基于面部系数的 VSG 方法

主动外观模型(active appearance model, AAM)是最常用的面部系数模型之一,它代表了人脸形状和纹理相关性及其变化。Fan 等[21]利用两层 BiLSTM 网络根据重叠的三音素输入估计嘴部区域的 AAM 系数,然后将其映射到人脸图像以产生逼真的说话视频。实验表明,BiLSTM 网络的性能要明显优于传统基于 HMM 的方法。类似地,如图5(c)所示,Taylor 等[67]引入了一种简单多层的 DNN 作为滑动窗口预测器,以基于固定长度音素序列作为输入来预测 AAM 系数。此外,借助有效的人脸重标定方法,该模型可以重新定位以驱动其他目标人脸模型。AAM 系数的主要实际限制是参考人脸 AAM 参数化在重标定到新对象时可能会导致潜在的错误。

除了2D 面部系数模型,通过主成分分析获得的3D 面部系数在 VSG 中更为常用[68-74]。例如,Pham 等[71,75]提出利用基于 CNN + RNN 的骨干架构将音频信号映射到 3D 人脸的 blendshape 系数[76]。然而,这些方法严重依赖于目标说话者的先验 3D 面部模型。Abdelaziz 等[72]将一个预训练的 DNN 的声学模型进行微调,实现将驱动音频映射到 3D blendshape 系数,因为他们认为预训练的声学模型在说话人独立的 VSG 任务上比随机初始化的模型具有更好的泛化能力。如图5(d)所示,Thies 等[68]为音频驱动的 VSG 提出了一个广义的 Audio2Expression 网络和一个专门的基于 UNet 的神经人脸渲染网络。所提出的 Audio2Expression 网络旨在使用基于 CNN 的主干架构和内容感知过滤网络,基于 DeepSpeech 音频特征来估计时间稳定的 3D blenshape 系数。通过这种方式,该模型能够将其他人的音频序列作为驱动源中合成说话视频。

除了 3D blendshape 系数,Kim 等[77-78]还引入了一种更密集的 3D 可变形模型[79](3D morphable model, 3DMM),用于面向视频的 face2face 转换。3DMM 系数包含刚性头部姿势参数、面部系数、表情系数、双眼注视方向系数和球面谐波照明系数。参考上面提到的基于 3DMM 的 face2face 转换管道,将驱动源从视频转换为音频片段(文本脚本),并将此管道迁移到 VSG 任务[4,41,73-74,69-81]。这些方法都有一个近似的框架,如图5(e)所示。该框架的流程图通常遵循四个步骤:①训练一个网络将驱动源映射到面部表情系数,因为视觉语音信息隐含在面部表情系数

中;②使用预训练的深度人脸重建模型获得目标人脸参考图像的 3DMM 系数;③结合目标参考图像的 3DMM 系数和预测的面部表情系数,得到混合 3DMM 系数;④使用 GPU 渲染或生成网络合成说话视频。

根据上述流程图,Song 等[69]设计了一个 Audio2Expression 网络,如图 5(e)所示。他们发现嵌入语音特征中的源身份信息会影响将语音映射到唇动的性能。因此,他们显式地添加了一个 ID-Removing 子网络,以便从驱动音频中去耦掉身份信息。同时,引入了一个 UNet 风格的生成网络,以生成由嘴部标志点的人脸图像序列生成。Yi 等[73]提出了一个基于 LSTM 的网络来将音频 MFCC 特征映射到面部表情和头部姿势,因为他们认为音频和头部姿势在短时间内是相关的。此外,他们提出了一种记忆增强 GAN 来将这些合成的视频帧进一步优化。Wu 等[80]提出了一种能够模仿任意说话风格的 VSG 方法。在映射阶段,他们引入了额外的说话风格参考视频作为输入,并使用深度 3D 重建模型来获取参考视频的风格编码。接下来,他们将音频特征与重构的风格编码连接起来,以预测风格化的 3DMM 系数。然而,上述基于 3DMM 的模型无法将视觉语音信息与眉毛和头部姿势等其他面部动作分离开。因此,Zhang 等[41]提出了一种流引导 VSG 框架,包括一个特定风格的动画生成器和一个流引导视频生成器,以合成高视觉质量的视频。此外,特定风格的动画生成器成功地将嘴唇动态与眉毛和头部姿势分开。Li 等[82]为文本驱动的 VSG 采用了类似的框架。Ji 等[4]提出了一种情感视频人像(emotional video portraits,EVP),以实现语音驱动的情感控制,用于说话人脸合成。与之前的方法不同,他们在 audio2expression 阶段采用了跨域重构[83]技术,将输入音频分解为内容表征和情感表征。

