基于语义增强与纹理-运动融合的说话人无关视觉配音方法

陈燚雷[1,3]，熊盛武[2*]

（1. 湖北经济学院数字金融创新湖北省重点实验室，湖北武汉 430205；2. 武汉理工大学计算机与人工智能学院，湖北武汉 430070；
3. 湖北经济学院信息工程学院，湖北武汉 430205）

摘 要： 说话人无关的视觉配音技术旨在通过语音信号驱动说话人脸视频中唇部区域的运动，实现音视频的高度同步与自然融合. 该技术不仅要求编辑后的视频具备良好的语音-视频同步性，还需保持面部纹理与身份特征的一致性. 然而，现有方法在处理存在自然头部运动的视频时，常出现修复区域与真实人脸区域纹理不一致的问题，导致生成质量下降. 为解决上述难题，本文提出了一种跨模态语义增强与3D人脸引导的运动纹理协同生成网络. 该方法以三维可变形人脸模型（3D Morphable Model，3DMM）作为中间表示，将任务分解为语音驱动的3D表情系数预测与运动-纹理协同的人脸渲染两个子任务. 首先，设计了跨模态语义增强的3DMM表情系数预测网络，通过引入Wav2Lip生成的语义图像序列与局部跨模态注意力机制，显著提升了语音-视频的同步率与几何一致性. 其次，提出3D人脸引导的运动纹理协同渲染网络，利用多参考人脸与3D重建人脸进行纹理补偿与细节增强，并构建多任务学习框架以保证修复区域与真实人脸的纹理一致性. 在VoxCeleb1和VoxCeleb2数据集上的大量实验表明，本文所提方法在生成保真度、运动鲁棒性和同步性方面均优于现有代表性方法. 与基线模型相比，本方法在VoxCeleb1数据集上实现了峰值信噪比（Peak Signal Noise Ratio，PSNR）提升7.76，学习感知图像块相似度（Learned Perceptual Image Patch Similarity，LPIPS）降低0.08，结构相似性指标（Structural Similarity Index Measure，SSIM）提升0.11，人脸关键点距离（Landmark Distance，LMD）降低1.10，音画同步评分（Lip-Sync Score，Sync）得分提高0.20；在VoxCeleb2数据集上，分别实现了PSNR提升7.12，LPIPS降低0.10，SSIM提升0.11，LMD降低1.10，Sync得分提高0.15. 实验结果充分验证了所提方法在复杂头部运动与多样身份条件下的有效性与优越性.

关键词： 视觉配音；说话人无关；跨模态注意力；3D人脸建模；运动纹理协同生成

基金项目： 国家重点研发计划（No.2022ZD0160604）

中图分类号： TP37 **文献标识码：** A **文章编号：** 0372-2112(2025)10-3608-14

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240685

Speaker-Independent Visual Dubbing Method Based on Semantic Enhancement and Texture-Motion Fusion

CHEN Yi-lei[1,3], XIONG Sheng-wu[2*]

(1. *Hubei Key Laboratory of Digital Finance Innovation*, *Hubei University of Economics*, *Wuhan*, *Hubei* 430205, *China*;
2. *School of Computer Science and Artificial Intelligence*, *Wuhan University of Technology*, *Wuhan*, *Hubei* 430070, *China*;
3. *School of Information Engineering*, *Hubei University of Economics*, *Wuhan*, *Hubei* 430205, *China*)

Abstract: Speaker-independent visual dubbing aims to edit the lip movements of talking face videos according to speech signals, ensuring high audio-visual synchronization and natural fidelity. This task not only requires accurate lip-sync performance but also demands consistent facial texture and identity preservation. However, existing methods often suffer from texture inconsistencies between the restored and original facial regions when natural head movements occur, leading to unstable generation quality. To address these challenges, this paper proposes a cross-modal semantic enhanced and 3D face-guided motion-texture synergistic generation network. Specifically, we adopt 3D morphable models (3DMM) as an intermediate representation and decompose the task into two submodules: cross-modal semantic enhanced 3DMM

收稿日期：2024-07-22；录用日期：2025-10-23；责任编辑：朱梅玉

*通讯作者：熊盛武

expression coefficient prediction and 3D face-guided motion-texture synergistic rendering. In the first stage, a cross-modal attention mechanism integrates Wav2Lip-generated semantic image sequences with audio features, significantly improving synchronization accuracy and geometric consistency. In the second stage, a 3D face-guided rendering network leverages multi-reference faces and reconstructed 3D geometry to enhance texture consistency under head motion, while a multi-task learning framework further refines visual fidelity between the restored and real facial regions. Extensive experiments on the VoxCeleb1 and VoxCeleb2 datasets demonstrate that the proposed method achieves superior performance in generation fidelity, motion robustness, and synchronization compared with state-of-the-art approaches. On VoxCeleb1, our method improves peak signal noise ratio (PSNR) by 7.76, reduces learned perceptual image patch similarity (LPIPS) by 0.08, increases structural similarity index measure (SSIM) by 0.11, decreases landmark distance (LMD) by 1.10, and improves lip-sync score (Sync) by 0.20 over the baseline. On VoxCeleb2, it improves PSNR by 7.12, reduces LPIPS by 0.10, increases SSIM by 0.11, decreases LMD by 1.10, and improves Sync by 0.15. These results verify the effectiveness and robustness of the proposed framework under complex head movements and diverse identities.

Key words: visual dubbing; speaker-independent talking face; cross-modal attention; 3D face modeling; motion-texture synergistic generation

Foundation Item(s): National Key Research and Development Program of China (No.2022ZD0160604)

1 引言

说话人无关的视觉配音(Speaker Independent Visual Dubbing,SIVD)方法,旨在通过输入语音信号驱动任意人物人脸视频中唇部区域的动态变化,实现音视频内容的语义一致与自然同步. 与传统方法相比,该类方法不依赖于特定说话人的建模或个性化调优,因而具有更强的泛化能力与适应性. 近年来,随着深度生成模型特别是扩散模型和生成对抗网络的快速发展,视觉配音技术得到了广泛关注,成为多模态生成领域的重要研究方向. 该技术在媒体内容创作、智慧教育、虚拟人构建、电影后期制作、人机交互等多个场景中展现出广阔的应用前景和商业潜力,正逐步推动语音驱动说话人脸视频合成从研究走向实际应用.

语音驱动的单样本说话人脸视频生成技术[1]正逐渐成为人工智能领域的研究热点,已有大量前期工作[2~6]围绕该任务展开. 当前的研究重点之一在于探索如何在仅给定一张静态人脸图像和一段语音信号的条件下,生成与语音内容高度同步、面部运动自然连贯的说话人脸视频. 不同于直接生成说话人脸视频的任务,说话人无关的视觉配音方法侧重于在已有视频的基础上,根据输入语音信号对目标视频中唇部区域的运动进行精准编辑,从而实现音视频内容在语义与时间维度上的一致性. 该任务不仅要求编辑结果具备较高的唇动同步精度,还需在保持原始视频中人物身份、姿态与表情一致性的前提下生成高度真实、自然的说话人脸序列. 因此,该任务具有更高的建模难度,需要结合其特定的目标约束与输入条件,专门设计具备语义建模与视觉保真能力的网络结构. 说话人无关的视觉配音任务中具有代表性的方法是Wav2Lip[7],其通过引入语音-视频同步判别器作为训练损失,有效提升了生成视频在音视频同步方面的性能,成为该方向的基础性工作. 随后,多项研究工作[8,9]在其框架基础上进行扩展,尝试通过更复杂的网络设计进一步提升同步准确率. Zhang等人[10]提出的SadTalker将音频映射至三维可变形人脸模型[11](3D Morphable Model, 3DMM)的表情与姿态系数,通过ExpNet学习表情变化、PoseVAE生成多风格头部运动,并结合三维感知渲染器生成说话人视频. 随着生成模型的发展,近年来也有学者将扩散模型引入语音驱动的人脸生成任务,Mukhopadhyay等人[12]提出的Diff2Lip方法基于音频条件的扩散生成机制,在自然场景下实现了高保真的说话人无关视觉配音. 然而,这类方法普遍采用生成模型直接对人脸区域进行重建,缺乏对面部结构与纹理细节的有效建模. 一方面,在处理存在显著头部运动的视频时,容易出现伪影或结构错位;另一方面,由于未显式引入人脸几何与纹理先验,生成图像常常缺乏真实感,特别是在唇部区域周围,面部纹理模糊、细节缺失,严重影响了视频的整体观感与质量. 上述问题不仅降低了模型的稳定性,也使得难以明确判断在不同姿态下生成失败的具体原因. 目前,只有少量研究工作[13,14]尝试引入原视频中的人脸结构信息,以增强视觉配音任务中的生成稳定性与保真度. 其中,Xie等人[13]采用人脸关键点作为中间表示,将说话人无关的视觉配音任务分解为语音驱动的关键点预测阶段与图像渲染阶段,从而实现语义建模与图像生成的解耦. 然而,该方法在渲染阶段容易引发身份泄露问题,即生成结果中出现参考图像中人物的身份特征,影响编辑视频的说话人一致性. Zhong等人[14]提出了IP_LAP方法以缓解身份泄露问题,该方法基于Transformer架构[15]构建个性化人脸关键点生成器,从而弱化中间表示对身份特征的依赖. 尽管在防止身份泄露方面取得了一定改进,但其渲染网络对参考人脸图像的结构和纹理信息利用不足,难以生成高质

量的面部细节. 总体而言,现有方法主要分为两类:其一,直接基于语音信号利用生成对抗网络实现跨模态人脸区域修复. 该类方法在存在显著头部运动的视频中易出现稳定性下降的问题,且当生成质量不足时缺乏可解释性,难以明确定位失败原因;其二,以人脸关键点为中间表示的分阶段建模方法,虽在一定程度上提升了模型结构的可控性,但潜在的身份泄露风险不可忽视. 此外,这两类方法在纹理细节重建方面均存在局限,生成的人脸区域与原视频真实面部纹理差异明显,难以满足高保真视频生成的要求.

