多尺度视觉增强语音驱动人脸生成

杨想彦[1]，梁慧慧[2]，陈希[2]，李凡[2]
（1. 新疆大学计算机科学与技术学院，830046，乌鲁木齐；2. 西安交通大学电子与信息学部，710049，西安）

摘要：针对现有语音驱动人脸生成方法在生成人脸视频清晰度和真实感方面的不足，设计了一种多尺度视觉增强的端到端人脸视频生成模型——VisClearTalk，并提出包含视觉增强模块的人脸解码器。首先，采用人脸编码器处理包含随机参考帧和下半张脸被遮挡的先验帧，从中提取面部特征；接着，采用语音编码器从音频中提取特征，以指导面部内容生成；随后，采用人脸解码器融合上述特征，并通过卷积模块初步重建面部图像；最后，采用视觉增强模块将多尺度卷积和残差融合，进一步增强人脸下半部分区域的细节和边缘信息，从而提升生成人脸视频的视觉质量。使用公开唇语识别数据集对 VisClearTalk 模型进行实验验证，结果表明：通过引入视觉增强模块，该模型有效提升了面部视觉内容的细腻程度和真实感，能够生成清晰自然的人脸视频；在性能指标方面，峰值信噪比达到 34.349 dB，结构相似度达到 0.933，可学习感知图像块相似度降低至 0.040。研究结果可为当前人脸生成需求提供一种解决方案。

关键词：语音驱动；人脸生成；视觉增强；视觉质量
中图分类号：TP391 **文献标志码**：A
DOI：10.7652/xjtuxb202506017 **文章编号**：0253-987X(2025)06-0167-10

Audio-Driven Talking Face Generation with Multi-Scale Visual Enhancement

YANG Xiangyan[1], LIANG Huihui[2], CHEN Xi[2], LI Fan[2]
(1. School of Computer Science and Technology, Xinjiang University, Urumqi 830046, China;
2. Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: To address the limitations of existing audio-driven talking face generation methods in terms of video clarity and realism, an end-to-end talking face generation method called VisClearTalk which incorporates multi-scale visual enhancement is proposed in this paper, and a face decoder with a visual enhancement module is proposed. First, the face encoder processed a random reference frame and a prior frame with the lower half of the face occluded to extract facial features. Simultaneously, the audio encoder extracted features from the audio to guide facial content generation. Subsequently, the face decoder integrated these features and performed an initial reconstruction of facial images through convolutional modules. Finally, the visual enhancement module employed multi-scale convolution and residual fusion to further enhance the details and edge information of the lower face region, improving the visual quality of the generated talking face videos. The VisClearTalk model was experimentally validated using public lip-reading

收稿日期：2024-11-09。 作者简介：杨想彦(1998—)，女，硕士生；李凡(通信作者)，男，教授，博士生导师。 基金项目：国家自然科学基金资助项目(62471376)。
网络出版时间：2025-02-07 网络出版地址：https://link.cnki.net/urlid/61.1069.T.20250207.1449.002

datasets, with both quantitative and qualitative results demonstrating that the introduction of the visual enhancement module effectively improves the fineness and realism of facial visual content, enabling the generation of clear and natural talking face videos. In terms of performance metrics, the peak signal-to-noise ratio reached 34.349 dB, structural similarity reached 0.933, and learnable perceptual image patch similarity was reduced to 0.040. The VisClearTalk model offers a viable solution for current talking face videos generation needs.

Keywords: audio-driven; talking face generation; visual enhancement; visual quality

语音驱动人脸生成是视听领域的重要研究课题之一[1],其能够将视觉和听觉信息有机整合,增强人类对信息的理解和感知。对于人脸生成技术,生成清晰逼真的唇部细节能够大幅提升用户的视听体验,从而帮助用户更好地理解语义[2]。然而,由于人们对面部细微的变化极为敏感,如果生成的视频中面部动作与语音不同步,或唇部细节不够清晰,会使用户产生明显的违和感[3]。因此,如何在确保唇形与语音自然同步的同时,提升生成人脸的真实感与纹理清晰度,仍是该领域面临的重大挑战[4]。

传统的人脸生成方法多是利用先验知识(如音频),对静态源图像进行变形,极易导致输出不准确。随着深度神经网络,特别是生成对抗网络(GAN)和卷积神经网络(CNN)的发展,促成了语音驱动人脸生成技术的进展。目前,语音驱动人脸生成方法主要分为两类。第一类方法利用中间模态表征,将音频输入与视频输出连接起来,基于 3D 人脸重构以及 GAN 网络,通过语音信息去预测三维可变形人脸(3DMM)或 2D 面部的地标点,从而实现驱动人脸图像的动态变化[5-12]。例如,Zhong 等[5]从部分遮挡的人脸面部和静态参考帧中提取 2D 地标点,利用基于 Transformer 的地标生成器,从音频中预测唇部和下颌的地标点以生成唇部运动序列;Chatziagapi 等[6]弥合了基于 GAN 网络的唇形同步与神经辐射场的 3D 面部建模之间的差距,采用 3DMM 表情空间来表示唇形,通过音频到表情的映射,生成人脸视频。此类方法生成的人脸表情和嘴唇,能够根据语音特征进行相应的动态变化;然而,由于输入的语音和中间表征本质上不包含视觉内容信息,且为避免生成的面部形变与原始语音信号之间出现误关联,许多方法仅能生成唇部运动或眨眼动作,因此在保留身份信息的情况下,很难从音频和中间模态表征中产生逼真的人脸视频。此外,在生成长时间流畅的唇音同步面部序列方面,仍存在一定困难。