3.2.3　基于面部轮廓顶点的 VSG 方法

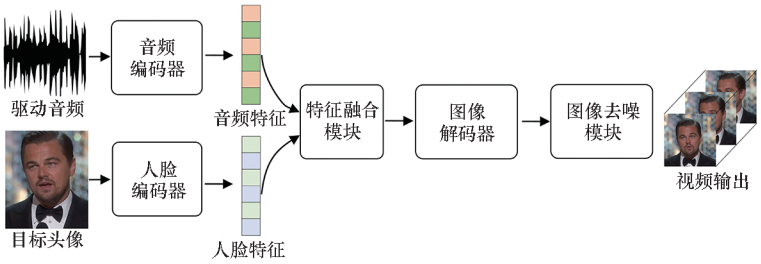
3D 面部轮廓顶点是 VSG 中另一个常用的 3D 面部模型。例如,Karras 等[5]使用简单的基于卷积神经网络的架构来学习从输入音频到目标人脸的 3D 顶点坐标(总共 15 066 个顶点)的非线性映射。为了使合成的视频更加自然,他们引入了额外的情绪编码作为对面部情绪状态的直观控制因子。然而,所提出的模型是专门针对特定说话人的。为了克服这个问题,如图 5 (f)所示,Cudeiro 等[39]将模型扩展到多个说话人。所提出的语音控制角色动画(voice operated character animation,VOCA)模型将 DeepSpeech 音频特征和说话人的 one-hot 向量连接起来,并输出 3D 顶点(总共 5 023 个顶点)位移量而非顶点坐标。VOCA 的关键贡献是附加的身份控制参数可以控制改变身份相关的视觉动态。Liu 等[84]基于 VOCA 提出了一个几何引导的密集透视网络(geometry-guided dense perspective network,GDPnet),具有来自不同视角的两个约束,以实现更鲁棒的人脸序列生成。Fan 等[85]提出了一种名为 FaceFormer 的基于 Transformer 的自回归 VSG 模型,用于对长期音频上下文信息进行编码并预测一系列 3D 人脸顶点。

Richard 等[86]为 VSG 提出了一个分类潜在空间,它基于跨模态损失来分解音频相关和音频不相关(眨眼、眉毛等面部表情)信息。然后,使用带有跳跃连接的 UNet 风格的网络架构来预测 3D 轮廓顶点坐标。由于模态解耦机制,面部不相关区域的合理运动是可控的,使合成的视频更加逼真。Lahiri 等[70]提出了一种说话人相关的 VSR 方法,它将音频到说话的人脸映射问题分解为 3D 人脸形状的预测和 2D 纹理图集的回归。为此,他们首先引入了标准化预处理阶段,以消除头部运动和光照变化的影响。然后,对几何解码器和自回归纹理合成网络进行训练,以分别学习顶点位移和相应的以唇为中心的图像纹理。最后,使用基于图形的视频渲染生成目标说话视频。

3.3　单阶段 VSG

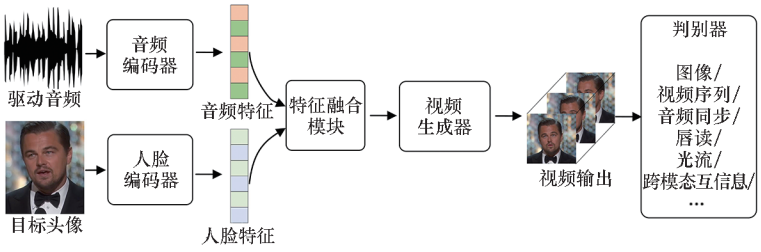
双阶段 VSG 框架在 2018 年之前一直占据主导地位。然而,双阶段 VSG 框架面临着复杂的处理流程、昂贵且耗时的面部参数建模、需要面部标志点检测和单目 3D 人脸重建等额外辅助技术等问题。因此,研究人员近期更加关注于探索单阶段(端到端)VSG 方法。单阶段 VSG 框架是一种不涉及任何人脸中间面部参数的端到端学习策略,直接将驱动源映射为唇动(面部)视频。

图 6 为几种代表性的单阶段 VSG 方法。Speech2Vid[52]是最早探索单阶段 VSG 框架的工作之一。如图 6(a)所示,它由四个子网络组成。其中,音频编码器旨在根据驱动音频提取语音特征;人脸编码器旨在根据目标参考图像提取身份特征;图像解码器融合语音和身份特征并合成图像序列。上述三个子网络形成自编码器架构,使用 L1 重建损失进行训练。此外,还引入了一个单独的预训练去模糊 CNN 作为后处理模块以提高图像质量。作为一项开创性工作,Speech2Vid 为说话人独立的 VSG 提供了研究基准,极大地推动了单阶段 VSG 的研究。



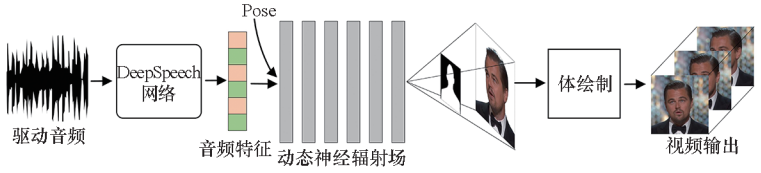
(a) 语音转视频[52]