为克服现有方法在自然头部运动场景下易出现唇部区域纹理不一致和音视频同步不稳的问题,本文提出了一种跨模态语义增强与3D人脸引导的运动纹理协同生成网络. 该方法建立在3DMM两阶段生成框架之上,将说话人无关的视觉配音任务分解为2个相互依赖的子任务:语音驱动的3DMM表情系数预测与3D人脸引导的运动-纹理协同渲染. 在语音建模阶段,构建了跨模态语义增强的3DMM表情系数预测网络. 该网络通过引入Wav2Lip生成的语义图像序列作为教师先验,并采用局部跨模态注意力机制对齐语音与视觉特征,从而在表情系数回归过程中同时强化了时序一致性与几何约束,显著提升了语音驱动下口型同步的稳定性与精确度. 在图像渲染阶段,针对仅依赖裁剪区域监督所导致的光流预测局部化和纹理细节不足的问题,本文提出了3D人脸引导的运动纹理协同渲染网络. 该模块由3个子网络组成:语义引导的运动流场建模网络(Semantic-guided Flow Modeling,SFM)、运动感知纹理调制网络(Texture Modulation Network,TMN)以及参考驱动的人脸修复网络(Reference Based Face Inpaint Network,RBFIN). 其中,SFM通过语义先验与3D几何约束预测全局非刚性运动流场,确保跨姿态条件下的几何对齐;TMN在运动引导下对多尺度纹理特征进行调制,实现运动与纹理信息的一致融合;RBFIN则综合利用3D重建与参考人脸的真实纹理信息,对唇部区域进行精细修复并保持身份一致性,还提出多尺度自适应融合模块(Multi-Scale Adaptive Fusion,MSAF)来有效地融合TMN的特征. 三者在协同优化下实现了全局运动建模与局部纹理还原之间的动态平衡,使得方法在复杂姿态和表情变化条件下依然能够生成高保真的说话人脸视频.

本文的主要贡献如下:

(1)提出了一种跨模态融合的语音预测3D人脸表情系数网络. 该方法在传统语音驱动预测框架中引入教师先验,利用Wav2Lip生成的语义图像序列作为辅助信息,实现了语音特征与视觉语义特征的跨模态融合. 通过这种设计,网络不仅提升了表情系数预测的精度,还为后续渲染提供了更可靠的动态先验.

(2)提出了一种多源融合修复网络,联合3D人脸先验、形变人脸和裁剪人脸图像,在解码端引入多尺度自适应融合模块,实现了对TMN输出纹理特征的多尺度建模与自适应注入. 该设计充分挖掘了运动特征与纹理特征的互补性,在保证几何一致性的同时有效恢复了高频细节,使生成结果在复杂头部运动场景下仍保持较高的纹理保真度与结构稳定性.

2 说话人无关的视觉配音相关工作

说话人无关的视觉配音方法分为2种主要技术路线:隐空间法和中间表示法. 其中,隐空间法主要使用语言信号在隐空间中扰动来编辑人脸图像. 中间表示法首先用语音信号预测中间表示的运动,然后使用中间表示生成人脸图像.

隐空间法:Prajwal等人[7]使用预训练的唇形同步鉴别器作为损失函数,首次合成了语音-视频同步率较高的说话人脸视频. Cheng等人[16]提出了一种新的系统VideoReTalking,将视觉配音任务分解为3个顺序子任务:人脸视频生成、语音-唇形同步和面部增强. 论文使用基于学习的方法解决了这3个步骤,并且所有模块都可以在顺序流水线中处理. Sun等人[17]研究使用音频和视觉信息来实现对任意目标视频的嘴部运动驱动,通过精心设计的注意力机制和细化网络,实现了高质量的语音-视频同步的生成结果. Song等人[18]研究快速适应新身份的人脸配音方法,提出了一种新的风格翻译网络,通过跨模态的自适应实例归一化[19](Adaptive Instance Normalization,AdaIN)模块将目标说话者的说话风格与源说话者的说话内容进行整合,从而实现了模型对新说话者的快速适应. Yang等人[20]研究了基于上下文感知的说话人脸配音方法,通过将目标视频的上下文信息与源视频的外观信息进行融合,实现了高质量的说话人脸视频编辑. 该方法能够根据目标视频的语义和情感上下文,自动合成与目标视频一致的说话人脸视频,具有较好的视觉一致性和自然度. Park等人[8]通过引入语音-视觉记忆模块,在训练过程中将语音特征与对应的唇部特征对齐,从而在推理时能够输出与音频对齐的唇部特征来提高音视频同步的质量. Guan等人[21]研究在保持高保真度的同时实现音视频同步的说话人脸视频生成,通过对基于Style-Gan的生成器进行简单但有效的修改,实现了高保真度的音视频同步. 通过引入面部细节的掩膜引导空间信息编码模块,通过调制卷积准确地修改唇部运动,同时通过样式空间和生成器优化实现个性化的唇同步. Ki等人[22]研究了个性化的说话人脸视频配音方法,以适应不同人物的说话风格. Wang等人[9]研究了使用唇读网络来提

高音视频同步的效果,提出了使用预训练的唇读识别专家模型来惩罚生成的视频中的错误口型,并使用了一种新的策略来评估唇读的可懂性. Mukhopadhyay 等人提出的 Diff2Lip[12]方法利用音频条件扩散模型生成与语音同步的唇部运动,在开放场景下实现了较高的视觉保真度. 该方法通过扩散去噪过程逐步生成符合语音语义的唇形帧,显著改善了图像质量. 但由于缺乏显式的几何约束,在复杂姿态或大范围头动条件下,生成结果仍存在结构漂移与纹理模糊等问题.

中间表示法:在说话人无关的视觉配音任务中,中间表示法作为重要技术路线,已经被广泛应用于音频与视觉的同步. 该方法主要通过面部关键点、3DMM 或扩散模型等中间表示进行音频与视觉特征的融合. Xie 等人[13]提出了基于人脸关键点的中间表示,将视觉配音任务分解为语音驱动的关键点预测和图像渲染. 该方法通过修改唇部关键点的运动来编辑说话人脸视频,但在渲染阶段可能引发身份泄露问题,即生成结果中可能出现参考图像中人物的身份特征,影响说话人一致性. Zhang 等人[10]提出的 SadTalker 方法基于音频驱动的 3DMM 模型,通过预测表情与姿态系数生成说话人脸视频,在头部运动的自然度方面表现较好. 但该方法语音与视觉特征对齐精度有限,唇部细节和纹理稳定性不足. Zhong 等人[14]提出了 IP_LAP 方法,采用基于 Transformer 的关键点预测与多参考图像渲染框架,通过修复人脸下半部分生成说话人无关的视频. 该方法有效改善了细节合成并减少伪影,但对参考图像的结构和纹理利用不足,在高保真度和身份保持方面仍存在局限. Guan 等人[23]提出通过修改 Style-Gan 生成器架构,将音频特征和视觉特征中的运动信息有效对齐,解决了音频与视觉的同步问题. 该方法显著提升了面部运动的同步性,特别是在生成自然的嘴唇运动和面部细节方面表现优异. Zhang 等人[24]提出 PersonaTalk,引入个性化注意力机制和跨模态风格传递网络,将目标说话者的个性特征与源视频音频内容进行整合,生成更自然且具有身份感的说话人脸动画. 该方法在不同人物的说话风格生成上具有较好的适应性,并在音视频同步精度和视觉一致性方面表现优秀.

不同于上述方法,本文提出的跨模态语义增强与 3D 人脸引导的运动-纹理协同生成方法结合了语义与几何两类先验. 该方法引入 Wav2Lip 生成的语义图像序列作为教师先验,并在 3DMM 框架下实现多源特征的自适应融合,在提升音视频同步精度的同时,改善了身份保持与纹理稳定性,能够在复杂头部运动条件下生成更自然一致的说话人脸视频.

3　提出方法

本文提出的融合多元先验知识的运动鲁棒说话人无关视觉配音方法,整体框架如图 1 所示. 以 3DMM 为中间表示,将任务分解为语音驱动的表情预测与多参考人脸引导的图像修复 2 个子任务,并在此基础上构建相应的子网络以实现高保真、身份一致的说话人脸视频生成.

