计算机科学
COMPUTER SCIENCE

基于注意力机制的音频驱动数字人脸视频生成方法

郭星星, 肖雁南, 温佩芝, 徐智, 黄文明

引用本文

郭星星, 肖雁南, 温佩芝, 徐智, 黄文明. 基于注意力机制的音频驱动数字人脸视频生成方法[J]. 计算机科学, 2026, 53(2): 245-252.

GUO Xingxing, XIAO Yannan, WEN Peizhi, XU Zhi, HUANG Wenming. Attention-based Audio-driven Digital Face Video Generation Method [J]. Computer Science, 2026, 53(2): 245-252.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241200067

基于注意力机制的音频驱动数字人脸视频生成方法

郭星星[1,2]　肖雁南[1,2]　温佩芝[1,2,3]　徐　智[1,2]　黄文明[1,2,3]
1 桂林电子科技大学计算机与信息安全学院　广西 桂林 541004
2 广西图像图形与智能处理重点实验室　广西 桂林 541004
3 桂林信息科技学院信息工程学院　广西 桂林 541004
(guoxingxing0328@163.com)

摘　要　音频驱动数字人脸视频生成的难点问题在于,如何将音频与视频两种不同模态的信息对齐,从而实现唇音同步。现有技术大多基于英文数据集开发,由于中文发音与英文发音存在差异性,直接将这些技术运用于中文音频驱动数字人脸视频生成时,存在牙齿模糊和视频清晰度不够的问题。基于GAN框架,提出了一种基于注意力机制的音频驱动数字人脸视频生成方法M-CSAWav2Lip。将MFCC和Mel Spectrogram融合,实现音频特征提取。利用MFCC的时间动态特性和Mel Spectrogram的频率分辨能力,全面捕捉语音信息的细微变化。在数字人脸生成过程中,采用基于注意力机制及残差连接的网络架构,通过加权通道和空间注意力机制强化特征的重要性,提高关键音频和视频特征的获取能力,实现有效编码和融合中文音视频信息,生成与语音内容相匹配的唇部动作和面部视频。最后,在自建的中文数据集及通用数据集上进行训练与测试。实验结果表明,所提方法生成的唇音同步数字人脸视频在精度和质量方面均有一定的提升。

关键词:音频驱动;唇音同步;音频特征提取;数字人脸生成;注意力机制
中图分类号　TP391

Attention-based Audio-driven Digital Face Video Generation Method

GUO Xingxing[1,2],XIAO Yannan[1,2],WEN Peizhi[1,2,3],XU Zhi[1,2] and HUANG Wenming[1,2,3]
1 School of Computer and Information Security,Guilin University of Electronic Technology,Guilin,Guangxi 541004,China
2 Guangxi Key Laboratory of Image and Graphics Intelligent Processing,Guilin,Guangxi 541004,China
3 School of Information Engineering,Guilin University of Information Technology,Guilin,Guangxi 541004,China

Abstract　The key challenge in audio-driven digital face video generation lies in aligning the information from two different modalities,audio and video,to achieve lip synchronization. Existing technologies have primarily been developed using English datasets. However,due to the phonetic differences between Chinese and English,directly applying these methods to Chinese audio-driven face video generation results in issues such as blurred teeth and insufficient video clarity. This paper proposes M-CSA-Wav2Lip,an audio-driven digital face video generation method based on a GAN framework and enhanced by an attention mechanism. The method combines MFCC and Mel Spectrograms for audio feature extraction. By leveraging the temporal dynamics of MFCC and the frequency resolution of Mel Spectrograms,the method captures subtle variations in speech information comprehensively. During the digital face generation process,a network architecture based on attention mechanisms and residual connections is employed. This architecture uses weighted channel and spatial attention mechanisms to enhance the importance of features,improving the ability to extract key audio and video features. This allows for the effective encoding and fusion of Chinese audio-video information,generating lip movements and facial videos that are consistent with the audio content. Finally,the model is trained and tested on both a custom Chinese dataset and a general dataset. Experimental results demonstrate that the generated lip-synced digital face videos show improvements in both accuracy and quality.

Keywords　Audio-driven,Lip synchronization,Audio feature extraction,Digital face generation,Attention mechanism

到稿日期:2024-12-09　返修日期:2025-03-08
基金项目:广西图像图形与智能处理重点实验室开放基金(GIIP2310);广西自然科学基金(2020GXNSFAA297186)
This work was supported by the Guangxi Key Laboratory of Image and Graphic Intelligent Processing Foundation Proiect(GIIP2310) and Guangxi Natural Science Foundation,China(2020GXNSFAA297186).
通信作者:肖雁南(xiaoyn@guet.edu.cn)

1 引言

随着计算机视觉和深度学习技术的迅速发展，音频驱动的数字人脸视频生成技术日趋成熟。其通过将目标人物的脸部图像和音频结合，生成唇形与音频高度同步的视频，极大地提高了数字人的真实感，甚至达到“以假乱真”的视觉效果。该技术可以大大提高元宇宙等虚拟现实内容的制作生产效率，在影视产业、数字媒体、虚拟播报、远程教育、智能服务等场景得到应用推广[1]，可以节省大量的人力和物力，具有广阔的应用前景。