不同于中间模态表征,第二类方法直接学习语音到视频帧的映射,基于语音和源视频驱动,通过样本训练建立语音信号与唇部运动的映射模型[13-18],有效缓解了上述问题。例如,为实现精准的唇音同步,Suwajanakorn 等[16]采用递归神经网络学习从原始音频特征到嘴型的映射。Zheng 等[17]提出了一种新的基于时间 CNN 网络的唇音同步方法,用于学习从音频信号到唇部动作的跨模态顺序映射。Prajwal 等[18]设计了一种新的唇音同步网络,以约束人脸生成过程中的唇部动作。此类方法生成的人脸表情和动作主要参考源视频中的人脸运动,能够自然地呈现出逼真的表情和流畅的唇部动作。然而,在镜头有限的情况下,唇部纹理细节与驱动音频的关联度较低,使得通过音频映射直接生成高频唇部的纹理细节变得非常困难[19],从而导致该类方法在生成图像的下半张脸区域时,常常出现模糊或伪影的问题。

针对这一问题,Chen 等[20]提出了两阶段模型 HyperLips。其中,第一阶段以人脸帧序列作为输入,在保留原始人脸帧序列人物面部运动信息的基础上,通过超级网络 HyperNet 控制语音到唇部的映射,生成基础人脸,以提高唇音同步;第二阶段使用面部地标作为指引,通过高分辨率解码器实现对基础人脸的细节增强。该模型能够有效提高唇音同步和面部清晰度,然而由于高分辨率解码器依赖于第一阶段生成的人脸和对应的人脸地标草图进行训练,如果参考帧中出现侧脸或人脸角度偏斜,导致人脸检测器无法准确检测到地标,则模型将无法正常生成清晰的人脸。因此,如何在保证唇音同步的同时提升人脸清晰度,生成清晰自然的人脸视频仍是一个值得挑战的问题。

为了应对以上挑战,本文基于 HyperLips 模型,设计了一种端到端训练的人脸生成模型 VisClearTalk,提出了包含视觉增强模块的人脸解码器,通过多尺度卷积,从低质量图像中提取不同尺度的面部纹理细节,经过残差融合,进一步增强人脸下半部分区域的细节和边缘信息,从而有效细化面部视觉内容的生成。与原模型相比,VisClearTalk

模型仅需一个阶段即可生成高质量的人脸，有效解决了因地标检测失败而无法生成清晰人脸的问题。实验结果表明，本文提出的方法能够从给定的参考视频中渲染出更高质量的面部运动，生成清晰自然的人脸视频。

1　人脸视频生成

1.1　基础模型 HyperLips

HyperLips 模型是在给定输入音频、视频参考帧以及下半部分人脸被遮挡姿态先验帧的情况下，通过逐帧生成被遮挡区域的面部，实现同步嘴唇运动，从而生成高保真度的人脸视频。

该模型包含基础人脸生成和高保真渲染两个阶段。在基础人脸生成阶段，首先，通过人脸编码器提取面部信息，同时通过语音编码器提取语音特征；接着，将提取的语音特征输入到 HyperNet，旨在通过语音特征控制唇部动作生成；然后，通过双卷积层对特征进行进一步融合和调整。最后，由人脸解码器将融合特征解码为完整的面部图像，生成基础人脸。在高保真渲染阶段，采用第一阶段生成的人脸数据和相应的人脸地标草图，训练高清解码器，以增强基础人脸渲染。

1.2　优化模型 VisClearTalk

本文设计了一种端到端的人脸生成网络模型 VisClearTalk，提出了一个包含视觉增强模块的人脸解码器，用于替换 HyperLips 模型中基础人脸生成阶段的解码器，结构如图 1 所示。VisClearTalk 模型包含一个生成器，以及唇音同步和视觉质量两个判别器。生成器将输入的音频信息转换为一系列唇部动作序列，判别器通过比较生成的唇部动作与真实视频中的唇部动作，评估生成器的输出效果。在生成器解码阶段，人脸解码器利用卷积模块对融合后的特征进行初步解码，生成模糊的人脸视频帧，再经过视觉增强模块对初步生成的人脸帧进行高质量修复，进一步提升细节和视觉效果。在预训练的唇音同步和视觉质量判别器的监督下，VisClearTalk 模型通过生成对抗网络进行训练，建立音频与视频之间的映射关系，并通过该映射，生成与输入音频高度匹配的唇部动作序列，以填补被遮挡的下半部分人脸区域，实现音视频的高精度同步。