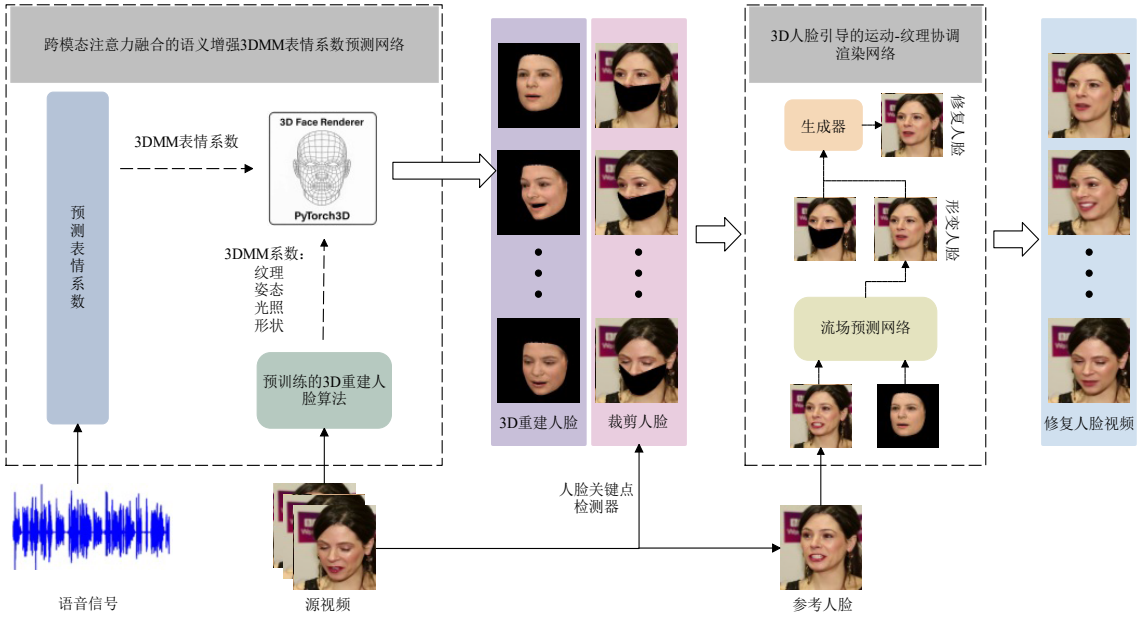

图 1　跨模态语义增强与 3D 人脸引导的运动-纹理协同生成网络

3.1　三维人脸形变模型

三维人脸形变模型用 3DMM 系数的集合 $\chi(I)=\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta},\boldsymbol{\gamma},\boldsymbol{p}\in\mathbf{R}^{257}$ 来表示人脸图像 I,其中,$\boldsymbol{\alpha}\in\mathbf{R}^{80}$ 为人脸身份系数向量,$\boldsymbol{\beta}\in\mathbf{R}^{64}$ 为人脸表情系数,$\boldsymbol{\delta}\in\mathbf{R}^{80}$ 为纹理

系数，$\boldsymbol{\gamma}\in\mathbf{R}^{27}$为光照系数，$\boldsymbol{p}\in\mathbf{R}^{6}$为包含旋转和平移的姿态向量，则人脸形状$S$和面部纹理$T$可以分别表示为式(1)和式(2)：

$$S=\overline{S}+B_{\mathrm{id}}\boldsymbol{\alpha}+B_{\mathrm{exp}}\boldsymbol{\beta} \tag{1}$$

$$T=\overline{T}+B_{\mathrm{tex}}\boldsymbol{\delta} \tag{2}$$

其中，$\overline{S}$和$\overline{T}$为平均形状和平均纹理，B_{id}、B_{exp}和B_{tex}分别为形状、表情和纹理的PCA基. B_{id}、B_{tex}采用2009年的巴塞尔人脸模型[25](2009 Basel Face Model)，B_{exp}采用FaceWareHouse[26]提供的人脸表情基. 本文使用3D人脸重建算法[27]预测人脸图像I的3DMM系数$\chi(I)$，假设人脸为兰伯曲面(Lambertian Surface)，并使用球面近似场景照明[28]，顶点v_i与法向量$\boldsymbol{n}_i$和纹理向量$\boldsymbol{t}_i$的辐照度为$C(\boldsymbol{n}_i,\boldsymbol{t}_i,\boldsymbol{\gamma})=\boldsymbol{t}_i\sum_{b=1}^{B^2}\boldsymbol{\gamma}_b\Phi_b(\boldsymbol{n}_i)$. 其中，$\Phi_b:\mathbf{R}^3\to\mathbf{R}$为SH基函数，$\boldsymbol{\gamma}_b$为SH系数，$B=3$为SH带数，人脸姿态由旋转角度和平移来表示，使用透视相机模型将3D人脸模型投影到图像平面上.

选择3DMM作为中间表示主要有3点优势：其一，3DMM系数可独立控制表情与姿态变化，避免身份泄露；其二，渲染得到的三维人脸同时提供几何与纹理先验，有助于渲染网络生成更真实的人脸图像；其三，三维重建人脸易于从自然场景图像中获取，并在复杂姿态下保持较强鲁棒性，能够有效提升网络的稳定性与泛化能力.

3.2 跨模态注意力融合的语义先验增强3DMM表情系数预测网络

在现有视觉配音方法中，基于生成对抗网络的唇部重建技术较为典型，其中Wav2Lip通过在裁剪人脸区域引入同步判别器，有效提升了音视频对齐精度. 但该方法在复杂场景下仍存在局限：一方面，面对大幅度头部运动或视角变化时易产生伪影和错位；另一方面，生成图像分辨率较低，纹理细节不足，难以保持身份一致性. 为此，本文提出语义先验增强的语音驱动三维表情系数预测网络，如图2所示.

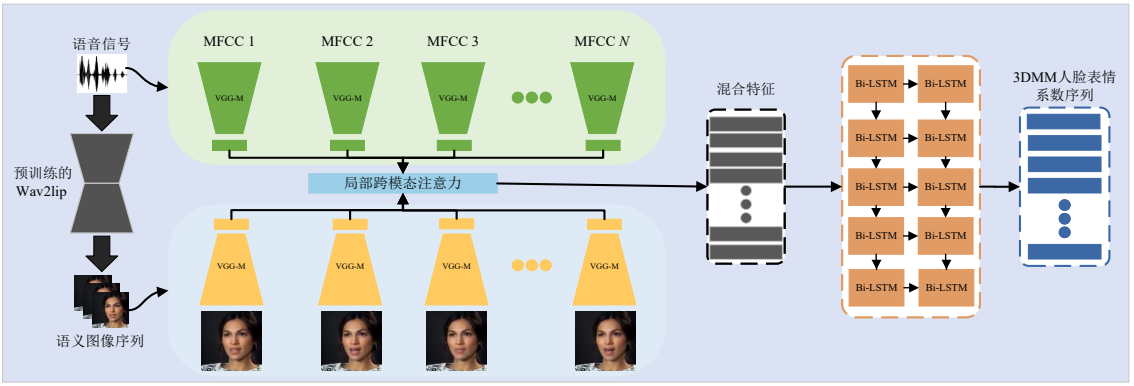

图2 跨模态注意力融合的语义先验增强3DMM表情系数预测网络

其中，网络以语音梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficients，MFCC)特征与Wav2Lip生成的语义图像序列为输入，预测3DMM表情参数序列，如式(3)所示：

$$A=\{a_1,a_2,\cdots,a_T\},\quad V=\{v_1,v_1,\cdots,v_T\} \tag{3}$$

输入编码：逐帧提取音频与视觉特征，采用VGG-M卷积逐帧提取MFCC语义特征F_a^t，采用VGG-M提取语义视觉特征F_v^t，如公式(4)所示：

$$\begin{aligned}F_a^t&=\mathrm{CNN}_{\mathrm{audio}}(a_t)\in\mathbf{R}^{d_a},\\F_v^t&=\mathrm{CNN}_{\mathrm{visual}}(v_t)\in\mathbf{R}^{d_v},t=1,2,\cdots,T\end{aligned} \tag{4}$$

局部跨模态注意力融合：为实现细粒度对齐，在时刻t构造视觉局部时间窗，如式(5)所示：

$$N(t)=t-k,\cdots,t,\cdots,t+k;\ |\mathcal{N}(t)|=2k+1 \tag{5}$$

将特征投影到统一维度d，并采用行向量/行堆叠的记法以保证点积维度匹配，如式(6)所示：

$$\begin{aligned}q^t&=\left(W_Q F_a^t\right)^{\mathrm{T}}\in\mathbf{R}^{1\times d},\\K^t&=\left[\left(W_K F_v^\tau\right)^{\mathrm{T}}\right]_{\tau\in\mathcal{N}(t)}\in\mathbf{R}^{(2k+1)\times d},\\V^t&=\left[\left(W_V F_v^\tau\right)^{\mathrm{T}}\right]_{\tau\in\mathcal{N}(t)}\in\mathbf{R}^{(2k+1)\times d}\end{aligned} \tag{6}$$

其中，$W_Q\in\mathbf{R}^{d\times d_a}$，$W_K,W_V\in\mathbf{R}^{d\times d_v}$为可学习参数. 采用缩放点积注意力在时间窗维度归一化，如式(7)所示：

$$\begin{aligned}\alpha^t&=\mathrm{Softmax}\left(\frac{q^t\left(K^t\right)^{\mathrm{T}}}{\sqrt{d}}\right)\in\mathbf{R}^{1\times(2k+1)},\\z^t&=\alpha^t V^t\in\mathbf{R}^{1\times d}\end{aligned} \tag{7}$$

其中，Softmax沿时间窗长度$(2k+1)$进行归一化，由此得到的跨模态融合特征记为z^t.

时序建模与表情系数预测：将$\{z^t\}(t=1,2,\cdots,T)$作为长度为T、维度为d的序列输入Bi-LSTM，如式(8)所示：

$$\{h_t\}(t=1,2,\cdots,T)=\text{Bi-LSTM}\left(\{z^t\}(t=1,2,\cdots,T)\right),\quad h_t\in\mathbf{R}^h \tag{8}$$

经线性映射得到帧级表情参数预测,如式(9)所示:

$$\beta_t^*=W_o h_t+b_o\in\mathbf{R}^m,\ t=1,2,\cdots,T \tag{9}$$

其中,m为表情参数维度,$W_o\in\mathbf{R}^{m\times h}$, $b_o\in\mathbf{R}^m$. 由式(3)~(9)给出的整体映射可记为式(10):

$$M_e:(A,V)\mapsto\beta^*=\{\beta_1^*,\beta_2^*,\cdots,\beta_T^*\} \tag{10}$$

为了确保生成结果的几何一致性,本文引入了形状约束和表情系数约束. 在训练过程中,首先通过将预测的3D人脸表情系数序列$\beta^*=\{\beta_1^*,\beta_2^*,\cdots,\beta_T^*\}$与真实的3DMM参数组合生成三维人脸网格,并通过可微投影生成对应的2D人脸关键点序列$l_t^*\in\mathbf{R}^{2N}$. 同理,基于真实的表情系数β_t也可以生成真实的2D关键点l_t.

