Computer Science and Application 计算机科学与应用, 2025, 15(1), 199-208
Published Online January 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.151020

Hans汉斯

基于语音驱动的人脸生成

李昊渊[1,2]

[1]河北地质大学信息工程学院，河北 石家庄
[2]河北地质大学人工智能与机器学习研究室，河北 石家庄

收稿日期：2024年12月23日；录用日期：2025年1月20日；发布日期：2025年1月28日

摘　要

语音驱动人脸生成旨在生成与参考人脸具有相同身份信息，与语音内容相对应的说话人脸视频。针对现有方法中生成人脸身份信息较差、脸部细节较差的问题，提出了一种基于关键点的语音驱动说话人脸视频生成模型LTFG-GAN。该模型首先将基于在语音识别领域微调的无监督预训练模型作为语音编码器，通过融合卷积与注意力机制预测人脸关键点；其次在人脸生成过程中加入交叉注意力机制获取原始参考人脸信息，通过条件卷积与空间自适应归一化将扭曲得到高维形变人脸信息与原始人脸信息融合；最终得到与语音同步的说话人脸视频。实验结果表明，上述方法对于人脸的生成有明显地提升。

关键词

人脸生成，深度学习，Wav2vec，交叉注意力机制，条件卷积

Speech-Driven Facial Generation

Haoyuan Li[1,2]

[1]College of Information Engineering, Hebei GEO University, Shijiazhuang Hebei
[2]Artificial Intelligence and Machine Learning Laboratory, Hebei University of Geosciences, Shijiazhuang Hebei

Received: Dec. 23rd, 2024; accepted: Jan. 20th, 2025; published: Jan. 28th, 2025

Abstract

Voice driven face generation aims to generate speech facial videos that have the same identity information as the reference face and correspond to the speech content. A speech driven facial video generation model based on landmarks, LTFG-GAN, is proposed to address the issues of poor facial identity information and facial details in existing methods. The model first uses an unsupervised pre trained model fine-tuned in the field of speech recognition as a speech encoder, and predicts

文章引用: 李昊渊. 基于语音驱动的人脸生成[J]. 计算机科学与应用, 2025, 15(1): 199-208.
DOI: 10.12677/csa.2025.151020

facial landmarks by integrating convolution and attention mechanisms; Secondly, a cross-attention mechanism is added to the face generation process to obtain the original reference face information. The distorted high-dimensional deformed face information is fused with the original face information through conditional convolution and spatial adaptive normalization; The final result is a speech synchronized facial video. The experimental results show that the above method has a significant improvement in face generation.

Keywords

Facial Recognition, Deep Learning, Wav2vec, Cross-Attention Mechanism, Conditional Convolution

1. 引言

语音驱动的说话人脸视频生成是指输入任意一段语音数据，结合参考图像，由计算机自动生成具有语音相应内容的说话人脸视频[1]。说话人脸视频生成在许多领域中具有重要价值广泛应用，例如视觉配音、数字助手、电影制作、虚拟交互等。说话人脸视频生成技术可以分为端到端生成和借助中介特征生成两种方法。

在端到端生成中，Chung 等人[2]采用了编码器–解码器(Encoder-Decoder)架构的卷积神经网络(Convolutional Neural Network)，以语音信号作为主导信息实现了唇音同步的人脸视频的端到端生成，具有开创性的意义；Prajwal 等人[3]提出了基于生成对抗网络(Generative Adversarial Network)的 LipGAN 模型，在生成器中对输入的语音和人脸进行低维度的特征提取后将特征拼接进行图像恢复，再通过鉴别器确定生成的人脸是否与输入的语音匹配，改善了生成的人脸效果；Prajwal 等人[4]在 LipGAN 模型的基础上进行了进一步的优化提出了 Wav2Lip 模型，使用了基于 SyncNet [5]模型改进的唇音同步专家鉴别器来检测生成视频中的唇部运动与语音中的内容是否同步，进一步提高了唇音同步的准确性；Kun [6]等人在 Wav2Lip 模型的基础上在编码器和解码器的连接处使用交叉注意力融合参考人脸信息，进一步提升了人脸生成的效果。端到端生成大多高效便捷，但往往需要训练额外的鉴别器来保证唇音同步，生成的人脸轮廓边界不清晰，出现下颚抖动现象，简单的编码与解码在生成牙齿等细微的结构时生成较为模糊。