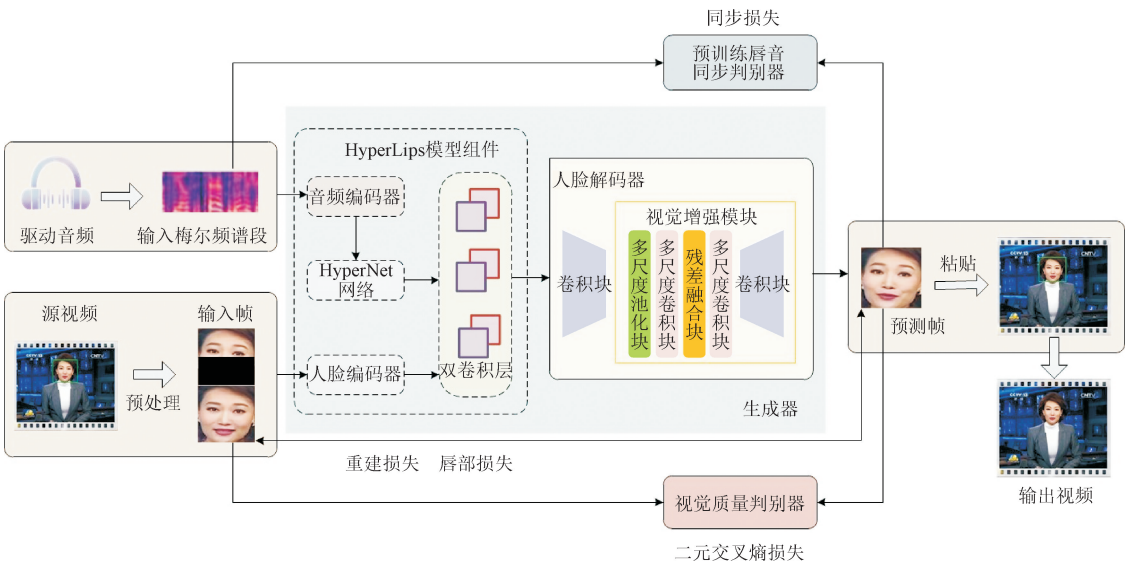


图 1　VisClearTalk 模型的网络结构

Fig. 1　Network structure of VisClearTalk model

1.3　人脸解码器

为了减少生成的人脸视频在下半部分人脸区域出现的模糊和伪影，本文设计了一个包含视觉增强模块的人脸解码器。该解码器由一个卷积块和一个视觉增强模块组成，其中视觉增强模块包括多尺度池化块、多尺度卷积块、残差融合块和一个卷积块，其核心功能是对第一个卷积块生成的低质量人脸帧进行高质量修复。解码器中的每个卷积块均由多个卷积层和转置卷积层堆叠而成。

多尺度池化块从初步生成的低质量人脸帧中提取不同尺度面部细节信息，确保模型能够捕捉到全局和局部特征，然后对输入特征执行最大池化操作，以提取显著特征，再将这些特征传递给多尺度卷积块进行进一步处理和整合。该网络结构见图 2，其中 C 为通道数。

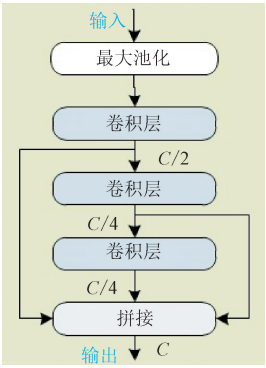


图 2 多尺度池化块网络结构

Fig. 2 Multi-scale pooling block network

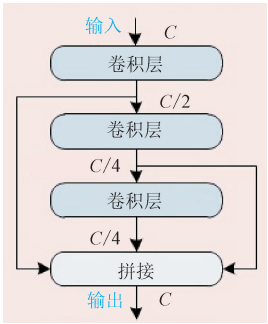


图 3 多尺度卷积块网络结构

Fig. 3 Multi-scale convolutional block network

多尺度卷积块接收特征图作为输入，并通过 3 个卷积层进行处理，其网络结构见图 3。第一个卷积层的输出通道数是输入通道数的 1/2，下来两个卷积层的输出通道数是输入通道数的 1/4。将这 3 个卷积层的输出在通道维度上进行拼接，既保持了输出与输入通道数相同，又融合了多层卷积后的特征信息。

残差融合块接受特征图作为输入，在多尺度池化块和多尺度卷积块的基础上嵌入残差单元，结合这些多尺度特征，进一步增强下半部分人脸细节和边缘信息。最终，将上采样后的输出特征图与多尺度卷积块输出特征图在通道维度上进行拼接。其具体的网络结构如图 4 所示。

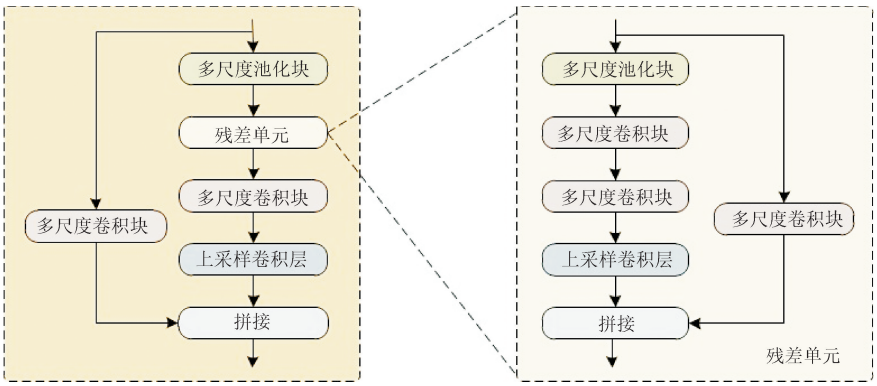


图 4 残差融合块网络结构

Fig. 4 Residual fusion block network

上述设计采用嵌套式模块思路，从模块到子网络再到完整网络，形成类似二阶 Hourglass[21] 的结构。左侧网络专注于保留原始分辨率下的全局特征，以辅助上采样过程；右侧残差单元则聚焦于 1/2 尺度特征的融合与细化。通过两者的协同作用，既能保持全局特征信息，又能增强局部低分辨率特征的细节表达。经过此过程融合的特征随后会被送入卷积块，以生成高质量的人脸帧。