表情系数约束:本文对表情系数采用L_1损失,即计算预测表情系数与真实表情系数之间的绝对误差,如式(11)所示:

$$\mathcal{L}_{\text{exp}}=\sum_{t=1}^{T}\left\|\beta_t-\beta_t^*\right\|_1 \tag{11}$$

其中,β_t为真实的表情系数,β_t^*为预测的表情系数,T为序列长度. 此损失函数确保表情系数的精确预测,并且增强模型对异常值的鲁棒性.

形状约束损失:为了保持几何一致性,本文对2D关键点采用L_1对齐约束,度量预测的2D关键点与真实2D关键点之间的几何差异,如式(12)所示:

$$\mathcal{L}_{\text{shape}}=\sum_{t=1}^{T}\left\|l_t-l_t^*\right\|_1 \tag{12}$$

其中,l_t为真实的2D关键点,l_t^*为预测的2D关键点,T为序列长度. 该损失函数直接度量预测关键点与真实关键点之间的几何差异,确保生成的人脸在几何形态上的一致性.

最终的总损失由表情系数回归损失和形状约束损失的加权和构成. 其中,λ_{shape}是形状约束损失的权重,通常设置为0.5,以平衡两者对最终优化目标的影响,如式(13)所示:

$$\text{L}=\mathcal{L}_{\text{exp}}+\lambda_{\text{shape}}\mathcal{L}_{\text{shape}} \tag{13}$$

该总损失函数能够同时优化表情系数与形状一致性,从而得到精确且结构合理的3D人脸表情.

3.3　3D人脸引导的运动-纹理协同渲染网络

为提升缺失区域的纹理保真度,本文提出一种基于三维人脸先验的运动-纹理协同融合渲染框架,如图3所示. 该框架包含的3个子网络SFM、TMN以及RBFIN三者协同作用,分别实现几何对齐、动态纹理补偿与多源细节融合,从而在保证结构一致性的同时提高纹理保真度与视觉一致性.

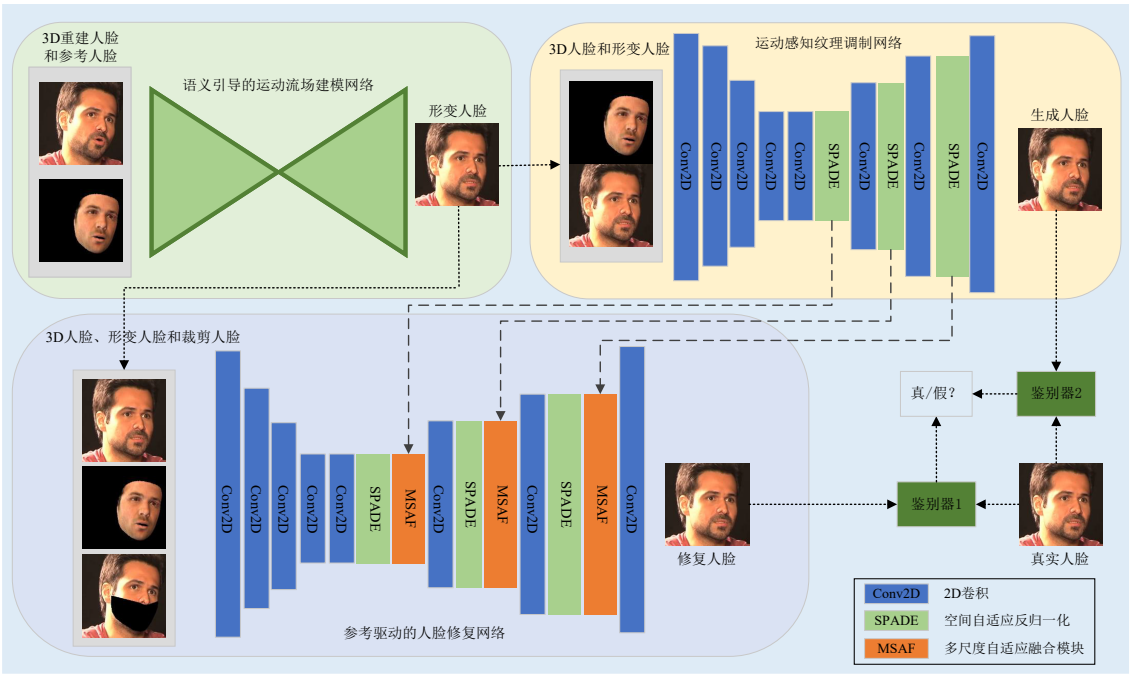

图3　3D人脸引导的运动-纹理协同融合人脸渲染网络

3.3.1　3D人脸引导的运动-纹理协调融合渲染网络

SFM:SFM以三维重建人脸$I_{3D}\in\mathbf{R}^{3\times H\times W}$为几何先验,结合参考图像$I_r\in\mathbf{R}^{3\times H\times W}$预测像素级光流场$\omega\in\mathbf{R}^{H\times W\times 2}$,并通过空间变换网络[29](Spatial Transform Network,STN)生成对齐图像I_{warp},其表达式如式(14)所示:

$$\omega=\mathrm{SFM}\left(I_{3\mathrm{D}},I_r\right),\quad I_{\mathrm{warp}}=\mathrm{STN}\left(\omega,I_r\right)\tag{14}$$

其中,光流场ω刻画了参考图像像素到目标图像像素的映射关系,保证跨视角、跨姿态条件下的几何一致性,为后续纹理补偿与修复提供稳定约束.

TMN:在对齐图像I_{warp}的基础上,TMN引入三维先验$I_{3\mathrm{D}}$,联合建模运动信息与纹理细节.网络输出颜色掩膜C_e、注意力掩膜A_e以及中间纹理特征F_{tex},其计算方式如式(15)所示:

$$\left(C_e,A_e,F_{\mathrm{tex}}\right)=\mathrm{TMN}\left(I_{\mathrm{warp}},I_{3\mathrm{D}}\right)\tag{15}$$

并利用注意力掩膜对颜色图进行加权调制,如式(16)所示:

$$I_e^*=A_e\odot C_e+\left(1-A_e\right)\odot I_{\mathrm{warp}}\tag{16}$$

该过程在保持几何一致性的同时增强了细节表达能力.与此同时,TMN的解码器部分在若干层中引入空间自适应反归一化模块[30](Spatially-Adaptive Denormalization, SPADE),以便在生成过程中将三维先验条件以空间自适应的方式注入特征,使运动与纹理的建模更加一致与精细.得到的中间特征F_{tex}为后续修复阶段提供了更高质量的纹理信息,这里的F_{tex}表示解码器的特征,可以是多个尺度也可以是单一尺度,后续为了叙述方便,对不同尺度不作区分.

RBFIN:以形变人脸I_{warp}、三维重建人脸$I_{3\mathrm{D}}$、裁剪人脸I_{crop}以及TMN输出的F_{tex}为输入,预测颜色掩膜与注意力掩膜,如式(17)所示:

$$\left(C_i,A_i\right)=\mathrm{RBFIN}\left(I_{\mathrm{warp}},I_{3\mathrm{D}},I_{\mathrm{crop}},F_{\mathrm{tex}}\right)\tag{17}$$

并生成修复结果I_i^*,如式(18)所示:

$$I_i^*=A_i\odot C_i+\left(1-A_i\right)\odot I_{\mathrm{crop}}\tag{18}$$

在解码阶段,设第k层的解码特征为$F_k\in\mathbf{R}^{C_k\times H_k\times W_k}$,TMN在对应尺度的纹理特征记为$F_{\mathrm{tex}}^{(k)}$.此外,解码器在部分层中引入SPADE,以空间自适应的方式将几何先验条件注入到生成过程,从而提升纹理与结构的一致性.为实现两者的有效融合[31],本文在解码端提出使用多尺度自适应融合模块(Multi-Scale Adaptive Fusion, MSAF).首先对$F_{\mathrm{tex}}^{(k)}$进行多尺度卷积变换,如式(19)所示:

$$T_{k,s}=\mathrm{Conv}_{s\times s}\left(F_{\mathrm{tex}}^{(k)}\right),\ s\in 3,5,7\tag{19}$$

然后,使用式(20)计算Softmax归一化得到尺度权重:

$$W_{k,s}=\frac{\exp\left(\gamma_{k,s}\right)}{\sum\limits_{t\in\{3,5,7\}}\exp\left(\gamma_{k,t}\right)},\ \sum_s W_{k,s}=1\tag{20}$$

通过1×1卷积和Sigmoid激活生成空间门控掩膜,如式(21)所示:

$$\begin{gathered}M_k=\sigma\left(\mathrm{Conv}_{1\times1}\left(\left[F_k,\ T_{k,3},\ T_{k,5},\ T_{k,7}\right]\right)\right)\\ M_k\in\left(0,1\right)^{1\times H_k\times W_k}\end{gathered}\tag{21}$$

最终,融合后的解码特征表示为式(22):

$$F_k^{\mathrm{fusion}}=F_k+\sigma\cdot\left(M_k\odot\sum_{s\in\{3,5,7\}}W_{k,s}T_{k,s}\right)\tag{22}$$

其中,σ为可学习缩放系数,用于调节纹理特征在解码器中的注入强度.若直接引入全部增强分支,可能导致训练早期梯度不稳定.通过在残差路径引入σ,网络能够在不同阶段自适应地控制增强作用:当分支特征有效时,σ倾向于增大其贡献;当分支特征无效或噪声较多时,σ会减弱其影响,从而避免性能退化.该机制保证了RBFIN在细节恢复与结构稳定性之间的动态平衡.