借助人脸中介特征生成主要分为三种。第一种为借助人脸或唇部的关键点生成人脸，Suwajanakorn 等人[7]通过时延的长短期记忆神经网络(Time-delayed Long Short Term Memory)先预测出语音中的嘴部关键点的坐标信息画出简易嘴部形状，再通过传统的计算机视觉算法对嘴部形状进行肌理的渲染，生成效果真实，但生成算法过于复杂，生成速度较慢。第二种为借助 3D 模型生成人脸，Yi 等[8]从 2D 人脸图像结合三维形变模型(3DMorphable Model)的方式构建 3D 人脸模型，再通过 GAN 生成在二维投影的人脸图像，使用 3DMM 的方式虽然使得生成的人脸视频面部和头部的姿势变得可控，但是数据集收集困难，且 3D 人脸建模的方法依赖 3DMM 人脸建模模型和视频重定时算法，导致生成的人脸视频不自然。Yudong [9]等人使用神经辐射场(Neural Radiance Field)作为主体框架并引入基于注意力机制的面部解耦模块来渲染说话人脸图像，生成的图像更加逼真，但是流程复杂、计算量巨大，需要对生成的目标说话者进行重新训练或微调。第三种为借助二维人脸的扭曲形变信息生成人脸，Zhimeng [10]等人提出仿射形变修复网

络 DINet，对编码器提取的人脸特征图采用空间形变的方式扭曲特征图来输入解码器再进行修复，形变操作将像素移动到合适位置，能够保留人脸纹理的细节，这种方法生成的纹理效果相较于其他方法最为还原，但是全部借助于参考人脸的形变信息，导致在生成某些参考信息中未出现的部位时效果较差。

为了使生成的口型不受声纹信息的干扰，本文使用微调后的无监督预训练模型 Wav2vec 2.0 [11]的编码器提取初步语音特征，设计了音频编码器对特征进行降维，并使用多层的 Conformer [12]模块来预测人脸关键点。为了使生成的人脸与人物身份一致，充分利用参考信息，本文利用交叉注意力机制来补充人脸的原始参考信息，并在多尺寸下利用条件卷积模块预测卷积核权重，将特征的语义融合过程可控。

2. LTFG-GAN 模型

2.1. 整体框架

LTFG-GAN 网络的整体架构如图 1 所示。本文采用 IP_LAP [13]作为整体架构的主干，网络由关键点生成器和人脸生成器构成。在关键点生成器中，本文使用 Wav2vec 提取语音特征并设计音频特征提取器来对提取的语音特征进行维度压缩，对目标人脸图像的 L^c 对应的 L^p 以及若干参考 L 通过一维卷积进行特征提取，将这三部分特征通过绝对位置编码和模态编码后拼接，输入基于 Conformer 的编码器中，通过融合理解不同模态的信息，准确预测 L^c，与 L^c 拼接后绘制人脸轮廓图 L^s。

人脸生成器由对齐模块、翻译模块构成，使用多尺度鉴别器[14]。对齐模块将参考图片的信息通过卷积得到原始人脸的特征信息，再通过扭曲等若干操作将关键点生成器中预测的 L^s 生成的扭曲人脸信息，之后，将这两部分信息传递给翻译模块，由翻译模块将原始输入信息与对齐模块提供的两种信息不断融合最终生成人脸。