目前，数字人脸视频生成技术主要依赖生成对抗网络(Generative Adversarial Network，GAN)来实现唇部动作与音频内容的高度匹配，但在面对复杂语音信号时，无法充分捕捉音频中的时域和频域的细节信息，导致语音与唇部动作之间的映射存在一定的偏差和局限性。人们对异步语音和面部动作细微变化的敏感性，以及语音对唇部运动的依赖，导致实现输入语音与合成视频流在时间上的精确对齐仍然是数字人脸视频生成研究中极具挑战性的问题。

Chen 等[2]提出的 ATVGnet 采用动态可调的像素级损失函数，并提出了一种基于回归的判别器网络来改善视听同步。然而，该模型在复杂语音场景中仍会出现唇音同步不准确的情况。Prajwal 等[3]提出的 Wav2Lip 模型在生成器的末端引入辅助嵌入网络分析视听一致性，并通过 GAN 实现唇部动作与音频内容的高度匹配。但运用于中文语音驱动时，该模型存在牙齿模糊和视频清晰度不够的问题。Zhou 等[4]提出的 PC-AVS 采用一种低维姿态编码设计，将视听信息分解为身份特征、语音内容和姿势空间，从而生成姿势可控且具有精确唇音同步的数字人脸视频。但其泛化能力有限，在训练数据不足的情况下生成的视频质量明显下降。Guo 等[5]提出的 AD-NeRF 通过端到端的方式将音频特征直接映射到动态神经辐射场进行肖像渲染，避免了中间信息的丢失。但由于训练数据和实际应用中语言不匹配，该模型生成的唇部动作不自然。Mukhopadhyay 等[6]提出的 Diff2Lip 架构则是基于扩散模型，根据音频特征逐步更新面部图像的每一帧，在一定程度上提升了动态细节的生成能力。但在复杂场景下，该模型很难复现面部的动态细节。

本文提出了一种融合注意力机制的音频驱动数字人脸视频生成模型——M-CSAWav2Lip(Mel Frequency Cepstral Coefficients-Channel Spatial Attention Mechanism Wav2Lip)。首先，为提升模型在中文环境下的性能，构建了针对中文音频处理的数据集 CLR(Chinese Lip Reading)。然后，在音频特征提取方面，将梅尔频谱倒谱系数(Mel Frequency Cepstral Coefficients，MFCC)[7]和梅尔频谱(Mel Spectrogram)[8]融合，充分利用时频特征和音频特性，以提供更全面的音频特征，实现音频信号与唇部动作的精准同步。最后，引入基于残差连接的注意力网络架构，通过通道和空间注意力机制[9]加权通道增强重要特征的比重，希望进一步提升模型对关键音视频特征的处理能力，实现中文音频信息的有效编码和融合。

本文的主要贡献如下：

1)构建了中文数据集 CLR，提出一种端到端的模型，采用更精细的音频预处理技术，无需任何中间表示或情感标签，即可生成富有表现力的人脸视频；

2)结合 MFCC 的时间动态特性和 Mel Spectrogram 的频率分辨率，捕捉语音信息的细微变化，实现语音信号与视频中唇部动作的同步；

3)采用通道空间注意力机制，有效提升模型对关键唇部动作和面部变化的识别能力，同时提高处理效率；

4)大量实验结果表明，本文提出的模型能够生成音唇动作同步且精确清晰的数字人脸视频。

2 相关工作

2.1 音频特征提取方式

音频特征提取是语音信号处理的关键环节，目的是将音频信号转化为可供计算机处理的数字特征。Mel Spectrogram 作为一种广泛应用于语音处理和音频分析的特征表示方法，能够有效捕捉语音信号的频率特征。人耳对频率的感知呈非线性特征，在低频部分表现出较高的敏感性，而对高频的感知能力相对较弱。Mel Spectrogram 的设计便采用了梅尔刻度对频率轴进行非线性映射，使低频部分具有较高的分辨率，高频部分的分辨率降低，增强对低频部分特征的辨别能力，从而更有效地捕捉信号的频率特征。

MFCC 作为另一种基于人耳听觉特性的非线性音频特征提取方法，结合了梅尔频率尺度和倒谱分析的优点，进一步提升了频率特征的提取精度。通过梅尔滤波器组将音频信号映射到梅尔频率轴上，可以更好地模拟人耳对不同频率的感知过程。接着，采用离散余弦变换(Discrete Cosine Transform，DCT)对梅尔频谱进行处理，去除频谱的冗余信息，保留最关键的倒谱特征，可以在增强语音信号表达能力的同时，更有效地捕捉语音的细节变化和动态特性。

2.2 注意力机制

通道注意力机制[10]通过对特征图中的不同通道进行加权，强化对重要通道的关注，减少不相关通道的干扰，从而提升模型的感知能力。其首先通过全局平均池化或全局最大池化操作，压缩输入特征图的空间信息，保留每个通道的全局描述，更好地捕捉到整体特征；然后利用全连接层或卷积层学习每个通道的重要性，生成相应的通道权重系数；最后依据学习到的权重对各通道进行加权，使得重要通道的响应值得到增强，突出与嘴唇运动和面部表情变化相关的音频特征。

空间注意力机制[11]侧重于关注特征图中重要的空间位置，提升关键区域的特征表达。其通过对输入特征图的空间维度信息进行压缩，生成空间注意力图，并利用该图对特征图进行加权，增强关键区域的特征表达。在面部视频生成过程中，空间注意力机制能够聚焦于嘴巴、眼睛等关键面部区域，增强这些区域的细节表现，提高面部表情的真实感。