3.3.2 损失函数

为保证所生成的人脸图像在结构一致性与纹理保真度上的综合性能,本文在渲染网络训练过程中引入多项损失函数,并采用分阶段权重策略进行端到端优化.整体目标是同时兼顾几何对齐、全局结构、细节纹理及感知真实性,从而提升人脸修复与合成效果.

重建与结构损失:首先,在像素空间引入基于L_1范数的重建损失,以约束生成图像与目标图像之间的差异.同时,利用结构相似性指标(Structural Similarity Index Measure, SSIM)保证全局结构一致性,如式(23)和式(24)所示:

$$\begin{aligned}L_{\mathrm{rec}}^i&=\left\|I_t-I_i^*\right\|_1,\\ L_{\mathrm{rec}}^e&=\frac{1}{\left|\Omega_c\right|}\sum_{x\in\Omega_c}\left\|I_t(x)-I_e^*(x)\right\|_1\end{aligned}\tag{23}$$

$$\begin{aligned}L_{\mathrm{SSIM}}^i&=1-\mathrm{SSIM}\left(I_t,I_i^*\right),\\ L_{\mathrm{SSIM}}^e&=1-\mathrm{SSIM}\left(I_t,I_e^*\right)\end{aligned}\tag{24}$$

其中,I_t为目标图像,I_i^*、I_e^*分别为修复结果与中间合成结果.

本文采用WGAN-GP(Wasserstein Generative Adversarial Network with Gradient Penalty)作为对抗损失函数,有效改善生成对抗网络的训练稳定性和提升生成图像的真实感,如式(25)所示:

$$\begin{aligned}L_{\mathrm{wgan}}^i&=-\mathrm{E}\left[D\left(I_t^*\right)\right],\\ L_{\mathrm{wgan}}^e&=-\mathrm{E}\left[D\left(I_e^*\right)\right]\end{aligned}\tag{25}$$

其中,$D(\cdot)$为判别器,其优化目标定义为式(26):

$$\begin{aligned}&\max_D\mathrm{E}\left[D\left(I_t\right)\right]-\mathrm{E}\left[D\left(I_t^*\right)\right]-\mathrm{E}\left[D\left(I_e^*\right)\right]\\ &-\lambda_{\mathrm{gp}}\mathbb{E}_{\hat{x}}\left(\left\|\nabla_{\hat{x}}D\left(\hat{x}\right)\right\|_2-1\right)^2\end{aligned}\tag{26}$$

其中,$\hat{x}=\epsilon I_t+\left(1-\epsilon\right)I_g$,$I_g\in\{I_t^*,I_e^*\}$,$\epsilon\sim\mathcal{U}(0,1)$和$\lambda_{\mathrm{gp}}>0$为梯度惩罚系数.该损失能够有效缓解训练不稳定性,增强生成结果的清晰度与自然度.

本方法引入一致性损失[32](Correspondence Wise Loss, CWL)来度量形变人脸图像I_{warp}与目标人脸图像I_t之间的位置误差,如式(27)所示:

$$L_{\text{corr}}=\left\|\mathcal{C}(I_t)-\mathcal{C}(I_{\text{warp}})\right\|_2 \tag{27}$$

其中，$\mathcal{C}(\cdot)$表示预训练的光流预测网络，通过使用一致性损失提升形变人脸图像和目标人脸图像之间的一致性关系，进而提升渲染网络的鲁棒性.

感知与风格损失[33]：为提升生成图像在高层语义特征上的一致性，引入基于VGG-19特征的感知损失，如式(28)所示：

$$\begin{aligned}L_p^i&=\sum_{l\in\mathcal{L}}\frac{1}{C_lH_lW_l}\left\|\varphi_l\left(I_i^*\right)-\varphi_l\left(I_t\right)\right\|_1\\L_p^e&=\sum_{l\in\mathcal{L}}\frac{1}{C_lH_lW_l}\left\|\varphi_l\left(I_e^*\right)-\varphi_l\left(I_t\right)\right\|_1\end{aligned} \tag{28}$$

其中，$\phi_l(\cdot)$为VGG-19第l层的特征映射，$\left(C_l,H_l,W_l\right)$表示特征图的维度参数.

此外，为进一步保持生成图像的纹理统计特性，在修复人脸图像I_i^*上引入基于Gram矩阵的风格损失，如式(29)所示：

$$L_{\text{style}}^i=\sum_{l\in\mathcal{L}_s}\frac{1}{\left(C_lH_lW_l\right)^2}\left\|\mathrm{G}\left(\phi_l\left(I_i^*\right)\right)-G\left(\phi_l\left(I_t\right)\right)\right\|_1 \tag{29}$$

其中，$\boldsymbol{G}(F)=\mathrm{FF}^{\mathrm{T}}$为Gram矩阵，$\mathcal{L}_s$为选取的特征层集合. 该损失能够有效约束高频纹理分布，增强唇周与面部细节的锐利度.

两个子网络的优化目标分别如式(30)和式(31)所示：

$$\begin{aligned}L_{\text{total}}^i=&L_{\text{rec}}^i+\lambda_{\text{SSIM}}L_{\text{SSIM}}^i+\lambda_pL_p^i\\&+\lambda_{\text{style}}L_{\text{style}}^i+\lambda_{\text{wgan}}L_{\text{wgan}}^i\end{aligned} \tag{30}$$

$$\begin{aligned}L_{\text{total}}^e=&L_{\text{rec}}^e+\lambda_{\text{SSIM}}L_{\text{SSIM}}^e+\lambda_pL_p^e\\&+\lambda_{\text{wgan}}L_{\text{wgan}}^e+\lambda_{\text{corr}}L_{\text{corr}}\end{aligned} \tag{31}$$

最终，渲染网络的整体损失定义为式(32)：

$$L_{\text{total}}=\varepsilon L_{\text{total}}^i+(1-\varepsilon)L_{\text{total}}^e \tag{32}$$

为了保证训练流场的稳定性，在训练初期只训练SFM和TMN，待网络训练稳定后，再联合训练SFM、TMN和RBFIN. 联合训练时，ε在训练过程中分阶段设定：在早期训练阶段取$\varepsilon=0.1$，以突出运动流场的学习；在跨模态流场预测稳定后取$\varepsilon=0.2$，以更注重人脸修复与纹理细化.

4　实验分析

4.1　数据集和实现细节

本文所提出的方法在以下3个数据集上训练和测试：LRW[34]、VoxCeleb1[35]和VoxCeleb2[36]. 本文实验使用LRW数据集训练教师先验知识融合的语音预测3D人脸表情系数网络，遵循与文献[2]相同的预处理过程，使用Wav2lip[7]根据语音信号生成对应的语义图像序列. 本文在VoxCeleb1和VoxCeleb2上训练多人脸先验引导的细节纹理增强渲染网络. VoxCeleb1数据集包含22 496个从YouTube视频中提取的说话人脸视频，本文遵循文献[37]中描述的预处理方法，经过预处理后得到420万张人脸图像用于训练. 然后，使用Deng等人[27]训练好的3D人脸重建模型提取相应的3D人脸系数，使用现代渲染引擎Pytorch3D渲染对应的3D重建人脸图像，使用Bulat等人[38]提出的人脸关键点检测器检测人脸关键点，并用OpenCV绘制裁剪人脸区域得到裁剪人脸. VoxCeleb2数据集共包含6 112位名人，训练集中有5 994位说话者，测试集中有118位，预处理后大约有2 000万张图像，使用与VoxCeleb1相同的预处理方式得到相应的裁剪人脸和3D重建人脸.

实现细节：本文使用Ranger优化器训练教师先验知识融合的语音预测3D人脸表情系数网络. Ranger是一种结合了RAdam[39]和LookAhead[40]的协同优化器，学习率lr=0.000 2，教师先验知识融合的语音预测3D人脸表情系数网络结构借鉴了Yi等人的工作[41]，设置batch=16，在LRW数据集上训练耗时1天. 使用Ranger优化器训练多人脸先验引导的细节纹理增强渲染网络，设置学习率为lr=0.000 1. 该网络的生成器结构为Unet结构，参考人脸图像和3D重建人脸图像大小调整为192×192，设置batch=2，在VoxCeleb1数据集上训练渲染网络耗时3天，在VoxCeleb2数据集上训练渲染网络耗时1周，所有模型使用PyTorch实现，使用1个24 GB Geforce 3090 GPU训练.

4.2　与对比方法的定量结果和定性结果

为全面评估所提方法的性能，本文选取了多种具有代表性的说话人脸生成方法进行对比. MakeItTalk[42]以单张人像为输入，利用音频驱动预测人脸关键点轨迹，并通过生成对抗网络渲染生成对应的说话人脸视频，属于典型的单样本生成方法；SadTalker[10]基于三维人脸参数建模，将音频映射至3DMM表情与姿态系数，并结合三维感知渲染器生成具有自然头部运动的说话人视频；Wav2Lip[6]构建跨模态对抗生成框架，引入唇形同步判别器以显示约束音频与视频的时序一致性，实现高精度的口型同步；IP_LAP[14]采用两阶段结构，结合人脸关键点先验与外观先验，在说话人无关条件下提升身份保持与视听同步性能；Diff2Lip[12]基于扩散模型使用语音信号引导唇部区域的图像修复.