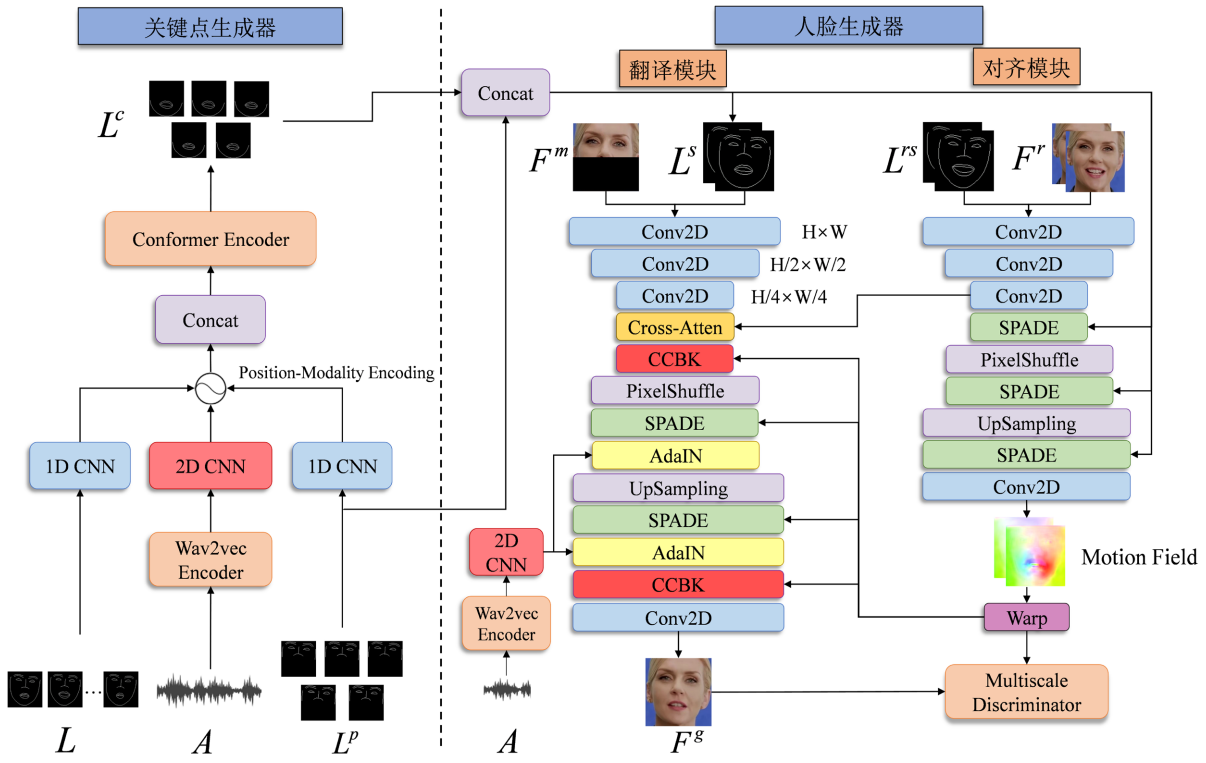

Figure 1. LTFG-GAN network
图 1. LTFG-GAN 网络总图

网络的生成过程具体定义如下：

$$\begin{cases} L^{s}_{t-k:t+k} = \Phi\left(A_{t-k:t+k}, L^{p}_{t-k:t+k}, L_{1:l}\right) \\ F^{g}_{t} = \Psi\left(F^{r}_{1:f}, L^{rs}_{1:f}, F^{m}_{t}, L^{s}_{t-k:t+k}, A_{t}\right) \end{cases} \tag{1}$$