3 音频驱动的唇音同步数字人脸视频生成方法

本文提出的 M-CSAWav2Lip 模型由生成器和判别器两

部分构成。生成器接收给定的面部图像帧（用于提供身份和姿势参考）以及语音信号（用于唇部动作参考）作为输入，负责生成对应的说话视频。判别器部分，唇形同步判别器通过同步损失函数来惩罚不准确的唇部动作，提高视频的精确性。视觉质量判别器的引入，旨在进一步提升生成视频的视觉质量。模型的具体架构如图 1 所示。

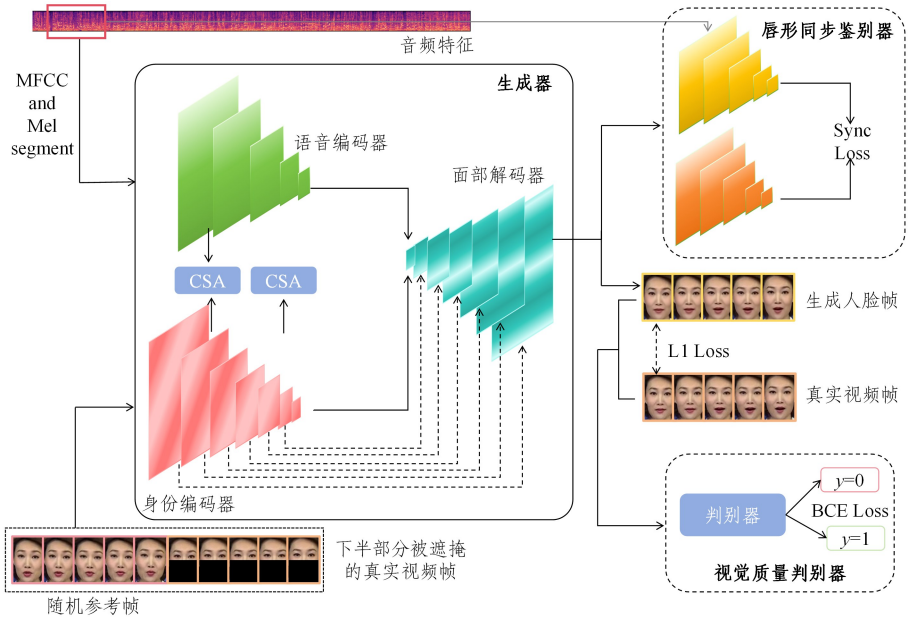

图 1　M-CSAWav2Lip 模型架构图

Fig. 1　Architecture diagram of M-CSAWav2Lip model

3.1　视频生成器的选择

在 Wav2Lip 模型[3]的基础上对生成器进行改进。采用二维卷积神经网络（2D-CNN）的编码器-解码器架构，其中包含 3 个模块：身份编码器、语音编码器和面部解码器。身份编码器通过残差卷积层将随机选择的参考帧与姿势先验帧结合。语音编码器采用 2D 卷积处理音频信号，将处理后的音频特征与面部特征融合。同时，语音编码器中不仅采用了改进的音频特征提取方法提高音频特征的准确性，而且引入通道空间注意力机制提升模型对关键音频特征的提取和表达能力。在身份编码器中引入通道空间注意力机制，增强对当前任务唇部动作和视觉特征的关注。面部解码器利用卷积层和反卷积上采样技术，将身份编码器和语音编码器提取的特征进行融合，输出与音频同步的数字人脸重建图像。

3.1.1　音频与唇部动作的对齐

通过充分利用 Mel Spectrogram 在捕捉频率分布和时间变化方面的优势，以及 MFCC 在表达音色和发音特性上的优势，进行音频特征的提取。这种特征融合策略可增强语音编码器在噪声环境下的语音信号识别能力和鲁棒性，确保音频信息在频域和时域上的完整表达。

音频特征提取的具体流程如图 2 所示。首先，通过短时傅里叶变换从原始音频中获取频谱数据 $X(t,f)$。然后，使用梅尔滤波器组 $h_m(f)$ 将频谱数据映射至梅尔尺度，得到梅尔频谱 $M(t,m)$，如式(1)所示：

$$M(t,m)=\sum_f X(t,f)\cdot h_m(f) \tag{1}$$

其中，m 是梅尔频带的索引，t 表示时间索引，f 表示频率。

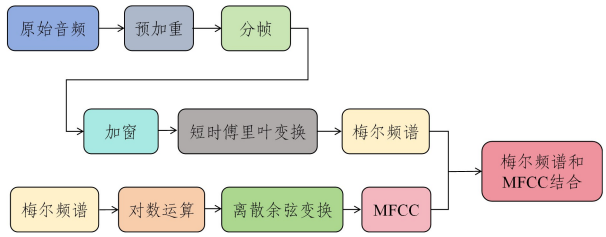

图 2　音频特征提取

Fig. 2　Audio feature extraction

接下来对梅尔频谱值 $M(t,m)$ 进行对数变换，模拟人耳对声音强度的非线性感知。对数变换后的梅尔频谱如式(2)所示：

$$L(t,m)=\log(M(t,m)) \tag{2}$$

随后，采用 DCT 对变换后的梅尔频谱进行处理（见式(3)），从中提取主要的倒谱特征，减少数据的冗余和噪声，保留最关键的倒谱信息。

$$C_k(t)=\sum_{m=1}^{M}L(t,m)\cdot\cos\left[\frac{\pi}{M}\cdot(m-1)\cdot(k-1)\right] \tag{3}$$