评估指标：本文使用SSIM和峰值信噪比(Peak Signal Noise Ratio，PSNR)来评估人脸图像质量，使用学习感知图像块相似度(Learned Perceptual Image Patch Similarity，LPIPS)来评估生成人脸图像与真实人脸图像之间的相似性[40]，使用人脸关键点距离(Landmark Distance，LMD)评估面部几何相似度[43]，使用身份余弦相似度[43](Cosine SIMilarity，CSIM)来评估身份相似度，使用音画同步评分(Lip-Sync score，Sync)来评估语音-视频同步的准确性[34]，使用平均表情距离(Average Ex-

pression Distance，AED）[41]来评估预测的人脸表情系数和真实人脸表情系数之间的差距.

表1和表2给出了本方法在VoxCeleb1与VoxCeleb2测试集上的定量评估结果. 由表中数据可见，与当前典型的语音驱动视频生成方法相比，本方法在各项指标上均取得较为优异的性能表现. 在图像质量相关指标方面，本方法在PSNR、SSIM等指标上整体优于对比方法，说明其在面部纹理重建与视觉保真度方面具有更好的生成能力；在LPIPS与LMD等反映感知一致性与几何精度的指标上，本方法亦呈现较低的误差，表明其在口型形变与面部几何一致性方面具备较高稳定性. 同时，在身份保持与音视频同步性指标（如CSIM与Sync）上，本方法也相对取得更高得分，验证了所引入的3D结构先验与多源信息融合机制对于提升跨人物与跨姿态生成质量的有效作用. 总体来看，本方法在两个测试集上均取得了较为全面的性能提升，尤其在存在大幅度头部运动与复杂表情变化的场景中，其生成结果在清晰度、几何一致性和同步性能等方面均表现稳定，说明所提出框架能够在多种条件下保持良好的泛化性与鲁棒性.

表1 本文提出方法和其他主流方法在VoxCeleb1测试集上定量结果比较

方法	PSNR ↑	LPIPS ↓	SSIM ↑	LMD ↓	CSIM ↑	Sync ↑
MakeItalk	28.84	0.28	0.64	3.21	0.97	4.81
Wav2Lip	33.64	0.12	0.86	2.41	0.98	5.28
SadTalker	30.75	0.21	0.80	3.54	0.98	5.14
IP_LAP	34.42	0.11	0.91	2.54	0.99	5.26
Diff2Lip	33.24	0.16	0.90	2.43	0.99	5.16
Ours	**38.31**	**0.04**	**0.95**	**1.62**	**0.99**	**5.42**

注：↑表示指标越高性能越好，↓表示指标越低性能越好. 加粗数字表示该表格中指标最佳值.

表2 本文提出方法和其他主流方法在VoxCeleb2测试集上定量结果比较

方法	PSNR ↑	LPIPS ↓	SSIM ↑	LMD ↓	CSIM ↑	Sync ↑
MakeItalk	30.47	0.19	0.77	3.42	0.97	5.31
Wav2Lip	33.57	0.14	0.88	2.64	0.98	5.48
SadTalker	30.75	0.16	0.82	2.84	0.98	5.34
IP_LAP	34.96	0.10	0.92	2.46	**0.99**	5.32
Diff2Lip	33.94	0.15	0.91	2.38	**0.99**	5.28
Ours	**38.45**	**0.04**	**0.96**	**1.54**	**0.99**	**5.62**

注：↑表示指标越高性能越好，↓表示指标越低性能越好. 加粗数字表示该表格中指标最佳值.

定性结果：在VoxCeleb1和VoxCeleb2测试集上的定性结果如图4和图5所示. 从整体视觉效果来看，本方法生成的人脸视频在唇部运动、身份一致性及细节还原方面均表现优异. 与语音驱动的单样本生成方法（MakeItalk、SadTalker）相比，本方法在保持语音-视频同步的同时，能够生成更自然流畅的头部运动和更稳定的面部结构，避免了姿态变化下的失真和漂移问题. Wav2Lip在唇部区域存在模糊现象，其对抗生成结构依赖局部修补机制，难以在复杂头动场景中保持纹理细节与唇形边界的清晰度. IP_LAP基于关键点的两阶段生成框架在一定程度上维持了身份一致性，但唇部动作的精确性不足，面部细节表现略显平滑. Diff2Lip利用音频条件扩散模型提升了生成的平滑度与真实性，但由于未充分引入参考图像的几何与纹理先验，其生成结果在面部细节上仍与真实图像存在差异.

相比之下，本方法在语义一致性和视觉保真度上表现更为突出. 通过融合语义先验与参考图像纹理特征，模型能够在复杂姿态变化下生成唇部边界清晰、表情自然且身份一致的人脸视频，整体视觉效果更接近真实视频. 这表明本方法在高保真度和时序稳定性方面均优于现有代表性方法.

4.3 消融实验

表3给出了在LRW测试集上的消融实验结果. 可以看出，语义图像信息、跨模态注意力机制以及形状约束损失均对三维表情系数的预测具有不同程度的提升作用. 相比基础模型，去除任一组件都会导致表情参数精度下降或音视频同步性减弱，说明语义先验能够增强面部结构约束，跨模态注意力有助于实现更有效的音视频对齐，而形状约束可进一步提高口型与表情的几何一致性. 完整模型在各项指标上均取得最优表现，验证了多模块协同设计对整体性能提升的重要性.

表4展示了在VoxCeleb1测试集上的渲染网络消融实验结果. 实验表明，运动-纹理协同渲染网络中的关键模块均对最终渲染质量具有显著贡献. 基础模型在缺少运动感知与特征增强机制时，生成结果在纹理保真度、结构一致性与面部动态稳定性方面均出现明显退化；移除特征增强网络后，面部细节的复原能力进一步减弱，局部区域纹理表达不足；在缺少一致性约束的情况下，跨帧表达易出现几何偏差，整体稳定性降低；多尺度自适应融合模块的去除会削弱模型整合不同尺度特征的能力，不利于细节保持；此外，取消分阶段训练策略会导致优化过程不够稳定，使网络更易陷入次优状态. 总体而言，完整模型在视觉质量、几何一致性与动态表现等方面均优于其消融版本，说明多模块协同设计与分阶段训练策略对提升渲染性能具有重要作用.

为验证所提出的MSAF的有效性，本文从2个方面开展消融实验：缩放系数σ的设置方式以及在不同特征尺度上的使用情况，如表5所示. 当σ固定为0或1时，性能均低于可学习情况. 其中，σ=0表示完全不使用特征增强，导致多感受野特征无法发挥作用；σ=1表示强