其中 Φ 表示关键点生成器，Ψ 表示人脸生成器，t 表示目标帧的序列号，k、l、f 表示各个特征信息的参考量。

2.2. Wav2vec 提取语音特征

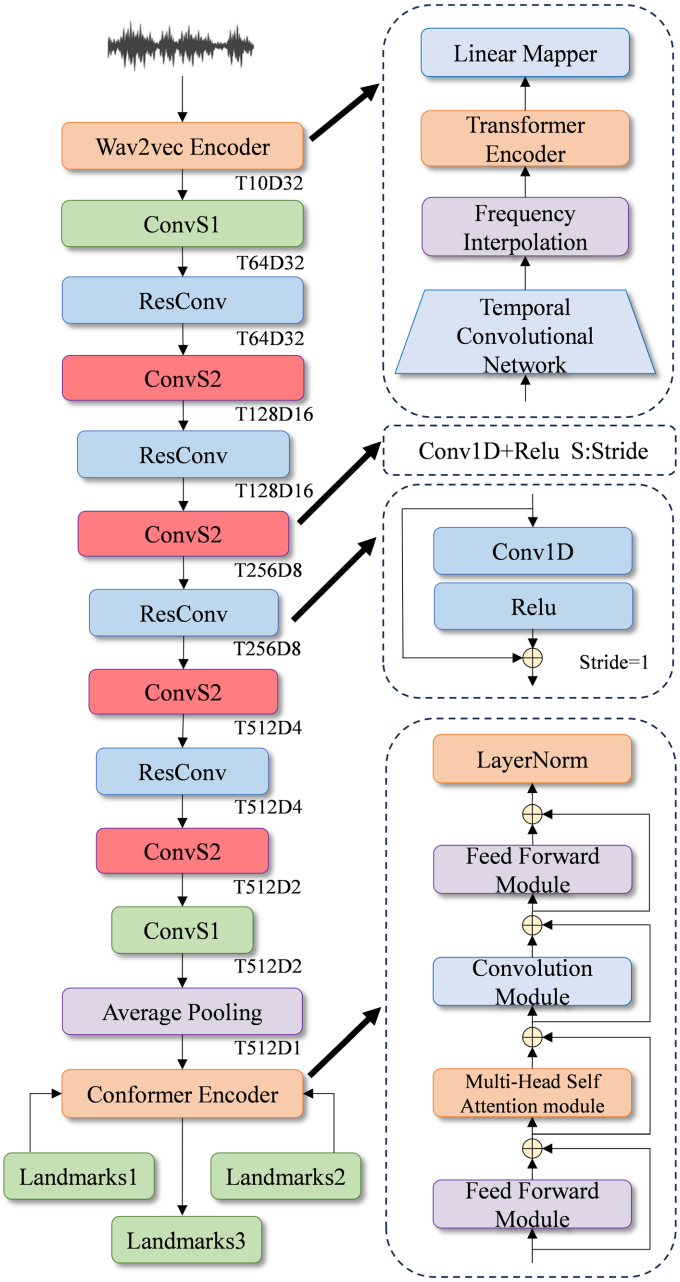

Figure 2. Audio encoder based on Wav2vec2.0
图 2. 基于 Wav2vec2.0 的音频编码器

语音特征可以分为反映了说话者独特声音特征的声纹信息和表达具体语义的内容信息[15]，对说话者语音特征的提取和处理对于准确地生成自然、流畅的口型至关重要。广泛使用基于梅尔频谱提取的梅尔频谱图(Mel-spectrogram)以及进一步从梅尔频谱图中计算得到的梅尔频率倒谱系数(Mel Frequency Cepstral Coefficients, MFCC)作为语音特征，梅尔频谱主要捕捉语音中低层次的频率信息，如音调、音色、共振峰等声音的物理特性有关的信息，提取的特征对于语音驱动人脸生成任务而言包含了无用的部分声纹信息，且梅尔频谱特征无法充分表征语音中的上下文依赖和更高层次的语义信息。为了使生成的人脸与自身的身份信息尽可能保持一致，提取的语音特征具有更高层次的语义信息、更好的可解释性以及更多的内容信息，本文使用无监督的预训练语音模型 Wav2Vec2.0 在语音识别领域(Automatic Speech Recognition，ASR)微调后的编码器提取音频特征，并使用结合了 Transformer 和卷积机制的 Conformer 结构提取特征来设计音频编码器，以获得更高层次的抽象表示，音频编码器的具体结构如图 2 所示。

2.3. 交叉注意力模块

交叉注意力机制[16]在计算机视觉任务和多模态处理任务中被广泛应用。在人脸生成器中，本文通过交叉注意力机制融合翻译模块中的内容特征图(输入向量在前向传播中的代称)与对齐模块传递的作为补充信息的参考特征图，将内容特征图的结构信息与参考特征图中带来的参考人脸特征信息进行融合，得到与参考特征图具有语义一致性、与内容特征图具有结构连贯性的人脸特征图。交叉注意力模块如图 3 所示。

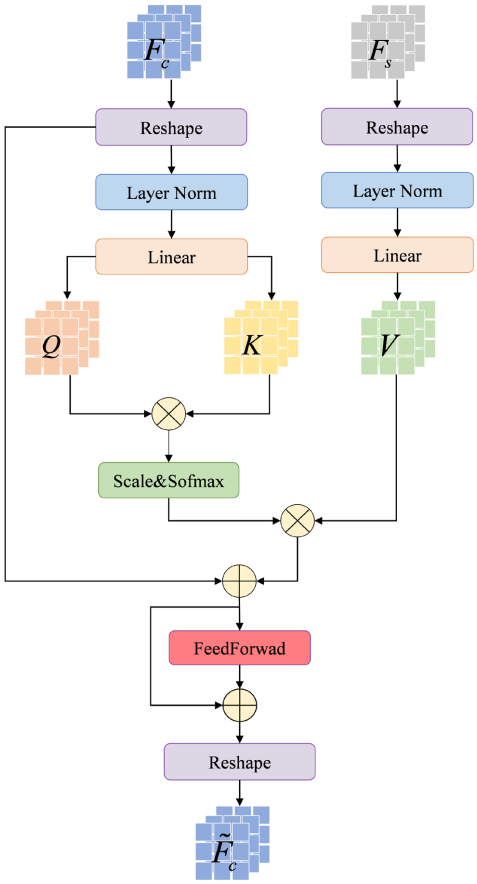

Figure 3. Cross attention module
图 3. 交叉注意力模块

在生成过程中，使用若干张参考特征图并对它们进行求和平均，能够综合利用多张特征图提供的不同细节信息，进而提供更丰富和更准确的特征表示。这使得模型能够更全面地理解输入数据，降低对单个特征图的依赖性，从而提高其稳健性和泛化能力。对 N 张参考特征图的求和平均的定义如下：