1.4 损失函数

在训练阶段，采用重建损失、唇部损失、唇音同步损失和对抗损失 4 种损失函数训练人脸生成网络。

重建损失函数通过约束生成的人脸帧和真实的人脸帧之间的 L1 损失来实现视觉重建，表示如下

$$L_{\mathrm{r}}=\frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{I}^{\mathrm{B}}-\boldsymbol{I}^{\mathrm{GT}}\right\|_1 \tag{1}$$

式中：$\boldsymbol{I}^{\mathrm{B}}$ 为生成的人脸帧；$\boldsymbol{I}^{\mathrm{GT}}$ 为真实的人脸帧；N 为样本个数。

为了更好地细化唇部区域，本文借助预训练的视觉几何组(VGG)网络[22]提取人脸特征，采用可学习感知图像块相似度[23]来量化生成唇部区域与真实唇部区域之间的高阶特征差异，并将这种差异表示为 φ。同时，通过结合真实唇部区域的掩码，利用 L1 损失对生成图像中与掩码匹配的唇部区域进行约束，确保局部区域的重建质量。唇部损失函数可表示为

$$L_{\text{lip}}=\frac{1}{N}\sum_{i=1}^{N}(\varphi(\boldsymbol{I}_{\text{lip}}^{\text{B}},\boldsymbol{I}_{\text{lip}}^{\text{GT}})+\|(\boldsymbol{I}^{\text{B}}-\boldsymbol{I}^{\text{GT}})\boldsymbol{I}_{\text{lip}}^{\text{mask}}\|_1) \tag{2}$$

式中：$\boldsymbol{I}_{\text{lip}}^{\text{B}}$、$\boldsymbol{I}_{\text{lip}}^{\text{GT}}$ 分别为根据生成的人脸帧和真实的人脸帧的唇部标记点裁剪得到的唇部区域图像；$\boldsymbol{I}_{\text{lip}}^{\text{mask}}$ 为真实唇部区域图像的掩码。

采用唇音同步判别器，惩罚生成器生成的唇部动作与输入音频不同步的情况。为了生成高真实感的人脸视频，对文献[18]中提出的判别网络进行深化处理，采用 Adam 优化器[24]，设置初始学习率为 1×10^{-4}，批次为 32，在中文唇音同步数据集 CMLR 上对改进后的判别网络进行了预训练。

一个由生成的连续 5 帧人脸图像构成的序列 $\boldsymbol{F}^{\text{V}}$，其中每个图像仅包含下半部分面部（即嘴巴区域），与一个特定的音频片段 $\boldsymbol{F}^{\text{A}}$ 相对应。$\boldsymbol{F}^{\text{A}}$ 和 $\boldsymbol{F}^{\text{V}}$ 经过判别网络提取的特征分别表示为 $\boldsymbol{f}_{\text{a}}$、$\boldsymbol{f}_{\text{v}}$，输出特征对的同步概率计算如下

$$d_{\text{sync}}(\boldsymbol{f}_{\text{a}},\boldsymbol{f}_{\text{v}})=\frac{\boldsymbol{f}_{\text{a}}\boldsymbol{f}_{\text{v}}}{\|\boldsymbol{f}_{\text{a}}\|_2\|\boldsymbol{f}_{\text{v}}\|_2} \tag{3}$$

唇音同步损失函数可表示为

$$L_{\text{sync}}=-\frac{1}{N}\sum_{i=1}^{N}\ln(d_{\text{sync}}(\boldsymbol{F}_i^{\text{A}},\boldsymbol{F}_i^{\text{V}})) \tag{4}$$

对抗损失可通过视觉质量判别器[20]，预测生成人脸帧在细节、清晰度等方面是否与真实人脸帧相当的概率，其损失函数表示如下

$$L_{\text{gen}}=E_{\boldsymbol{I}^{\text{B}}}[\log(D(\boldsymbol{I}^{\text{B}}))] \tag{5}$$

$$L_{\text{disc}}=E_{\boldsymbol{I}^{\text{GT}}}[\log(1-D(\boldsymbol{I}^{\text{GT}}))]+L_{\text{gen}} \tag{6}$$

式中：L_{gen} 为判别器对生成人脸帧的预测准确度；L_{disc} 为对抗损失；$E_{\boldsymbol{I}^{\text{B}}}$、$E_{\boldsymbol{I}^{\text{GT}}}$ 分别为生成人脸帧和真实人脸帧的期望；$D()$ 为判别器对样本的预测概率。