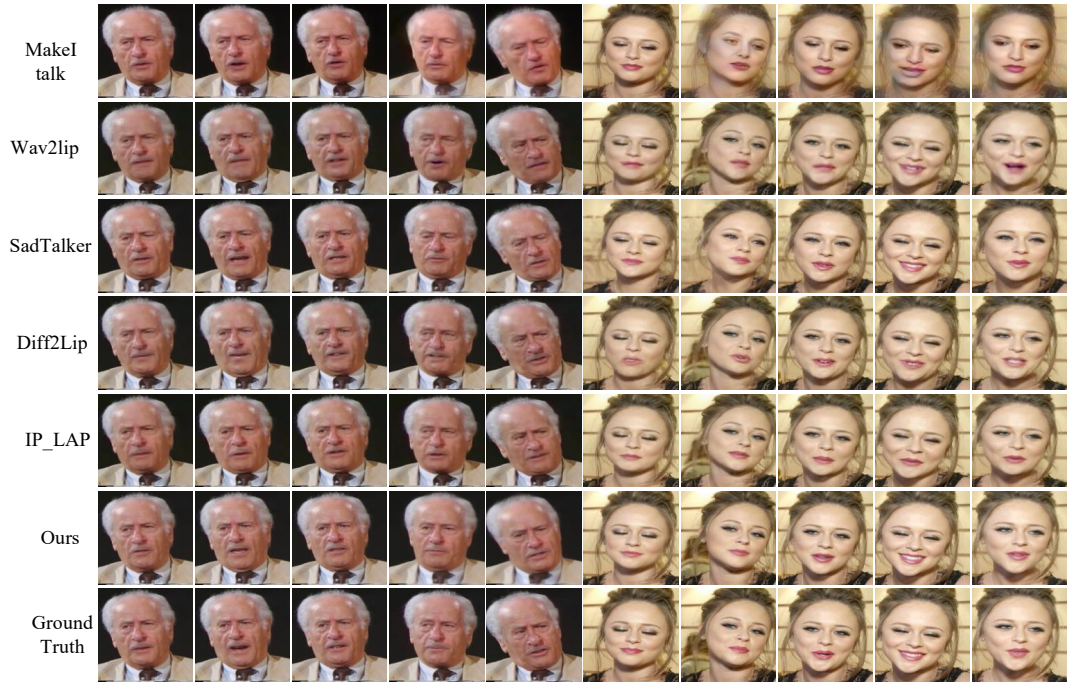

图4　在VoxCeleb1数据集上和其他方法定性结果

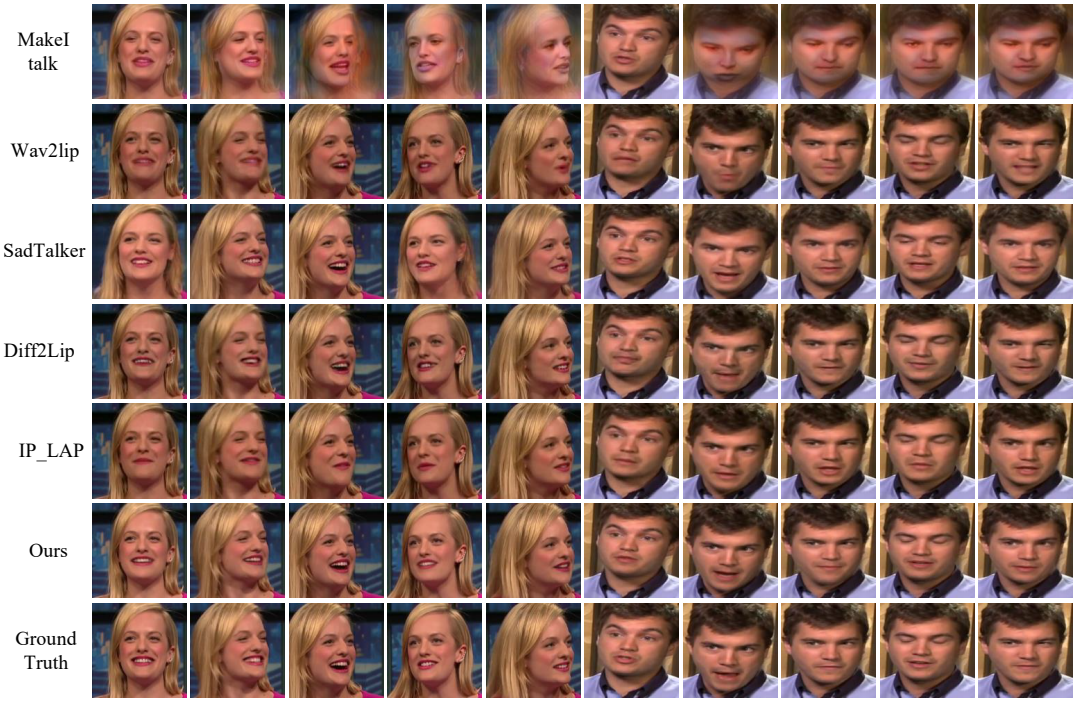

图5　在VoxCeleb2数据集上和其他方法定性结果

制完全使用特征增强,虽然能够引入更多信息,但会造成训练过程不稳定,降低渲染网络的性能. 相比之下,当σ设为可学习参数时,网络能够在[0,1]区间内根据输入自适应分配不同分支的权重,从而在保证训练稳定性的同时有效提升渲染质量和几何一致性. 这表明所提出的自适应缩放机制能够增强模型的泛化能力和鲁棒性.

表6给出了在不同特征尺度上使用多尺度自适应融合模块的对比结果,从结果可以看出,仅在底层特征(48×48)引入时能够提供一定的全局几何约束,但缺乏细节刻画,导致SSIM与LMD表现不佳;仅在高层特征(192×192)引入时可以增强局部纹理与细节表现,但对

表3 在LRW数据集上研究语音预测3D表情系数消融实验

方法	AED ↓	LMD ↓	Sync ↑
基本模型	0.096	2.71	4.55
w/o.语义图像信息	0.094	2.64	5.44
w/o.跨模态注意力融合	0.082	2.12	5.81
w/o.形状约束损失	0.089	2.34	5.66
Full-model	**0.081**	**2.09**	**5.83**

注：↑表示指标越高性能越好，↓表示指标越低性能越好.加粗数字表示该表格中指标最佳值.

表4 在Voxceleb1数据集上研究3D人脸引导的运动-纹理协同渲染网络消融实验

方法	PSNR ↑	LPIPS ↓	SSIM ↑	LMD ↓	CSIM ↑
基本模型	37.33	0.054	0.939	1.91	0.99
w/o. 特征增强网络	37.85	0.049	0.944	1.74	0.99
w/o. 对应一致性损失	38.11	0.044	0.949	1.68	0.99
w/o. 多尺度自适应融合模块	38.21	0.041	0.951	1.66	0.99
w/o. 分阶段训练策略	37.94	0.044	0.942	1.76	0.99
Full-model	**38.31**	**0.040**	**0.954**	**1.62**	**0.99**

注：↑表示指标越高性能越好，↓表示指标越低性能越好.加粗数字表示该表格中指标最佳值.

表5 在Voxceleb1数据集上研究多尺度自适应融合模块缩放系数α消融实验

方法	PSNR ↑	LPIPS ↓	SSIM ↑	LMD ↓	CSIM ↑
σ=0	37.33	0.054	0.939	1.91	0.99
σ=1	37.94	0.044	0.942	1.76	0.99
σ可学习	**38.31**	**0.040**	**0.954**	**1.62**	**0.99**

注：↑表示指标越高性能越好，↓表示指标越低性能越好.加粗数字表示该表格中指标最佳值.

整体结构约束不足，几何精度仍有限. 相比之下，在底层与高层特征同时使用该模块时，PSNR、LPIPS、SSIM及LMD等指标均取得最优性能，表明多尺度特征的互补性在保证全局结构一致性的同时有效提升了细节保真度，从而显著增强了渲染网络的整体性能与稳定性.

表6 在Voxceleb1数据集上研究在不同特征尺度使用多尺度自适应融合模块消融实验

方法	PSNR ↑	LPIPS ↓	SSIM ↑	LMD ↓	CSIM ↑
仅在高层特征使用	38.20	0.044	0.951	1.65	0.99
仅在底层特征使用	38.17	0.043	0.950	1.66	0.99
全部尺度使用	**38.31**	**0.040**	**0.954**	**1.62**	**0.99**

注：↑表示指标越高性能越好，↓表示指标越低性能越好.加粗数字表示该表格中指标最佳值.

图6给出了不同网络配置下的人脸生成结果. 从结果可以看出，基础模型能够完成缺损区域的人脸修复，但由于对参考人脸图像特征的利用不足，生成的人脸纹理细节较为模糊，难以保持身份一致性. 当去除运动特征增强而仅依赖流预测模型时，网络能够利用参考人脸的部分纹理信息实现局部补全，但其学习到的流场主要局限于局部形变，导致口唇区域的动态表现和整体纹理的一致性仍然不足. 相比之下，引入运动特征增强后，模型能够学习全局非刚性形变，更加充分和有效地利用参考人脸图像特征，从而在口唇运动、牙齿结构及面部纹理等方面获得更自然清晰的生成结果. 实验表明，运动特征增强网络在提升全局结构建模能力的同时，也显著增强了修复网络的纹理细节生成能力，有效提高了渲染结果的质量与稳定性.

从图7的实验结果可以观察到，LMD随着帧序列的变化而呈现显著波动. 当头部发生较大幅度运动时，关键点间距离显著增加，反映了人脸几何结构的不稳定性. 这类几何扰动通常会对生成图像的结构一致性与细节还原造成不利影响. 然而，通过SSIM的变化趋势可以看出，即使在关键点距离增大的情况下，生成图像的SSIM值依然保持在0.96以上，整体处于较高水平.特别是右图所示的散点相关分析表明，LMD与SSIM存在一定的负相关(Pearson相关系数约为–0.796)，即关键点偏差的增加可能导致图像结构相似性略有下降，但整体幅度较小. 综上所述，该结果表明：尽管头部姿态变化会对生成图像的几何一致性产生一定影响，但本文提出的方法在跨姿态条件下依然能够维持较高的图像质量，尤其是在SSIM指标维持高水平的情况下，说明模型在运动适应性与纹理保持方面具有良好的鲁棒性与稳定性.

为评估所提出语义先验驱动的多阶段音视频人脸生成框架的计算复杂度，对系统中4个主要子网络的参数规模及计算量进行了统计分析. 4个子网络分别为：跨模态注意力融合的语义先验增强3DMM表情系数预测网络，语义引导的运动流场建模网络，运动感知纹理调制网络，以及参考驱动的人脸修复网络. 各子网络的复杂度指标如表7所示. 由表7可见，整体框架参数总量约为1.01×10^8，计算复杂度约为8.69×10^{10} FLOPs，显存占用约为690 MB. 其中，运动感知纹理调制网络与参考驱动的人脸修复网络为主要计算负载，分别占总FLOPs的约45%和36%；语义引导的运动流场建模网络计算量最小，仅占约3%；跨模态注意力融合的语义先验增强3DMM表情系数预测网络虽参数量较小，但通过跨模态注意力实现语音与视觉语义的高效对齐，对表情预测的准确性具有关键作用.

本方法在修复过程中仍存在一定局限性. 由于网络主要依赖静态帧级特征进行修复，缺乏对时序信息的充分建模，当输入视频中存在快速表情变化或大幅度头部运动时，修复结果的稳定性会受到影响. 如图8所示，在一些连续帧中，裁剪的人脸区域由于唇部运动幅度较大而出现局部偏差，导致渲染网络在修复过程中

图6　不同配置下人脸生成结果

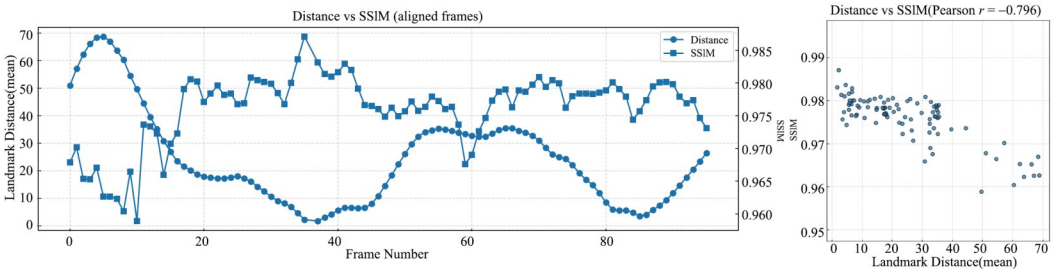

图7　帧序列中人脸关键点距离与SSIM的相关性分析

表7　各子网络计算复杂度统计

网络	Params/M	FLOPs/G	Madd/G	Memory/MB
跨模态注意力融合的语义先验增强3DMM表情系数预测网络	12.50	14.30	28.60	120.0
语义引导的运动流场建模网络	38.68	2.27	5.91	17.9
运动感知纹理调制网络	12.53	39.23	78.39	267.2
参考驱动的人脸修复网络	37.68	31.08	62.08	285.0
整个模型	101.17	86.88	174.98	690.2

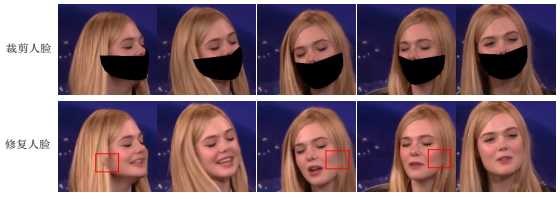

图8　多人脸先验引导的信息融合渲染网络修复失败的案例展示

生成不一致的纹理补丁,图中红色方框处出现伪影. 由于缺少跨帧约束机制,这种不稳定修复在局部区域容易产生伪影,从而影响整体视频的一致性与自然感知.