(a) Speech2Vid[52]



(b) 基于 GAN 的单阶段视觉语音生成[87]

(b) GAN based one-stage VSG[87]



(c) 音频驱动神经辐射场视觉语音生成[88]

(c) Audio driven neural radiance fields for VSG[88]

(d) 基于动态卷积核的视觉语音生成[89]

(d) VSG with dynamic convolution kernels[89]

图 6　代表性的单阶段 VSG 方法

Fig. 6　Representative one-stage VSG methods

然而,Speech2Vid 在训练期间只使用 L1 重构损失,这对于 VSG 来说效率不高,有以下几点原因:①L1 重建损失是在整个人脸上进行的,人脸的自发运动主要发生在人脸的上半部分,这导致视觉语音生成被弱化;②由于 Speech2Vid 在各时刻的生成是时序独立的,它通常会产生不太连贯的视频序列;③未充分考虑生成的视频与驱动音频之间的一致性需求。

3.3.1　基于 GAN 的单阶段 VSG 方法

为了克服 Speech2Vid 的诸多限制,许多研究人员试图通过利用生成对抗训练策略来提高 VSG 性能。如图 6(b)所示,基于 GAN 的单阶段 VSG 方法通常由三个子网络模块组成,即编码器模块、生成器模块和判别器模块。

以音频驱动的 VSG 为例,一段音频自然会耦合各种信息,如语音、情感、说话风格等。正如我们在 1.2 节中所强调的,信息耦合给 VSG 带来了巨大的挑战。为了改善这个问题,Zhou 等[87]提出了一种新的 VSG 框架,称为分离式视听系统(disentangled audio-visual system, DAVS)。与之前的 VSG 方法相比,他们更关注解耦的语音和身份特征提取,这是基于有监督的对抗训练实现的。但是,DAVS 在训练阶段依赖于额外的单词标签和说话人身份标签。Sun 等[90]通过在自监督的对比学习框架内学习语音和身份特征来改进模型,解决了需要额外标注的问题。Si 等[91]

在预训练的情感识别教师网络和预训练的人脸识别教师网络的帮助下,利用知识蒸馏从音频输入中分离出情感特征、身份特征和语音特征。

最近,一些工作试图将额外的面部可控动力学(如情绪和头部姿势)编码到生成管道中,以生成更自然自发的说话面部。例如,文献[92 - 93]引入了额外的情感编码器,文献[94]将隐式头部姿态编码设计到生成管道中。

考虑到仅使用图像重建损失的缺点,基于 GAN 的方法专注于为 VSG 定制更有效的判别器来强化 VSG 学习目标。例如,Prajwal 等[45,95]引入了一个简单的音视频同步判别器,用于强化生成视频与驱动源的同步性。此外,Chen 等[46]提出了一种音视导数相关损失来优化特征空间中两种模态的一致性,并提出三流融合 GAN 判别器来强化生成的说话视频与音频信号的相关性。

对于时间相关的视频生成,文献[42,96 - 97]利用自回归风格的 VSG 生成器网络来生成说话人脸。使用帧判别器和序列判别器,在两个尺度上优化生成的面部动态。基于文献[42],Song 等[98]进一步引入了 VSR 判别器,以提高生成的说话视频的唇部运动精度。其消融研究表明,额外的 VSR 判别器有助于实现更明显的唇部运动生成,这也证明了 VSR 和 VSG 是对偶和相互促进的。此外,Chen 等[99]开发了 DualLip 系统,通过利用任务对偶性来共同提升 VSR 和 VSG 性能,并证明 VSR 和 VSG 模型都可以在额外未标记数据的帮助下得到增强。

除上述判别器外,光流判别器[100]、面部动作单元判别器[101]和跨模态互信息估计器[102]也被用于优化生成的视频与驱动源的跨模态一致性。

3.3.2　其他单阶段 VSG 方法

此外,近几年出现了一些其他的单阶段 VSG 方法。例如,受神经辐射场[103](neural radiance field, NeRF)的启发,Guo 等[88]提出了 VSG 的音频驱动神经辐射场(audio driven neural radiance fields, AD-NeRF)模型。如图 6(c)所示,AD-NeRF 以 DeepSpeech 音频特征作为条件输入,学习隐式神经场景表示函数将音频特征映射到动态神经辐射场以进行说话人脸渲染。此外,AD-NeRF 通过学习两个单独的神经辐射场,同时对头部区域和说话人上半身进行建模。然而,AD-NeRF 不能很好地适应不匹配的驱动音频和说话人。如图 6(d)所示,与之前基于特征聚合的特征融合策略不同,Ye 等[89]提出了带有动态卷积核(dynamic convolution kernels, DCKs)的全卷积神经网络用于跨模态特征融合,从音频中提取特征并将特征排列为全卷积网络的动态卷积核。这种方法的网络构架简单而有效,实时性能显著提高。