综上所述，人脸说话视频生成阶段的训练损失可表示为

$$L=\lambda_{\text{r}}L_{\text{r}}+\lambda_{\text{lip}}L_{\text{lip}}+\lambda_{\text{sync}}L_{\text{sync}}+\lambda_{\text{disc}}L_{\text{disc}} \tag{7}$$

式中：λ_{r} 为面部重建损失惩罚权重；λ_{lip} 为唇部损失惩罚权重；λ_{sync} 为唇音同步损失惩罚权重；λ_{disc} 为对抗损失惩罚权重。上述权重总和为 1。

2　实验与讨论

2.1　参数设置

在数据预处理阶段，将采样率为 16 kHz 的音频信号转换为包含 16 个时间帧和 80 个梅尔频带的梅尔频谱图。同时，采用人脸检测器对帧率为 25 帧/s 的视频逐帧检测人脸区域，并将其裁剪调整为 128×128像素的人脸视频帧。基于前人经验[18,20]，将面部重建损失权重 λ_{r}、唇部损失权重 λ_{lip}、唇音同步损失权重 λ_{sync}、对抗损失权重 λ_{disc} 分别设置为 0.57、0.2、0.03 和 0.2。

模型训练阶段，VisClearTalk 模型的输入包括随机连续的 5 帧参考图像 $\boldsymbol{I}_i^{\text{R}}\in\mathbf{R}^{3\times128\times128}$，$i=1,2,\cdots,5$，随机连续的 5 帧下半部分被遮挡的人脸图像 $\boldsymbol{I}_i^{\text{M}}\in\mathbf{R}^{3\times128\times128}$，以及对应的音频梅尔频谱图 $\boldsymbol{A}_i^{\text{M}}\in\mathbf{R}^{16\times80}$。采用 B 表示批处理大小，对于人脸编码器，输入为 $\boldsymbol{I}_i^{\text{R}}$ 和 $\boldsymbol{I}_i^{\text{M}}$ 的拼接，其大小为 $B\times6\times128\times128$；对于音频编码器，输入为 $\boldsymbol{A}_i^{\text{M}}$，其大小为 $B\times1\times80\times60$。人脸解码器的输出为 5 帧预测图像 $\boldsymbol{I}_i^{\text{p}}\in\mathbf{R}^{3\times128\times128}$，其大小为 $B\times3\times128\times128$。

使用预训练好的唇音同步判别网络计算同步损失，设置批次大小为 16，使用 Adam 优化器，设置初始学习率为 0.000 1。所有实验均在 NVIDIA GTX3090 GPU 上进行。

2.2　数据集构建

所有实验均在 CMLR 数据集[25]上进行。CMLR 是句子级别的中文普通话唇语识别数据集，由新闻联播中 11 位主持人的 102 072 条视频片段组成。将数据集按照 7∶1∶2 的比例分为训练集、验证集和测试集，从测试集中随机抽取 80 个视频用于定量评估算法。

2.3　评价指标

采用峰值信噪比 ρ[26]、结构相似度 s[27]和可学习感知图像块相似度 l[23]作为性能指标，来评价生成人脸帧和真实人脸帧之间的相似度。为了公平比较，首先，通过 HyperLips 模型中的人脸检测器，检测视频帧中的人脸；接着，根据对应方法所需的分辨率，调整人脸区域；然后，采用对应方法生成人脸后，将其粘贴回原始视频中；最后，在生成的视频和对应的真实视频中裁剪出 128×128 像素的人脸，逐帧计算得到上述 3 个指标。

2.4　实验结果

将本文提出的 VisClearTalk 模型与现有语音驱动人脸生成模型进行比较。HyperLips 模型通过一个控制嘴唇运动的超网络和一个高分辨率解码器来生成人脸视频。DINet 模型采用一种变形修复网络，用于生成高分辨率人脸视频。Wav2Lip 模型通过对抗性训练学习的编码器-解码器模型生成人脸

视频。IP_LAP 模型提出了一个两阶段的框架，首先从语音中预测出对应的面部地标，然后将这些面部地标转换为视频帧。

2.4.1 对比实验结果

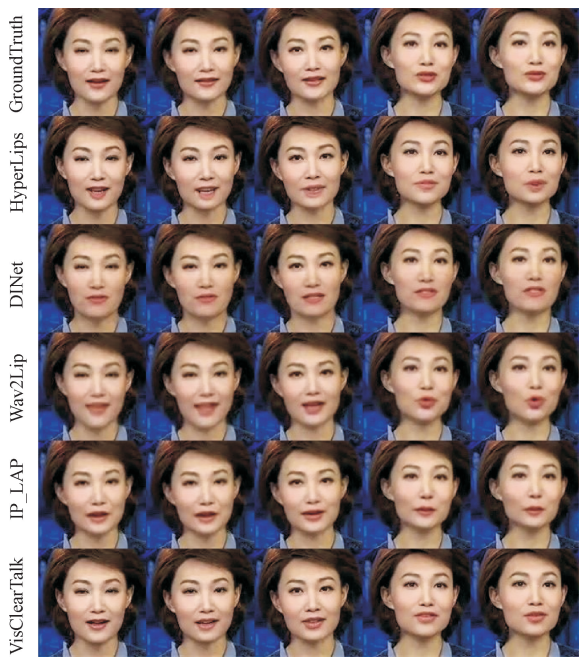
表 1 给出了采用 CMLR 数据集时不同模型的对比结果，其中 ρ、s 越大越好，l 越小越好。由表 1 可见，VisClearTalk 模型所生成的人脸在这 3 个指标上都明显优于其他方法，表明本文提出的视觉增强模块有利于高质量的人脸渲染。