5　结论

本文提出了一种多元先验融合的运动鲁棒说话人无关视觉配音方法,以3DMM作为中间表示,将任务分解为跨模态语音驱动的3D人脸表情预测与多源先验引导的人脸修复两部分. 该方法通过引入语义图像序列实现语音与视觉特征跨模态融合,提高了音视频同步性能;通过结合三维几何约束与参考人脸纹理,实现了在头部运动场景下的稳定修复和细节保持. 实验结果表明,所提方法在音视频同步精度和人脸图像质量方面具有一定优势,对提升编辑后说话人脸视频的保真度具有积极作用.

参考文献

[1] LIU M Y, HUANG X, YU J, et al. Generative adversarial networks for image and video synthesis: Algorithms and applications[J]. Proceedings of the IEEE, 2021, 109(5): 839-862.

[2] CHEN L L, MADDOX R K, DUAN Z Y, et al. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 7824-7833.

[3] ESKIMEZ S E, MADDOX R K, XU C, et al. Noise-resilient training method for face landmark generation from speech[J]. IEEE/ACM Transactions on Audio, Speech, and

Language Processing, 2019, 28: 27-38.
[4] VOUGIOUKAS K, PETRIDIS S, PANTIC M. Realistic speech-driven facial animation with GANs[J]. International Journal of Computer Vision, 2019, 128(5): 1398-1413.
[5] CHEN L L, CUI G F, LIU C L, et al. Talking-head generation with rhythmic head motion[C]//Computer Vision - ECCV 2020. Cham: Springer, 2020: 35-51.
[6] DAS D, BISWAS S, SINHA S, et al. Speech-driven facial animation using cascaded GANs for learning of motion and texture[C]//Computer Vision - ECCV 2020. Cham: Springer, 2020: 408-424.
[7] PRAJWAL K R, MUKHOPADHYAY R, NAMBOODIRI V P, et al. A lip sync expert is all you need for speech to lip generation in the wild[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 484-492.
[8] PARK S J, KIM M, HONG J, et al. SyncTalkFace: Talking face generation with precise lip-syncing via audio-lip memory[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(2): 2062-2070.
[9] WANG J D, QIAN X Y, ZHANG M L, et al. Seeing what you said: Talking face generation guided by a lip reading expert[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 14653-14662.
[10] ZHANG W X, CUN X D, WANG X, et al. SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 8652-8661.
[11] EGGER B, SMITH W A, TEWARI A, et al. 3D morphable face models—Past, present, and future[J]. ACM Transactions on Graphics (ToG), 2020, 39(5): 1-38.
[12] MUKHOPADHYAY S, SURI S, GADDE R T, et al. Diff2Lip: Audio conditioned diffusion models for lip-synchronization[C]//2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2024: 5280-5290.
[13] XIE T Y, LIAO L C, BI C, et al. Towards realistic visual dubbing with heterogeneous sources[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 1739-1747.
[14] ZHONG W Z, FANG C W, CAI Y Q, et al. Identity-preserving talking face generation with landmark and appearance priors[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 9729-9738.
[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2023-08-02)[2025-10-01]. https://arxiv.org/abs/1706.03762.
[16] CHENG K, CUN X D, ZHANG Y, et al. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild[C]//SIGGRAPH Asia 2022 Conference Papers. New York: ACM, 2022: 1-9.
[17] SUN Y S, ZHOU H, WANG K, et al. Masked lip-sync prediction by audio-visual contextual exploitation in transformers[C]//SIGGRAPH Asia 2022 Conference Papers. New York: ACM, 2022: 1-9.
[18] SONG L, WU W, FU C, et al. Audio-driven dubbing for user generated contents via style-aware semi-parametric synthesis[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(3): 1247-1261.
[19] HUANG X, BELONGIE S. Arbitrary style transfer in real-time with adaptive instance normalization[C]//2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 1510-1519.
[20] YANG S L, WANG W, LING J, et al. Context-aware talking-head video editing[C]//Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 7718-7727.
[21] GUAN J Z, ZHANG Z W, ZHOU H, et al. StyleSync: High-fidelity generalized and personalized lip sync in style-based generator[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 1505-1515.
[22] KI T, MIN D C. StyleLipSync: Style-based personalized lip-sync video generation[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2024: 22784-22793.
[23] GUAN J Z, XU Z L, ZHOU H, et al. Resyncer: Rewiring style-based generator for unified audio-visually synced facial performer[EB/OL]. (2024-08-06)[2025-10-01]. https://arxiv.org/abs/2408.03284.
[24] ZHANG L H, LIANG S, GE Z P, et al. PersonaTalk: Bring at-tention to your persona in visual dubbing[EB/OL]. (2024-09-09)[2025-10-01]. https://arxiv.org/abs/2409.05379.
[25] PAYSAN P, KNOTHE R, AMBERG B, et al. A 3D face model for pose and illumination invariant face recognition[C]//Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance. New York: ACM, 2009: 296-301.

[26] CAO C, WENG Y, ZHOU S, et al. Facewarehouse: A 3d facial expression database for visual computing[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 20(3): 413-25.

[27] 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]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2020: 285-295.

[28] RAMAMOORTHI R, HANRAHAN P. An efficient representation for irradiance environment maps[C]//Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. New York: ACM, 2001: 497-500.

[29] JADERBERG M, SIMONYAN K, ZISSERMAN A. Spatial transformer networks[EB/OL]. (2016-02-04) [2025-10-02]. https://arxiv.org/abs/1506.02025.

[30] PARK T, LIU M Y, WANG T C, et al. Semantic image synthesis with spatially-adaptive normalization[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 2332-2341.

[31] 姜文涛, 高原, 袁姮, 等. 门控机制的图像分类网络[J]. 电子学报, 2024, 52(7): 2393-2406.
JIANG W T, GAO Y, YUAN H, et al. Image classification network of gating mechanism[J]. Acta Electronica Sinica, 2024, 52(7): 2393-2406. (in Chinese)

[32] GENG D, HAMILTON M, OWENS A. Comparing correspondences: Video prediction with correspondence-wise losses[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 3355-3366.

[33] JOHNSON J, ALAHI A, LI F F. Perceptual losses for real-time style transfer and super-resolution[C]//Computer Vision - ECCV 2016. Cham: Springer, 2016: 694-711.

[34] CHUNG J S, ZISSERMAN A. Lip reading in the wild[M]// Computer Vision - ACCV 2016. Cham: Springer International Publishing, 2017: 87-103.

[35] NAGRANI A, CHUNG J S, XIE W, et al. Voxceleb: Large-scale speaker verification in the wild[J]. Computer Speech & Language, 2020, 60: 101027.

[36] CHUNG J S, NAGRANI A, ZISSERMAN A. VoxCeleb2: Deep speaker recognition[EB/OL]. (2018-06-27) [2025-09-20]. https://arxiv.org/abs/1806.05622.

[37] SIAROHIN A, LATHUILIÈRE S, TULYAKOV S, et al. First order motion model for image animation[EB/OL]. (2020-10-01)[2025-10-01]. https://arxiv.org/abs/2003.00196.

[38] BULAT A, TZIMIROPOULOS G. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230, 000 3D facial landmarks)[C]//2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 1021-1030.

[39] LIU L Y, JIANG H M, HE P C, et al. On the variance of the adaptive learning rate and beyond[EB/OL]. (2021-10-26)[2025-10-01]. https://arxiv.org/abs/1908.03265.

[40] ZHANG M R, LUCAS J, HINTON G, et al. Lookahead optimizer: K steps forward, 1 step back[EB/OL]. (2019-12-03)[2025-10-20]. https://arxiv.org/abs/1907.08610.

[41] YI R, YE Z P, ZHANG J Y, et al. Audio-driven talking face video generation with learning-based personalized head pose[EB/OL]. (2020-03-05) [2025-10-20]. https://arxiv.org/abs/2002.10137.

[42] ZHOU Y, HAN X, SHECHTMAN E, et al. MakeltTalk: speaker-aware talking-head animation[J]. ACM Trans Graph, 2020, 39(6): 1-15.

[43] ZAKHAROV E, SHYSHEYA A, BURKOV E, et al. Few-shot adversarial learning of realistic neural talking head models[C]//2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2020: 9458-9467.

作者简介

陈燚雷　男,1992年3月出生于湖北省武汉市. 现为湖北经济学院数字金融创新湖北省重点实验室研究员. 主要研究方向为计算机视觉和少样本人脸视频生成.
E-mail: 290306@whut.edu.cn

熊盛武　男, 1966年11月出生于湖北省咸宁市. 现为武汉理工大学计算机科学与人工智能学院及武汉学院跨学科人工智能研究所教授. 主要研究方向为智能计算、机器学习和模式识别.
E-mail: xiongsw@whut.edu.cn