3.4　方法总结及性能对比

本小节将讨论大规模数据集上的代表性 VSG,并总结当前 VSG 方法的主要问题。因为 VSG 方法有不同的实现要求(驱动源、额外的技术、不同的标注需求、特定的数据集等)和配置(训练集、学习范式、嘴唇或全脸生成、背景、姿势和情绪控制等),以完全统一和公平的方式比较 VSG 方法是不切实际的。

尽管如此,将一些具有代表性的 VSG 方法及其要求、配置整合到一个表格中还是很有价值的。因此,如表 2 所示,我们总结了一些代表性 VSG 方法在常用基准数据集 LRW 上测试的性能和实验设置。

为了让读者对 VSG 方法在不同框架下的性能有一个大致的了解,表中列出了三种常用的定量评价指标,即 PSNR、SSIM 和 LMD。值得注意的是,尽管上述三个指标在 VSG 中使用最为广泛,但它们还不够有效和完善。虽然近几年提出了许多专门针对 VSG 的定量评价指标,但以下问题需要进一步研究。

1)在 VSG 研究的早期,主要使用定性评估,例如可视化和用户偏好研究。然而,定性评估不稳定且不可重复。

2)许多工作都试图建立公认定量 VSG 评估基准,导致现有的 VSG 评估基准非常分散。Chen 等[27]试图对 VSG 的评价指标进行整合,并根据所需的属性为 VSG 设计了统一的基准。为了促进 VSG 的发展,研究人员应该在 VSG 评估基准上付出更多的努力。

3)定量评价和定性评价的结果有时会相互冲突。例如,一些工作[63,96,98]发现 PSNR 和 SSIM 都受到引入的图像或视频判别器的负面影响。然而,这些判别器的引入明显提高了用户对视频真实感和视觉质量的评价。

4)在实际应用中,实时性是对 VSG 的另一个实质性要求。然而,目前大多数 VSG 方法都忽略了实时性。因此,实时性能也是未来需要考虑的重要评价指标。

表2 代表性VSG方法性能对比

Tab. 2 Comparison of some representative VSG methods

框架 & 方法		实验设置			LRW				参考文献
		输入	训练集	额外需求	副产品	PSNR/dB	SSIM	LMD	
双阶段VSG	基于标志点	A + I	LRW/GRID	标志点检测		30.91	0.81	1.37	[62]
		A + I	LRW/GRID/VoxCeleb	标志点检测 图像生成网络 人脸识别网络	头部姿态可控	19.53	0.63	—	[63]
	基于系数	A + I	LRW	人脸重构网络	头部姿态可控	30.94	0.75	1.58	[73]
单阶段VSG	基于自编码器	A + I	LRW/VoxCeleb	人脸识别网络		28.06	0.46	2.25	[52]
	基于GAN	A + I	GRID	VSR网络		23.08	0.76	—	[96]
		A + I	LRW/GRID	FlowNet网络		28.65	0.53	1.92	[46]
		A + V	LRS2			33.40	0.96	0.60	[95]
		A + I	TCD-TIMIT LRW	VSR网络		27.43	0.92	3.14	[98]
		A + V	VoxCeleb LRW	单词/身份标签		26.70	0.88	—	[87]
		A + I	LRW/GRID			32.08	0.92	1.21	[102]
	其他	A + V	自建数据集	音频编码网络		31.98	0.81	1.44	[89]

4 总结与展望

本文对基于深度学习的视觉语音生成方法进行了全面回顾。总结了现实挑战和当前发展现状,包括数据集、评估指标、代表性方法、最佳性能、实际问题等。我们对VSG方法进行了系统概述,并讨论了它们的潜在联系、贡献和缺点。考虑到许多实际问题仍未解决,VSG的研究和应用仍有足够的机会。因此,本文在此提供一些想法并讨论潜在的未来研究方向。

1)多语种VSG。现有的音视数据集大多是单语种的。一般来说,英语是最通用也研究最多的语言。但在一些实际场景中,如空中交通管制和国际会议,需要多语种交流。尽管多语种音频语音识别已经得到广泛的探索,但多语种视觉语音生成却很少受到关注。

2)视觉语音识别与生成的协同。识别与生成是视觉语音分析领域两个最基本的问题,二者形式对偶且互相促进。基于这种特性,现有少量工作采用对偶学习或生成对抗训练来促进二者共同提高。当前,视觉语音表示跨模态互学习算法可做到同时适配于这两个任务,但相关工作不多。后续研究应基于现有跨模态学习算法进一步探究面向多任务的协同式视觉语音分析相关方法。

3)VSG的实际应用拓展。虽然近几年视觉语音分析相关问题的进展显著,但无论是识别还是生成,相关方法效果及适用范围还非常有限,距离实际应用部署还有较大差距。除VSG以外,视觉语音生成与识别能够为很多其他的学术或实际应用服务,比如鸡尾酒会问题(语音分离)、Deepfake检测、活体检测等。后续的工作应考虑基于现有算法将SVG延伸至更多有价值有前景的应用领域中。