表 1 不同人脸生成模型的结果对比

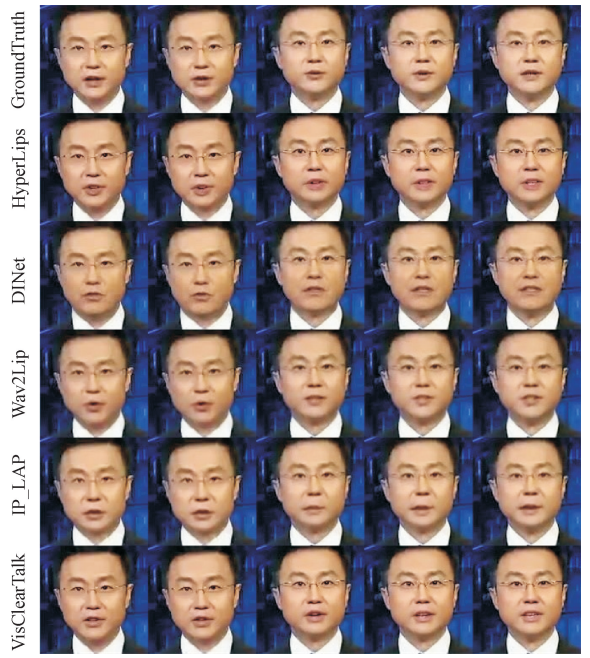
Table 1 Comparison results of different face generation models

模型	ρ/dB	s	l
HyperLips	33.822	0.918	0.050
DINet	33.729	0.929	0.043
Wav2lip	32.445	0.913	0.055
IP_LAP	33.543	0.903	0.051
VisClearTalk	34.349	0.933	0.040

为了验证 VisClearTalk 模型生成的人脸视频质量，邀请 20 名志愿者对采用不同模型生成的人脸视频进行评估。图 5 给出了所有模型在 CMLR 数据集[25]上的生成示例，其中每组模型均采用同一时刻随机连续拍摄的 5 帧视频图像。表 2 列出了图 5 中生成示例的量化性能指标，其中 $\bar{\rho}$、$\bar{s}$、$\bar{l}$ 分别表示视频 ρ、s、l 的平均值。



(a)女性



(b)男性

图 5 不同模型在 CMLR 数据集上的生成示例

Fig. 5 Generation examples of different generation methods on the CMLR datasets

表 2 图 5 生成示例的量化性能指标

Table 2 Quantitative performance metrics for the generated examples in Fig. 5

模型	$\bar{\rho}$/dB		$\bar{s}$		$\bar{l}$	
	女	男	女	男	女	男
HyperLips	33.293	32.863	0.914	0.917	0.059	0.053
DINet	33.133	33.739	0.921	0.927	0.037	0.037
Wav2lip	32.547	32.546	0.912	0.928	0.050	0.048
IP_LAP	33.187	33.570	0.897	0.912	0.056	0.053
VisClearTalk	33.999	33.748	0.939	0.948	0.035	0.034

从 CMLR 测试数据集中随机选取 20 个生成的视频，与 20 个真实的人脸说话视频混合，让志愿者判断哪些视频是真实的，哪些是通过网络生成的。计算所有志愿者判断结果的辨识误差[15]，列于表 3。其中，辨识误差越小，表示志愿者越难以分辨视频是生成的还是真实的。可以看出，相比于其他模型，志愿者对 VisClearTalk 生成的人脸视频更加难以准确区分，表明 VisClearTalk 模型的视频质量性能最为优异。

鉴于本文所提出的 VisClearTalk 模型能够使用训练集以外的音频和人脸视频，生成自然的人脸视频，为了测试这一性能，随机选取 LRS2 数据集[28]中两个不同身份的人脸视频，采用相同的语音信号驱动不同的人脸生成模型，并对比各模型的生成结果。

表 3　视频质量评估结果对比

Table 3 Comparion of video quality evaluation results

模型	辨识误差/%
HyperLips	35.00
DINet	52.75
Wav2lip	38.50
IP_LAP	43.50
VisClearTalk	31.25

图 6 展示了真实帧及 HyperLips、DINet、Wav2Lip、IP_LAP 和 VisClearTalk 模型生成的人脸帧。从视觉效果上看，真实视频帧和 VisClearTalk 模型生成的人脸视频帧之间的差异较小，远小于真实视频与其他 4 个模型生成的视频帧之间的差异。此外，VisClearTalk 生成的人脸序列视频更加真实可信，且在生成结果中没有出现伪影，可以清晰地呈现人脸，甚至可以清晰地看到牙齿。表 4 列出了图 6 中不同模型生成示例的量化性能指标。