其中，$C_k(t)$ 表示时间 t 的第 k 个 MFCC 系数，M 是梅尔频带数量，k 为 MFCC 系数的索引。

最后，将 Mel Spectrogram 和 MFCC 在时间维度上截断，确保两者在时间尺度上对齐。将截断后的 Mel Spectrogram 和 MFCC 特征沿特征轴拼接，即将两者的特征向量按列连接，形成一个综合特征向量，实现时间维度的一致性和对音频特征的全面覆盖。

将融合后的音频特征频谱图与随机图像参考帧及对应的真实帧进行级联。具体而言，将音频特征的频谱图作为额外通道，沿通道维度与视频帧进行拼接，从而将音频和视频的特

征联合编码，生成包含音视频信息的高维特征表示。该级联过程旨在将音频和视频的特征信息融合，确保音频驱动的唇部动作与视频帧中的面部特征精准对齐。在此基础上，通过学习音频特征与唇部动作之间的映射关系，并基于训练数据建立音频与相应唇部动作之间的映射，实现 Mel Spectrogram 和 MFCC 融合后提取的音频特征与生成视频中唇部动作的同步。随后，采用卷积神经网络(Convolutional Neural Network，CNN)对拼接后的特征进行降维处理，生成音视频特征向量，并将这些特征向量映射至唇部运动。为确保音频驱动的唇部运动转化为相应的图像帧，进一步通过一系列转置卷积层处理这些特征向量，并将结果矩阵投影至重建的图像帧片段上，恢复视频细节，确保视频中唇部运动与音频之间精确匹配。最后，在身份编码器和面部解码器之间引入残差跳跃连接，促进信息的有效传递和梯度流动，从而提升模型的稳定性和训练效率。

3.1.2 注意力机制的引入

所提方法在生成器的语音编码器 256 通道和身份编码器 32 以及 256 通道处都引入通道和空间注意力机制来优化特征整合和信息流，从不同维度提升模型对关键信息的捕捉能力。双重注意力机制在音频特征和视频特征的处理过程中发挥了重要作用，使音频信息和视觉信息之间的潜在关联可以得到有效挖掘，进而提高了多模态信息融合的准确性和表达能力。

通道注意力机制采用全局平均池化将输入特征图 $X\in \boldsymbol{R}^{H\times W\times C}$ 压缩至单一空间维度，生成压缩特征 F_{avg}：

$$F_{\text{avg}}=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}X(i,j,c),c=1,2,\cdots,C \tag{4}$$

其中，H 和 W 分别为特征图的高度和宽度，C 为通道数。

通过全局平均池化，特征图在空间维度上被压缩，只保留每个通道的全局信息。接下来，压缩后的特征 F_{avg} 通过卷积压缩层和扩展层进行序列处理，从而生成通道注意力权重 M_c，如式(5)所示：

$$M_c=\varepsilon(\boldsymbol{W}_2\delta(\boldsymbol{W}_1F_{\text{avg}}+b_1)+b_2) \tag{5}$$

其中，$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 为可学习的权重矩阵，b_1 和 b_2 为偏置项，ε 表示 Sigmoid 激活函数，δ 表示 ReLU 激活函数。

空间注意力机制通过评估每个空间位置的重要性，突出图像的关键区域。首先，空间注意力机制对输入特征图 X 在每个空间位置上的通道信息进行最大池化 F_{max}(见式(6))和平均池化 F_{avg}(见式(7))，使特征图在空间维度上保持原始大小，但通道数被压缩为 1。接着，将最大池化和平均池化后的特征图进行拼接，形成合并特征 F_{conc}(见式(8))。合并后的特征图 F_{conc} 通过卷积层进行处理，生成空间注意力权重图 M_s(见式(9))。

$$F_{\text{max}}=MaxPool(X) \tag{6}$$

$$F_{\text{avg}}=AvgPool(X) \tag{7}$$

$$F_{\text{conc}}=Conc(F_{\text{max}},F_{\text{avg}}) \tag{8}$$

$$M_s=\varepsilon(Conv(F_{\text{conc}})) \tag{9}$$

其中，$Conc(F_{\text{max}},F_{\text{avg}})$ 表示对 F_{max} 和 F_{avg} 进行拼接操作，$Conv()$ 表示卷积操作。

本文同时采用通道注意力机制 M_c 与空间注意力机制 M_s，通过在不同维度上的加权和调整，确保音频和视觉信息的无缝对接。具体实施过程如图 3 所示。首先对输入的特征图采用通道注意力机制，对各个通道的重要性进行加权调整；接着将处理后的特征图输入空间注意力机制，对空间维度上的关键区域进行加权增强，最终得到优化后的特征图。在此基础上，生成器采用损失函数来衡量生成视频帧和真实视频帧之间的像素级差异。对于给定的 N 对生成帧 L_g 与真实帧 L_G，重建损失函数 L_{recon} 为：

$$L_{\text{recon}}=\frac{1}{N}\sum_{i=1}^{N}\parallel L_g-L_G\parallel \tag{10}$$