4)虚拟角色的VSG技术。该研究方向主要来源于虚拟现实技术发展以及元宇宙平台建设的需求。元宇宙作为一种新兴的互联网应用和社交平台,近来备受关注。虚拟人物建模是元宇宙的一项关键技术。随着元宇宙技术的飞速发展,面向虚拟角色的VSG技术也应运而生。然而,现有的VSG方法大多侧重于现实说话者,当前面向虚拟角色的VSG研究的相关工作还很有限。与计算机图形学技术相结合的VSG方法是未来一个非常有潜力的研究方向。

5)隐私保护视觉语音分析技术。由于VSG涉及人脸相关的隐私信息,这导致难以构建公开可用的大规模音视数据集,也阻碍了VSG方法的进一步发展。为了解决这个问题,可用的隐私保护技术(例如联邦学习、同态加密以及安全多方计算等)可能会有所帮助。然而,据我们所知,隐

私保护的 VSG 研究尚未真正起步,值得未来进一步探索和突破。

参考文献(References)

[1] CHEN T. Audiovisual speech processing[J]. IEEE Signal Processing Magazine, 2001, 18(1): 9 – 21.

[2] MCGURK H, MACDONALD J. Hearing lips and seeing voices[J]. Nature, 1976, 264(5588): 746 – 748.

[3] JIA Y, ZHANG Y, WEISS R J, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018: 4485 – 4495.

[4] JI X Y, ZHOU H, WANG K, et al. Audio-driven emotional video portraits[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 14075 – 14084.

[5] KARRAS T, AILA T M, 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): 1 – 12.

[6] HALIASSOS A, VOUGIOUKAS K, PETRIDIS S, et al. Lips don't lie: a generalisable and robust approach to face forgery detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 5037 – 5047.

[7] AKHTAR Z, MICHELONI C, FORESTI G L. Biometric liveness detection: challenges and research opportunities[J]. IEEE Security and Privacy, 2015, 13(5): 63 – 72.

[8] REKIK A, BEN-HAMADOU A, MAHDI W. Human machine interaction via visual speech spotting[C]// Proceedings of International Conference on Advanced Concepts for Intelligent Vision Systems, 2015: 566 – 574.

[9] SUN K, YU C, SHI W N, et al. Lip-interact: improving mobile device interaction with silent speech commands[C]// Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology, 2018: 581 – 593.

[10] ZHOU Y, HAN X T, SHECHTMAN E, et al. MakeItTalk: speaker-aware talking-head animation[J]. ACM Transactions on Graphics, 2020, 39(6): 1 – 15.

[11] GARRIDO P, VALGAERTS L, SARMADI H, et al. VDub: modifying face video of actors for plausible visual alignment to a dubbed audio track[J]. Computer Graphics Forum, 2015, 34(2): 193 – 204.

[12] CAPPELLETTA L, HARTE N. Viseme definitions comparison for visual-only speech recognition[C]// Proceedings of 19th European Signal Processing Conference, 2011.

[13] POMIANOS G, NETI C, GRAVIER G, et al. Recent advances in the automatic recognition of audiovisual speech[J]. Proceedings of the IEEE, 2003, 91(9): 1306 – 1326.

[14] KIRCHHO K. Robust speech recognition using articulatory information[D]. Bielefeld: Bielefeld University, 1999.

[15] POTAMIANOS G, GRAF H P, COSATTO E. An image transform approach for HMM based automatic lipreading[C]// Proceedings of International Conference on Image Processing, 1998: 173 – 177.

[16] MATTHEWS I, COOTES T F, BANGHAM J A, et al. Extraction of visual features for lipreading[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(2): 198 – 213.

[17] DEENA S, HOU S B, GALATA A. Visual speech synthesis by modelling coarticulation dynamics using a non-parametric switching state-space model[C]//Proceedings of International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, 2010: 1 – 8.

[18] ANDERSON R, STENGER B, WAN V, et al. Expressive visual text-to-speech using active appearance models[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013: 3382 – 3389.

[19] KIM T, YUE Y S, TAYLOR S, et al. A decision tree framework for spatiotemporal sequence prediction[C]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015: 577 – 586.

[20] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84 – 90.

[21] FAN B, WANG L J, SOONG F K, et al. Photo-real talking head with deep bidirectional LSTM[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015: 4884 – 4888.

[22] CHUNG J S, ZISSERMAN A. Lip reading in the wild[C]// Proceedings of the Asian Conference on Computer Vision, 2016: 87 – 103.

[23] CHUNG J S, SENIOR A, VINYALS O, et al. Lip reading sentences in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 3444 – 3453.

[24] CHUNG J S, NAGRANI A, ZISSERMAN A. VoxCeleb2: deep speaker recognition[EB/OL]. (2018 – 06 – 18)[2022 – 05 – 20]. https://arxiv.org/pdf/1806.05622.pdf.