(a)女性　　(b)男性

图 6　真实帧与不同模型生成结果的对比

Fig. 6 Comparison of ground truth and generated results from different models

表 4　图 6 生成示例的量化性能指标

Table 4 Quantitative performance metrics for the generated examples in Fig. 6

模型	$\bar{\rho}$/dB		$\bar{s}$		$\bar{l}$	
	女	男	女	男	女	男
HyperLips	34.535	31.651	0.922	0.865	0.057	0.066
DINet	33.363	31.892	0.908	0.831	0.059	0.080
Wav2lip	32.085	30.857	0.932	0.893	0.060	0.066
IP_LAP	33.104	31.946	0.891	0.814	0.064	0.076
VisClearTalk	34.581	32.133	0.932	0.893	0.052	0.048

2.4.2　消融实验结果

为了验证本文提出的核心组件视觉增强模块以及损失函数在人脸视频生成任务中的有效性，在 CMLR 数据集上进行了相应的消融实验。

设计 4 个对比实验：①w/o 残差单元：将残差融合块中的残差单元替换为多尺度卷积块，移除其嵌套结构，使残差融合块简化为仅保留一阶结构；②w/o 视觉增强：移除人脸解码器中的视觉增强模块，即人脸编码器仅保留第一个卷积块；③w/o L_{lip}：在训练阶段消除唇部损失 L_{lip}；④w/o L_{sync}：在训练阶段消除唇音同步损失 L_{sync}。其中，w/o 表示移除某个指定模块的模型版本。

表 5 给出了消融实验的结果对比，图 7 给出了实验的结果示例，其中每组实验均采用同一时刻随机连续拍摄的 2 帧视频图像。实验结果显示：加入残差单元能够捕捉到更丰富的结构信息，生成包含更多高频细节的图像，进一步改善了生成效果；当模型中加入视觉增强模块时，评估人脸帧生成质量的各项指标均有显著提升，与未加入视觉增强模块的

原始模型相比，峰值信噪比提高了 0.493 dB，结构相似度提高了 0.013，表明视觉增强模块的加入有利于生成高质量的人脸视频。此外还可以看出，唇部损失 L_{lip} 和唇音同步损失 L_{sync} 的加入，也有效提升了模型的整体性能。

表 5 视觉增强模块有效性消融实验结果对比

Table 5 Comparison of ablation experiment results on the effectiveness of the visual clear module

方法	ρ/dB	s	l
w/o 残差单元	34.189	0.910	0.045
w/o 视觉增强	33.856	0.920	0.049
w/o L_{lip}	34.143	0.918	0.046
w/o L_{sync}	34.157	0.911	0.045
VisClearTalk	34.349	0.933	0.040

(a)女性

(b)男性

图 7 消融实验结果示例

Fig. 7 Examples of ablation experiment results

2.5 讨 论

VisClearTalk 模型通过引入视觉增强模块，在生成过程中能够更精准地捕捉细节，与模型[5, 18-20]相比，在生成质量与视觉一致性方面展现出显著优势。然而，该模型仍存在一些局限性需要进一步探讨和解决。

首先，尚未将模型中的视觉增强模块与文中参与对比的其他模型[5,18-20]进行整合，以开展全面的性能比较，因此其相对优势仍有待进一步验证。未来的研究将着重于将视觉增强模块与这些方法结合，以评估该模块在不同模型中的有效性。

其次，当前实验所用数据集主要集中在特定类型的场景或对象上，缺乏足够的多样性来充分反映现实世界的广泛变化。这种局限性可能会限制模型在未见数据上的泛化能力，影响其在不同应用场景中的表现和适用性。未来的研究将计划在更大规模且更为多样化的数据集上训练 VisClearTalk，以提升其泛化能力，确保模型在复杂场景中的稳定性。

3 结 论

针对语音驱动人脸生成任务中存在的清晰度不足与真实感欠缺问题，提出了一种基于多尺度视觉增强的端到端生成模型 VisClearTalk。该模型在 HyperLips 框架基础上，通过引入视觉增强模块与残差融合机制，显著提升了生成视频的细节保真度与视觉自然性。通过对比实验验证，得到以下主要结论。

(1)提出的 VisClearTalk 模型通过引入多尺度卷积和残差融合机制，采用视觉增强模块细化了人脸下半部分区域的边缘与纹理特征，与未加入视觉增强模块的原模型相比，生成图像的峰值信噪比提升了 0.493 dB，结构相似度提升了 0.013。

(2)在 CMLR 测试集上，VisClearTalk 模型生成图像的峰值信噪比达到 34.349 dB，结构相似度达到 0.933，可学习感知图像块相似度降至 0.040。相较于 HyperLips 模型，峰值信噪比提高了 0.527 dB，结构相似度提高了 0.015，可学习感知图像块相似度降低了 0.010。

(3)在 LRS2 数据集上实现跨身份生成任务中，VisClearTalk 模型生成的视频帧在视觉质量及定量指标上均优于现有的语音驱动人脸生成模型，且生成结果中未出现伪影，面部细节甚至牙齿均清晰可见。

参考文献：

[1] YU Lingyun, YU Jun, LI Mengyan, et al. Multimodal inputs driven talking face generation with spatial-temporal dependency [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(1): 203-216.

[2] TOSHPULATOV M, LEE W, LEE S. Talking human face generation: a survey [J]. Expert Systems with Applications, 2023, 219: 119678.