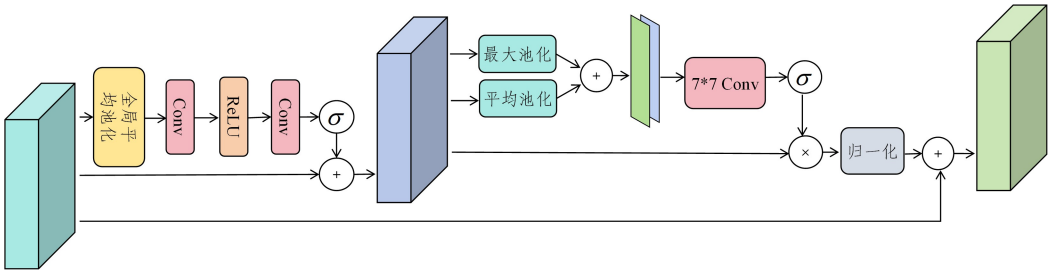

图 3 通道空间注意力机制

Fig. 3 Channel spatial attention mechanisms

3.2 生成视频的判别

判别器作为 GAN 的组成部分之一，旨在对生成的视频帧与真实视频帧进行校对，使生成器产生更加逼真和自然的视频输出。在这个过程中，判别器接收生成帧和真实帧作为输入，通过多层卷积网络提取面部特征，进而评估嘴唇运动是否与输入的音频信号紧密同步，确保视频和视觉内容的一致性。判别器分为唇形同步判别器和视频质量判别器。

3.2.1 唇形同步判别器

唇形同步判别器基于 SyncNet[12] 改进而来，首先采用 Mel Spectrogram 和 MFCC 融合的方式提取音频特征，与生成的视频帧进行融合处理。相比单一的特征提取方法，该方法能够更全面地捕捉音频的细节特征，提升在复杂音视频环境下模型的灵敏度和鲁棒性。

唇形同步判别器由 2D 卷积网络构成。音频信号首先通过 Mel Spectrogram 和 MFCC 进行预处理，从中提取音频特征；视频帧[13]则通过 CNN 提取特征，生成视频向量。接着使用全连接层将音频向量 $\boldsymbol{a}$ 和视频向量 $\boldsymbol{v}$ 映射到相同的维度 D_{sync}，确保两者具有相同的特征维度。然后计算音频向量和

视频向量之间的点积，评估音视频之间的同步性。点积计算结果经过余弦相似度归一化，从而量化音视频之间的相似度，并为每个样本生成音频-视频对同步的概率值 P_{acc}，如式(11)所示。

$$P_{acc}=\frac{\boldsymbol{v}\cdot\boldsymbol{a}}{\max(\|\boldsymbol{v}\|_2\cdot\|\boldsymbol{a}\|_2,\varepsilon)} \tag{11}$$

其中，ε 表示 Sigmoid 激活函数，将音频和视频向量的点积结果映射到[0,1]范围内。

最后，利用二元交叉熵损失函数[3]对模型进行优化，确保生成视频中的音视频同步更加精确。该判别器能够对不同时间偏移下的同步性进行稳定评估，识别微小的同步误差，从而提升视频生成质量的整体真实感。

训练期间，为确保唇形同步判别器和生成器的效果对齐，两者需要处理 5 个连续帧[3]。通过最小化唇形同步判别器提供的同步损失函数来提高生成器生成帧的唇形同步质量，确保生成的视频帧在唇音同步方面达到更高的质量。同步损失函数 L_{sync} 如式(12)所示：

$$L_{sync}=\frac{1}{N}\sum_{i=1}^{N}-\log(P_{acc}^{i}) \tag{12}$$

其中，P_{acc} 由式(11)计算得到，N 表示生成帧的数量。在生成器的训练过程中，唇形同步判别器的权重保持不变。

3.2.2 视觉质量判别器

除了音唇同步精度外，生成视频的视觉质量同样至关重要。视觉质量判别器旨在确保生成的视频帧在视觉上更自然且无明显的生成痕迹，不对唇形同步做任何更改。该判别器由多个卷积块堆叠而成，每个卷积块包含一个卷积层、一个实例归一化层和一个 Leaky ReLU 激活函数，后两者对训练的稳定性至关重要。生成器的目标函数 L_{gen} 旨在提升生成视频的质量，如式(13)所示。判别器的训练目标是最大化目标函数 L_{disc}，如式(14)所示。

$$L_{gen}=E_{x'\sim L_g}[\log(1-D(x'))] \tag{13}$$

$$L_{disc}=E_{x\sim L_G}[\log(D(x))]+L_{gen} \tag{14}$$

其中，E 表示期望，用于计算生成器和判别器在训练过程中的期望损失；x'表示生成样本，L_g 是生成器生成的视频帧；x 表示真实样本，L_G 是真实视频帧；$D(x)$ 表示判别器对输入样本 x 为真的概率值；L_{disc} 通过优化判别器，使其能够区分生成帧和真实帧。

生成器的最终优化目标是最小化重建损失函数 L_{recon}、同步损失函数 L_{sync} 和对抗损失函数 L_{gen} 的加权和，如式(15)所示：

$$L_{total}=(1-w_s-w_g)\cdot L_{recon}+w_s\cdot L_{sync}+w_g\cdot L_{gen} \tag{15}$$

其中，w_s 是同步惩罚权重，w_g 是对抗损失权重，根据大量实验结果最终将 w_s 和 w_g 设置为 0.03 和 0.07。

4 实验研究与分析

4.1 实验准备

4.1.1 数据集

采用自建的 CLR 数据集和英国广播公司(BBC)的 LRS2 数据集[14,16]对本文模型进行训练，并将其与其他方法在不同语言和环境条件下的有效性和适应性进行对比和评估。CLR 数据集首先从 CMLR 数据集[17-18]中选取了 11 位新闻联播主播的视频片段，包含 6 位男性和 5 位女性主播，29 198 个中文口语句子，每句不超过 29 个汉字，总计约 14 h。同时，在网络上收集了介绍桂林旅游景点的视频，视频总长约为 1 h，包含 1 459个中文口语句子，每句不超过 20 个汉字。使用 Premiere Pro 剪辑，保证人脸居中且视频长度不超过 3 s。LRS2 数据集包含新闻、访谈等多种节目类型，涵盖数千个口语句子，句子长度不超过 100 个字符。两个数据集的视频帧率统一调整为 25 FPS，音频采样率调整为 16 kHz。