[25] YANG S, ZHANG Y H, FENG D L, et al. LRW-1000: a naturally-distributed large-scale benchmark for lip reading in the wild[C]//Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition, 2019: 1 – 8.

[26] MATTHEYSES W, VERHELST W. Audiovisual speech synthesis: an overview of the state-of-the-art[J]. Speech Communication, 2015, 66: 182 – 217.

[27] CHEN L L, CUI G F, KOU Z Y, et al. What comprises a good talking-head video generation?: a survey and benchmark[EB/OL]. (2020 – 05 – 07)[2022 – 05 – 20]. https://arxiv.org/pdf/2005.03201v1.pdf.

[28] 陈小鼎, 盛常冲, 匡纲要, 等. 唇读研究进展与展望[J]. 自动化学报, 2020, 46(11): 2275 – 2301.
CHEN X D, SHENG C C, KUANG G Y, et al. The state of the art and prospects of lip reading[J]. Acta Automatica Sinica, 2020, 46(11): 2275 – 2301. (in Chinese)

[29] MORI M, MACDORMAN K F, KAGEKI N. The uncanny valley from the field[J]. IEEE Robotics & Automation Magazine, 2012, 19(2): 98 – 100.

[30] CHUNG S W, KANG H G, CHUNG J S. Seeing voices and hearing voices: learning discriminative embeddings using cross-modal self-supervision[EB/OL]. (2020 – 04 – 29)[2022 – 05 – 20]. https://arxiv.org/pdf/2004.14326.pdf.

[31] SHENG C C, PIETIKÄINEN M, TIAN Q, et al. Cross-modal self-supervised learning for lip reading: when contrastive learning meets adversarial training[C]//Proceedings of the

29th ACM International Conference on Multimedia, 2021: 2456 – 2464.

[32] CHUNG S W, CHUNG J S, KANG H G. Perfect match: self-supervised embeddings for cross-modal retrieval[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(3): 568 – 576.

[33] 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 Pt 1): 2421 – 2424.

[34] CZYZEWSKI A, KOSTEK B, BRATOSZEWSKI P, et al. An audio-visual corpus for multimodal automatic speech recognition[J]. Journal of Intelligent Information Systems, 2017, 49(2): 167 – 192.

[35] ANINA I, ZHOU Z H, ZHAO G Y, et al. OuluVS2: a multi-view audiovisual database for non-rigid mouth motion analysis[C]// Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015: 1 – 5.

[36] AFOURAS T, CHUNG J S, SENIOR A, et al. Deep audio-visual speech recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12): 8717 – 8727.

[37] SUWAJANAKORN S, SEITZ S M, KEMELMACHER-SHLIZERMAN I. Synthesizing Obama: learning lip sync from audio[J]. ACM Transactions on Graphics, 2017, 36(4): 1 – 13.

[38] AFOURAS T, CHUNG J S, ZISSERMAN A. LRS3-TED: a large-scale dataset for visual speech recognition [EB/OL]. (2018 – 09 – 03)[2022 – 05 – 20]. https://arxiv.org/pdf/1809.00496.pdf.

[39] CUDEIRO D, BOLKART T, LAIDLAW C, et al. Capture, learning, and synthesis of 3D speaking styles [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 10093 – 10103.

[40] WANG K, WU Q Y, SONG L S, et al. MEAD: a large-scale audio-visual dataset for emotional talking-face generation[C]//Proceedings of the European Conference on Computer Vision, 2020: 700 – 717.

[41] ZHANG Z M, LI L C, DING Y. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 3660 – 3669.

[42] VOUGIOUKAS K, PETRIDIS S, PANTIC M. End-to-end speech-driven facial animation with temporal GANs [EB/OL]. (2020 – 05 – 23)[2022 – 05 – 20]. https://arxiv.org/pdf/1805.09313.pdf.

[43] AMOS B, LUDWICZUK B, SATYANARAYANAN M. OpenFace: a general-purpose face recognition library with mobile applications[EB/OL]. (2016 – 01 – 01)[2022 – 05 – 20]. http://reports-archive.adm.cs.cmu.edu/anon/anon/2016/CMU-CS-16-118.pdf.

[44] ZAKHAROV E, SHYSHEYA A, BURKOV E, et al. Few-shot adversarial learning of realistic neural talking head models[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 9458 – 9467.

[45] 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, 2020: 484 – 492.

[46] CHEN L L, LI Z H, MADDOX R K, et al. Lip movements generation at a glance [C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 520 – 535.

[47] BREGLER C, COVELL M, SLANEY M. Video rewrite: driving visual speech with audio[C]//Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, 1997: 353 – 360.

[48] GARRIDO P, VALGAERTS L, REHMSEN O, et al. Automatic face reenactment[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 4217 – 4224.

[49] FU S L, GUTIERREZ-OSUNA R, ESPOSITO A, et al. Audio/visual mapping with cross-modal hidden Markov models[J]. IEEE Transactions on Multimedia, 2005, 7(2): 243 – 252.