[3] 张哲，齐春，张钊强，等. 人脸超分辨率重建中投影空间的选择方法 [J]. 西安交通大学学报，2018，52(8)：43-48.
ZHANG Zhe, QI Chun, ZHANG Zhaoqiang, et al. A selection method of projection space for face super-resolution reconstruction [J]. Journal of Xi'an Jiao-

tong University，2018，52(8)：43-48.

[4] 张云佐，郭亚宁，李文博. 融合时空切片和双注意力机制的视频摘要方法 [J]. 西安交通大学学报，2022，56(12)：127-135.
ZHANG Yunzuo，GUO Yaning，LI Wenbo. Video summarization method based on spatiotemporal slice and dual attention mechanism [J]. Journal of Xi'an Jiaotong University，2022，56(12)：127-135.

[5] ZHONG Weizhi，FANG Chaowei，CAI Yinqi，et al. Identity-preserving talking face generation with landmark and appearance priors [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway，NJ，USA：IEEE，2023：9729-9738.

[6] Chatziagapi A，Athar S，JAIN A，et al. LipNeRF：what is the right feature space to lip-sync a NeRF? [C]//2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG). Piscataway，NJ，USA：IEEE，2023：1-8.

[7] ZHANG Wenxuan，CUN Xiaodong，WANG Xuan，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 (CVPR). Piscataway，NJ，USA：IEEE，2023：8652-8661.

[8] 孙威. 中文文本驱动的人脸说话视频生成方法研究与实现 [D]. 南京：东南大学，2022.

[9] ESKIMEZ S E，ZHANG You，DUAN Zhiyao. Speech driven talking face generation from a single image and an emotion condition [J]. IEEE Transactions on Multimedia，2022，24：3480-3490.

[10] YE Zhenhui，JIANG Ziyue，REN Yi，et al. GeneFace：generalized and high-fidelity audio-driven 3D talking face synthesis [EB/OL]. (2023-01-31) [2024-05-15]. https://arxiv.org/abs/2301.13430.

[11] YE Zhenhui，HE Jinzheng，JIANG Ziyue，et al. GeneFace++：generalized and stable real-time audio-driven 3D talking face generation [EB/OL]. (2023-05-01) [2024-06-20]. https://arxiv.org/abs/2305.00787.

[12] CHEN Lele，MADDOX R K，DUAN Zhiyao，et al. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway，NJ，USA：IEEE，2019：7824-7833.

[13] YE Zipeng，XIA Mengfei，YI Ran，et al. Audio-driven talking face video generation with dynamic convolution kernels [J]. IEEE Transactions on Multimedia，2023，25：2033-2046.

[14] 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 Thirty-Sixth AAAI Conference on Artificial Intelligence and Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence and the Twelveth Symposium on Educational Advances in Artificial Intelligence. Palo Alto，CA，USA：AAAI Press，2022：2062-2070.

[15] 侯冰莹. 基于生成式对抗网络的语音驱动人脸生成的研究 [D]. 武汉：华中师范大学，2023.

[16] SUWAJANAKORN S，SEITZ S M，KEMELMACHER-SHLIZERMAN I. Synthesizing Obama：learning lip sync from audio [J]. ACM Transactions on Graphics，2017，36(4)：95.

[17] ZHENG Ruobing，ZHU Zhou，SONG Bo，et al. A neural lip-sync framework for synthesizing photorealistic virtual news anchors [C]//2020 25th International Conference on Pattern Recognition (ICPR). Piscataway，NJ，USA：IEEE，2021：5286-5293.

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

[19] ZHANG Zhimeng，HU Zhipeng，DENG Wenjin，et al. Dinet：deformation inpainting network for realistic face visually dubbing on high resolution video [C]//AAAI Conference on Artificial Intelligence. Palo Alto，CA，USA：AAAI Press，2023，37：3543-3551.

[20] CHEN Yaosen，YAO Yu，LI Zhiqiang，et al. HyperLips：hyper control lips with high resolution decoder for talking face generation [J]. Applied Intelligence，2024，55(2)：145.

[21] NEWELL A，YANG Kaiyu，DENG Jia. Stacked hourglass networks for human pose estimation [C]//Computer Vision-ECCV 2016. Cham：Springer International Publishing，2016：483-499.

[22] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10)[2024-10-12]. https://arxiv.org/abs/1409.1556.

[23] ZHANG R，ISOLA P，EFROS A A，et al. The unreasonable effectiveness of deep features as a perceptual metric [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ，USA：IEEE，2018：586-595.

[24] KINGMA D P，BA J. Adam：a method for stochastic

optimization [EB/OL]. (2017-01-30)[2024-10-20]. https://arxiv.org/abs/1412.6980.

[25] ZHAO Ya, XU Rui, SONG Mingli. A cascade sequence-to-sequence model for Chinese mandarin lip reading [C]//Proceedings of the 1st ACM International Conference on Multimedia in Asia. New York, USA: ACM, 2020: 32.

[26] HORÉ A, ZIOU D. Image quality metrics: PSNR vs. SSIM [C]//2010 20th International Conference on Pattern Recognition. Piscataway, NJ, USA: IEEE, 2010: 2366-2369.

[27] WANG Zhou, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity [J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[28] 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.

（编辑 李慧敏）