4.1.2 评价指标

数字人生成的视频主要从视觉质量和音唇同步两个方面进行评估。本文采用当今流行的技术指标 LSE-D(Lip Sync Error Distance)和 LSE-C(Lip Sync Error Confidence)[3]来衡量音唇同步的质量。LSE-D 表示唇部图像向量和音频向量之间距离的平均误差度量，LSE-D 值越低，表明音唇同步质量越高。LSE-C 则提供平均置信度得分，LSE-C 值越高，表示音频和视频的相关性越强。视频质量则采用经典的指标 PSNR(Peak Signal to Noise Ratio)[19]和 SSIM(Structural Similarity Index)[20]来评估。PSNR 用于衡量图像间的相似度。SSIM 从亮度、对比度和结构 3 个维度来评价图像的视觉相似性。较高的 PSNR 和 SSIM 均表明生成图像具有较好的视觉质量。

4.1.3 训练环境

模型的训练和验证均在 Linux 环境中进行，使用 NVIDIA RTX A4000 显卡，内存为 16 GB。训练中采用 Adam 优化器[21]优化学习率，加速模型收敛。设置判别器的批处理大小为 32，学习率为 1×10^{-3}，判别器总训练时间约为 24 h；生成器的批处理大小设置为 48，学习率为 1×10^{-4}，训练时长为 96 h。引入通道空间注意力机制后，生成器的训练时间缩短至 72 h 左右，在训练约 30 epochs 后观察到嘴唇开始变形，此时模型已能有效学习并生成准确的面部状态。

4.2 实验结果对比分析

4.2.1 视觉质量对比分析

将本文提出的方法与 ATVGnet，Wav2Lip，PC-AVS，AD-NeRF 和 Diff2Lip 这 5 种 2D 数字人脸视频生成方法进行实验比较，所有方法均在相同的实验配置下进行测试。图 4 展示了其中 2 组生成结果。图 4 中第一行是录制的真实视频帧，第二行是原始的音频信号，其余 6 行分别为采用不同音频驱动方法生成的数字人视频。ATVGnet[2]采用级联策略，通过音频输入预测低维面部关键点，但在快速唇部动作的同步上表现欠佳。Wav2Lip[3]致力于实现与音频精确同步的唇部动作，唇形正确但清晰度不够。PC-AVS[4]分离并重组身份、语音内容和面部姿态，以提高生成视频的清晰度，整体效果较好，但存在同步误差和唇形动作不精确的问题。AD-NeRF[5]最终生成的视频帧唇部周围和脸部有明显的间隙。Diff2Lip[6]采用条件 GAN 确保生成的高一致性，但生成的视频帧存在颗粒度问题，导

致唇形表现不够流畅。从图 4 中放大的嘴部图形可见，本文方法在视频质量和唇音同步方面表现出的唇形动作与真实视频帧的匹配度更高，生成的视频在清晰度上均优于上述方法，整体效果更加流畅。