[50] XIE L, LIU Z Q. Realistic mouth-synching for speech-driven talking face using articulatory modelling [J]. IEEE Transactions on Multimedia, 2007, 9(3): 500 – 510.

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

[52] JAMALUDIN A, CHUNG J S, ZISSERMAN A. You said that?: synthesising talking faces from audio[J]. International Journal of Computer Vision, 2019, 127(11/12): 1767 – 1779.

[53] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014: 2672 – 2680.

[54] WU Y, JI Q. Facial landmark detection: a literature survey[J]. International Journal of Computer Vision, 2019, 127(2): 115 – 142.

[55] KUMAR R, SOTELO J, KUMAR K, et al. ObamaNet: photo-realistic lip-sync from text[C]// Proceedings of the 31st Conference on Neural Information Processing Systems, 2017.

[56] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 5967 – 5976.

[57] SINHA S, BISWAS S, BHOWMICK B. Identity-preserving realistic talking face generation [C]//Proceedings of the International Joint Conference on Neural Networks (IJCNN), 2020: 1 – 10.

[58] DAS D, BISWAS S, SINHA S, et al. Speech-driven facial animation using cascaded GANs for learning of motion and texture[C]// Proceedings of the European Conference on Computer Vision, 2020: 408 – 424.

[59] HANNUN A Y, CASE C, CASPER J, et al. Deep speech: scaling up end-to-end speech recognition[EB/OL]. (2014 – 12 – 17)[2022 – 05 – 20]. https://arxiv.org/pdf/1412.5567.pdf.

[60] JALALIFAR S, HASANI H, AGHAJAN H. Speech-driven facial reenactment using conditional generative adversarial networks[EB/OL]. (2018 – 03 – 20)[2022 – 05 – 20]. https://arxiv.org/pdf/1803.07461.pdf.

[61] MIRZA M, OSINDERO S. Conditional generative adversarial nets[EB/OL]. (2014 - 11 - 06)[2022 - 05 - 20]. https://arxiv.org/pdf/1411.1784.pdf.

[62] CHEN L L, MADDOX R K, DUAN Z Y, et al. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 7824 - 7833.

[63] WANG S Z, LI L C, DING Y, et al. Audio2Head: audio-driven one-shot talking-head generation with natural head motion[EB/OL]. (2020 - 07 - 20)[2022 - 05 - 20]. https://arxiv.org/pdf/2107.09293.pdf.

[64] WANG S Z, LI L C, DING Y Q, et al. One-shot talking face generation from single-speaker audio-visual correlation learning[EB/OL]. (2021 - 12 - 06)[2022 - 05 - 20]. https://arxiv.org/pdf/2107.09293.pdf.

[65] LU Y X, CHAI J X, CAO X. Live speech portraits: real-time photorealistic talking-head animation[J]. ACM Transactions on Graphics, 2021, 40(6): 1 - 17.

[66] CHUNG Y A, GLASS J. Generative pre-training for speech with autoregressive predictive coding[C]//Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: 3497 - 3501.

[67] TAYLOR S, KIM T, YUE Y S, et al. A deep learning approach for generalized speech animation[J]. ACM Transactions on Graphics, 36(4): 1 - 11.

[68] THIES J, ELGHARIB M, TEWARI A, et al. Neural voice puppetry: audio-driven facial reenactment[C]//European Conference on Computer Vision, 2020: 716 - 731.

[69] SONG L S, WU W, QIAN C, et al. Everybody's talkin': let me talk as you want[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 585 - 598.

[70] LAHIRI A, KWATRA V, FRUEH C, et al. LipSync3D: data-efficient learning of personalized 3D talking faces from video using pose and lighting normalization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 2754 - 2763.

[71] PHAM H X, WANG Y T, PAVLOVIC V. End-to-end learning for 3D facial animation from speech[C]// Proceedings of the 20th ACM International Conference on Multimodal Interaction, 2018: 361 - 365.

[72] ABDELAZIZ A H, THEOBALD B J, BINDER J, et al. Speaker-independent speech-driven visual speech synthesis using domain-adapted acoustic models[C]//Proceedings of the 2019 International Conference on Multimodal Interaction, 2019: 220 - 225.

[73] YI R, YE Z P, ZHANG J Y, et al. Audio-driven talking face video generation with learning-based personalized head pose[EB/OL]. (2018 - 02 - 24)[2022 - 05 - 20]. https://arxiv.org/pdf/2002.10137.pdf.

[74] YAO X W, FRIED O, FATAHALIAN K, et al. Iterative text-based editing of talking-heads using neural retargeting[J]. ACM Transactions on Graphics, 40(3): 1 - 14.

[75] PHAM H X, CHEUNG S, PAVLOVIC V. Speech-driven 3D facial animation with implicit emotional awareness: a deep learning approach[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017: 2328 - 2336.

[76] CAO C, WENG Y L, ZHOU S, et al. FaceWarehouse: a 3D facial expression database for visual computing[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(3): 413 - 425.