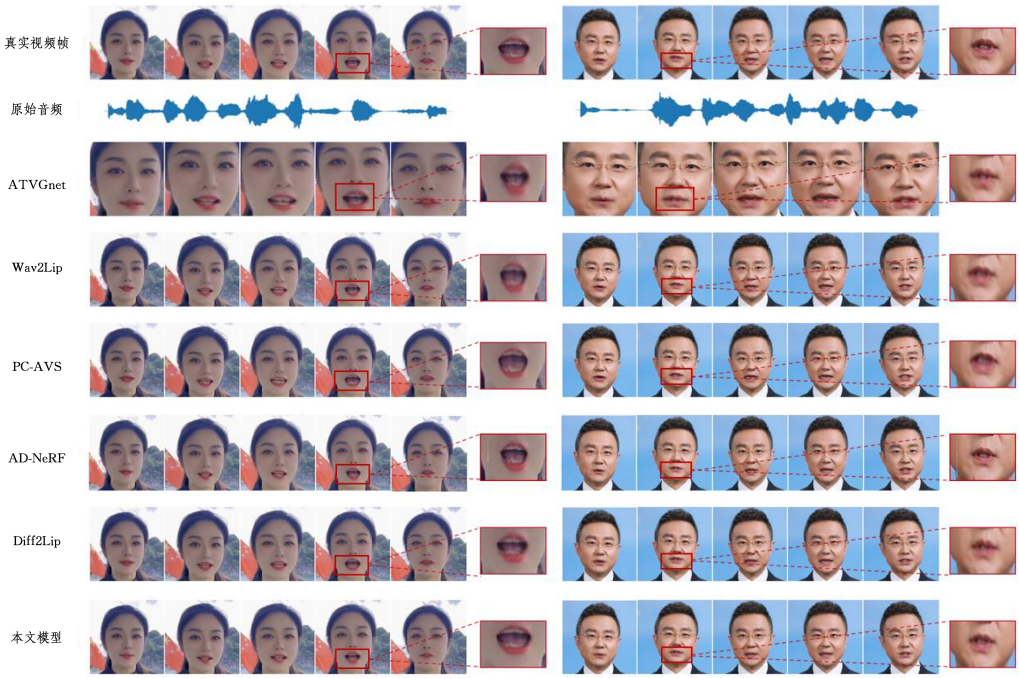

图 4 不同方法生成结果的比较

Fig. 4 Comparison results of different methods

4.2.2 技术指标对比分析

采用不同数据集对各模型的性能进行测试对比，各项评价指标结果如表 1 所列，表中“↓”表示该指标值越低越好，“↑”表示该指标值越高越好。由表 1 可见，在 CLR 数据集上，本文模型在所有评价指标上均优于其他 5 种方法，特别是在 LSE-D 和 LSE-C 指标上较 Wav2Lip 模型分别提高了 1.187和 1.276，相较于其他几种方法也有不同程度的提升，且和真实视频的性能指标较为接近。这是因为本文模型采用中文数据集进行训练，因此在处理中文音频和视频的同步问题上更为精确。相比之下，其他 5 种方法均基于英文数据集进行训练，对中文的处理效果并不理想。在 LRS2 数据集上，Wav2Lip 在 LSE-C 指标上由于直接使用 SyncNet，因此表现较好。本文模型在身份编码器的 256 通道中引入通道空间注意力机制，增强了对音视频同步关键面部区域的关注，改善了视觉连贯性，因此在 SSIM 和 PSNR 指标上优于其他方法，显示出其在视觉细节和图像质量上的优势。本文模型在 LSE-D 指标上也优于 Wav2Lip，原因是其采用的 Mel Spectrogram 和 MFCC 融合的音频特征提取方法有效增强了音频特征的识别能力，提升了音唇同步精度。

表 1 在 CLR 和 LRS2 数据集上的结果对比

Table 1 Comparison results between CLR and LRS2 datasets

方法	CLR				LRS2			
	LSE-D ↓	*LSE-C* ↑	*SSIM* ↑	*PSNR* ↑	*LSE-D* ↓	*LSE-C* ↑	*SSIM* ↑	*PSNR* ↑
真实视频	8.063	6.510	1.000	N. A.	7.934	6.508	1.000	N. A.
ATVGnet(2019)	9.869	4.931	0.792	30.454	8.823	5.584	0.785	30.627
Wav2Lip(2020)	9.505	5.212	0.835	31.274	8.491	**6.362**	0.837	31.454
PC-AVS(2021)	9.524	5.016	0.764	30.090	8.643	6.138	0.792	30.281
AD-NeRF(2021)	9.492	5.003	0.781	30.264	8.631	6.192	0.815	30.761
Diff2Lip(2023)	9.362	5.164	0.816	30.392	8.513	6.298	0.843	30.997
本文模型	**8.318**	**6.488**	**0.890**	**31.594**	**8.364**	6.340	**0.851**	**31.536**

4.2.3 消融实验

为评估音频特征提取方式及通道和空间注意力机制对生成视频质量的影响，本文进行了消融实验，结果如表 2 所列。其中，MFCC-Wav2Lip 代表采用 MFCC 进行音频特征的提取，M-Wav2Lip 表示 Mel Spectrogram 与 MFCC 融合的音频特征提取方法，CAT-Wav2Lip 表示模型仅采用通道注意力机制，SAT-Wav2Lip 指仅采用空间注意力机制，CSA-Wav2Lip 表示模型结合通道和空间注意力机制。

表 2 CLR 上消融实验的对比结果

Table 2 Comparative results of ablation experiments on CLR

方法	*LSE-D* ↓	*LSE-C* ↑	*SSIM* ↑	*PSNR* ↑
Wav2Lip	9.505	5.212	0.835	31.274
MFCC-Wav2Lip	9.357	5.291	0.839	31.498
M-Wav2Lip	8.916	5.549	0.846	32.901
CAT-Wav2Lip	8.748	5.831	0.862	33.179
SAT-Wav2Lip	8.792	5.749	0.870	33.032
CSA-Wav2Lip	8.634	5.891	0.887	33.319
本文模型	**8.318**	**6.488**	**0.890**	**33.594**

由表 2 结果可知，改变音频特征提取方式可以提升模型的性能。单独使用 Mel Spectrogram 时，梅尔滤波器组在降低频率维度过程中会丢失高频细节，限制了模型在精细同步嘴唇动作时对音频细节的分析能力。此外，Mel Spectrogram 难以反映音频中的微妙时间变化，而精细的时间动态变化对于数字人脸视频生成中的唇音同步至关重要。MFCC 能够提供更丰富的语音信息和音色特征，并通过一阶和二阶差分增强对时间变化的捕捉能力，但在频谱信息的表达方面存在局限。因此，结合 Mel Spectrogram 和 MFCC 能够有效弥补 Mel Spectrogram 在捕捉高频细节和时间变化以及 MFCC 在频谱信息表达上的不足。表 2 的结果还显示，本文采用的特征融合策略提升了模型在复杂环境中的同步精度和鲁棒性，增强了音频特征的表达能力，使模型更适合高精度的唇音同步应用。

实验表明，本文采用通道注意力机制强化对关键通道的特征学习和表示，突出了包含丰富信息的特征。引入空间注意力机制聚焦于输入特征图的关键空间位置，强调视频帧中与说话者嘴唇区域相关的部分。两者结合使模型能够在多维度上精细解析数据，特别是在光照条件多变、面部角度及表情持续变化的复杂环境中，能有效提升面部特征的理解和重建能力，增强视频的视觉吸引力和自然真实感。

4.2.4 主观评价

考虑到人们对音唇同步的高敏感性，人工评价在数字人脸视频生成技术中具有一定的实际参考意义。为此，设计了一份评估问卷，邀请 20 名参与者对不同方法在 CLR 数据集上训练生成的数字人脸视频进行主观评价。问卷包括 3 个评价指标：同步精度、视觉质量、整体感观（所生成视频的整体表现）。每个指标的评分范围为 1～5 分，分数越高表明参与者的满意度越高。表 3 所列数据为各项指标的平均得分，结果显示参与者对本文模型生成的视频给予了较高的评价。

表 3 不同音频驱动方法的用户研究

Table 3 User studies of different audio-driven methods

方法	同步精度	视觉质量	整体感官
ATVGnet	2.831	2.935	2.916
Wav2Lip	3.437	3.359	3.408
PC-AVS	3.317	3.108	3.219
AD-NeRF	3.392	3.307	3.165
Diff2Lip	3.419	3.401	3.412
本文模型	**3.921**	**3.516**	**3.734**

结束语 本文提出的基于注意力机制的音频驱动数字人脸视频生成模型 M-CSAWav2lip，旨在实现中文唇型和语音的精确同步，提高数字人脸生成视频的质量。该模型通过构建中文数据集 CLR 进行训练，采用 Mel Spectrogram 和 MFCC 融合的音频特征提取技术，有效提升了视频中关键音频特征的识别精度和非语言声音的处理能力；在关键部位引入通道和空间注意力机制，增强面部特征提取能力，使得生成的数字人脸表情更为连贯和自然。实验结果表明，无论在客观还是主观评价指标上，M-CSAWav2lip 模型均展示出较好的性能。但本文模型主要研究音频和唇部动作的同步性以及生成视频的清晰度，在面部情绪表达细节和躯体同步动作方面仍有待提升，在未来的研究中须进一步探索更加细致的音视频特征同步技术和情绪表达模型，使得生成的数字人脸视频能展现出人类更真实自然的情感表达。

参 考 文 献

[1] WANG J, QIAN X, ZHANG M, et al. Seeing what you said: Talking face generation guided by a lip reading expert[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 14653-14662.

[2] CHEN L, MADDOX R K, DUAN Z, 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. 2019: 7832-7841.

[3] 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.

[4] ZHOU H, SUN Y, 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. 2021: 4176-4186.

[5] GUO Y, CHEN K, 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. 2021: 5784-5794.

[6] MUKHOPADHYAY S, SURI S, GADDE R T, et al. Diff2lip: Audio conditioned diffusion models for lip-synchronization[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024: 5292-5302.

[7] MISTRY D S, KULKARNI A V. Overview: Speech Recognition Technology, Mel-Frequency Cepstral Coefficients (MFCC), Artificial Neural Network (ANN)[J/OL]. https://www.ijert.org/research/overview-speech-recognition-technology-mel-frequency-cepstral-coefficients-mfcc-artificial-neural-network-ann-IJERTV2IS100586.pdf.

[8] TRAN T, LUNDGREN J. Drill Fault Diagnosis Based on the Scalogram and Mel Spectrogram of Sound Signals Using Artificial Intelligence[J]. IEEE Access, 2020, 8: 203655-203666.

[9] LI H, QIU K, CHEN L, et al. SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images[J]. IEEE Geoscience and Remote Sensing Letters, 2020, 18(5): 905-909.

[10] QIN Z, ZHANG P, WU F, et al. Fcanet: Frequency channel attention networks[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 783-792.

[11] GUO M H, XU T X, LIU J J, et al. Attention mechanisms in computer vision: A survey[J]. Computational Visual Media, 2022, 8(3): 331-368.

[12] CHUNG J S, ZISSERMAN A. Out of time: automated lip sync in the wild[C]// Computer Vision－ACCV 2016 Workshops: ACCV 2016 International Workshops. 2017: 251-263.

[13] JI Y, YU Y Q. Optimization algorithm for speech facial video

generation based on dense convolutional generative adversarial networks and keyframes[J]. Journal of Jilin University(Engineering and Technology Edition), 2025, 55(3): 986-992.

[14] 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.

[15] SON C J, SENIOR A, VINYALS O, et al. Lip reading sentences in the wild[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6447-6456.

[16] CHUNG J, ZISSERMAN A. Lip reading in profile[C]// Ritish Machine Vision Conference. British Machine Vision Association and Society for Pattern Recognition, 2017.

[17] ZHAO Y, XU R, SONG M. A cascade sequence-to-sequence model for chinese mandarin lip reading[C]// Proceedings of the 1st ACM International Conference on Multimedia in Asia. 2019: 1-6.

[18] ZHAO Y, XU R, WANG X, et al. Hearing lips: Improving lip reading by distilling speech recognizers[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 6917-6924.

[19] PARK S J, KIM M, HONG J, et al. Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022: 2062-2070.

[20] LIANG B, PAN Y, GUO Z, et al. Expressive talking head generation with granular audio-visual control[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 3387-3396.

[21] DUCHI J, HAZAN E, SINGER Y. Adaptive subgradient methods for online learning and stochastic optimization[J]. Journal of machine learning research, 2011, 12(7): 2121-2159.

GUO Xingxing, born in 1998, postgraduate. Her main research interest is digital image processing.

XIAO Yannan, born in 1990, postgraduate, engineer. His main research interests include artificial intelligence and image-based 3D reconstruction.

(责任编辑:何杨)