[77] KIM H, GARRIDO P, TEWARI A, et al. Deep video portraits[J]. ACM Transactions on Graphics, 37(4): 1 - 14.

[78] KIM H, ELGHARIB M, ZOLLHÖFER M, et al. Neural style-preserving visual dubbing[J]. ACM Transactions on Graphics, 38(6): 1 - 13.

[79] DENG Y, YANG J L, XU S C, et al. Accurate 3D face reconstruction with weakly-supervised learning: from single image to image set[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019: 285 - 295.

[80] WU H Z, JIA J, WANG H Y, et al. Imitating arbitrary talking style for realistic audio-driven talking face synthesis[C]//Proceedings of the 29th ACM International Conference on Multimedia, 2021: 1478 - 1486.

[81] ZHANG C X, ZHAO Y F, HUANG Y F, et al. FACIAL: synthesizing dynamic talking face with implicit attribute learning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 3847 - 3856.

[82] LI L C, WANG S Z, ZHANG Z M, et al. Write-a-speaker: text-based emotional and rhythmic talking-head generation[C]// Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, 2021: 1911 - 1920.

[83] ABERMAN K, WU R D, LISCHINSKI D, et al. Learning character-agnostic motion for motion retargeting in 2D[J]. ACM Transactions on Graphics, 2019, 38(4): 1 - 14.

[84] LIU J Y, HUI B Y, LI K, et al. Geometry-guided dense perspective network for speech-driven facial animation[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(12): 4873 - 4886.

[85] FAN Y R, LIN Z J, SAITO J, et al. FaceFormer: speech-driven 3D facial animation with transformers[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 18749 - 18758.

[86] RICHARD A, ZOLLHÖFER M, WEN Y D, et al. MeshTalk: 3D face animation from speech using cross-modality disentanglement[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 1153 - 1162.

[87] ZHOU H, LIU Y, LIU Z W, et al. Talking face generation by adversarially disentangled audio-visual representation[C]// Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, 2019: 9299 - 9306.

[88] GUO Y D, CHEN K Y, LIANG S, et al. AD-NeRF: audio driven neural radiance fields for talking head synthesis[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 5764 - 5774.

[89] YE Z P, XIA M F, YI R, et al. Audio-driven talking face video generation with dynamic convolution kernels[J]. IEEE Transactions on Multimedia, 2023, 25: 2033 - 2046.

[90] SUN Y S, ZHOU H, LIU Z W, et al. Speech2Talking-Face: inferring and driving a face with synchronized audio-visual representation[C]//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 2021: 1018 -

1024.

[91] SI S J, WANG J Z, QU X Y, et al. Speech2Video: cross-modal distillation for speech to video generation[EB/OL]. (2021-07-10)[2022-05-20]. https://arxiv.org/pdf/2107.04806.pdf.

[92] SADOUGHI N, BUSSO C. Speech-driven expressive talking lips with conditional sequential generative adversarial networks[J]. IEEE Transactions on Affective Computing, 2021, 12(4): 1031-1044.

[93] ESKIMEZ S E, ZHANG Y, DUAN Z Y. Speech driven talking face generation from a single image and an emotion condition[J]. IEEE Transactions on Multimedia, 2021, 24: 3480-3490.

[94] ZHOU H, SUN Y S, WU W, et al. Pose-controllable talking face generation by implicitly modularized audio-visual representation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 4174-4184.

[95] PRAJWAL K R, MUKHOPADHYAY R, PHILIP J, et al. Towards automatic face-to-face translation[C]//Proceedings of the 27th ACM International Conference on Multimedia, 2019: 1428-1436.

[96] VOUGIOUKAS K, PETRIDIS S, PANTIC M. Realistic speech-driven facial animation with GANs[J]. International Journal of Computer Vision, 2020, 128(5): 1398-1413.

[97] ESKIMEZ S E, MADDOX R K, XU C L, et al. End-to-end generation of talking faces from noisy speech [C]// Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: 1948-1952.

[98] SONG Y, ZHU J W, LI D W, et al. Talking face generation by conditional recurrent adversarial network[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019: 919-925.

[99] CHEN W C, TAN X, XIA Y C, et al. DualLip: a system for joint lip reading and generation[C]//Proceedings of the 28th ACM International Conference on Multimedia, 2020: 1985-1993.

[100] ZENG D, LIU H, LIN H, et al. Talking face generation with expression-tailored generative adversarial network [C]// Proceedings of the 28th ACM International Conference on Multimedia, 2020: 1716-1724.

[101] CHEN S, LIU Z L, LIU J X, et al. Talking head generation with audio and speech related facial action units[EB/OL]. (2021-10-19)[2022-05-20]. https://arxiv.org/pdf/2110.09951.pdf.

[102] ZHU H, HUANG H B, LI Y, et al. Arbitrary talking face generation via attentional audio-visual coherence learning[C]// Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, 2021: 2362-2368.

[103] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. NeRF: representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2020, 65(1): 99-106.

(编辑：梁慧，杨琴)