$$F_s = \frac{1}{N}\sum_{i=1}^{N} F_s^i \tag{2}$$

其中，内容特征图为 $F_c \in \mathbb{R}^{C\times T}$，若干张参考特征图为 $F_s^i \in \mathbb{R}^{C\times T}$，$i \in \{1,2,\cdots,N\}$，$T = H \times W$，$C$、$H$、$W$ 表示特征图的通道数、长、宽，通过求和平均得到合并参考特征图 F_s，由 F_c 和 F_s 得到注意力权重矩阵，定义如下：

$$\begin{cases} Q = Linear\left(LN\left(F_c\right)\right) \\ K = Linear\left(LN\left(F_c\right)\right) \\ V = Linear\left(LN\left(F_s\right)\right) \end{cases} \tag{3}$$

其中，LN 为 $LayerNorm$，查询矩阵 $Q \in \mathbb{R}^{h\times d\times T}$，关键字矩阵 $K \in \mathbb{R}^{h\times d\times T}$，值矩阵 $V \in \mathbb{R}^{h\times d\times T}$，$h$ 为注意力头的个数，d 为注意力头的维度。

在上式的计算过程中，Q 和 K 由内容特征图生成，V 由合并参考特征图生成。Q 表示生成过程中要重点关注的信息，是内容特征图中某片区域或某条特征的表示，即查询哪些特征需要与其它特征进行匹配；K 表示所有可能的匹配特征，用来衡量内容特征图中 Q 与其它部分之间的关联度；V 表示匹配的具体特征值，提供了参考特征图中的细节信息，例如提取的牙齿细节、下颚边缘、纹理细节等深度信息；V 根据 Q 和 K 之间的关联程度进行加权处理，最终生成输出特征向量。

注意力权重计算 Q，K，V 在交叉注意力块的定义如下：

$$F_c' = softmax\left(\frac{LN\left(Q\right)\otimes LN\left(K^{\mathrm{T}}\right)}{\sqrt{d}}\otimes LN\left(V\right)\right) + F_c \tag{4}$$

$$\tilde{F}_c = FeedForward\left(LN\left(F_c\right)\right) + F_c' \tag{5}$$

其中，$\otimes$ 为矩阵点乘。

交叉注意力机制帮助生成器聚焦于关键的信息部分，动态地调整不同特征之间的权重，根据输入数据的情况自适应地融合内容特征图和参考特征图，提高模型对于不同区域之间的语义关联性的捕捉能力。

2.4. 条件卷积

在传统的卷积神经网络中，单个卷积核在整张图像的所有空间位置上滑动应用，提取局部特征。相同的卷积核权重在整张图像中共享，而没有考虑图像中不同位置的语义信息或不同样本之间的语义差异。传统的卷积操作对于结合语义布局信息的图像生成来说不够灵活和有效，而条件卷积以语义特征图为条件来预测卷积核，使得布局信息可以更明确和有效地控制图像生成过程[17]。

为了使卷积层能够感知语义特征图中不同部位人脸信息的语义标签，将不同尺寸的布局信息融入图像生成过程，翻译模块使用条件卷积在最低维度与最高维度下融合相对应尺寸的内容特征图 $M_f \in \mathbb{R}^{C\times H\times W}$ 与语义特征图 $M_s \in \mathbb{R}^{C\times H\times W}$ (扭曲人脸信息)，条件卷积具体操作如图 4 所示。

条件卷积结合了深度可分离卷积的思想，将普通卷积分解为条件深度卷积和点卷积，条件深度卷积对 M_s 的每个通道执行空间滤波来预测轻量级深度方向的卷积核权重 $W_c \in \mathbb{R}^{C\times D\times H\times W}$，并将预测的卷积核权重与内容特征图点乘，定义如下：

$$M'_f = \sum_{i=1}^{D}\left(unfold\left(M_f\right)\otimes W_c\right)_{:,i,:,:} \tag{6}$$

其中，$D = k \times k$，k 为卷积核大小，$\otimes$ 为矩阵点乘。

点卷积对 M_f 每个像素位置的所有通道进行线性组合，将通道间的信息重新融合，并结合由 M_s 生成的条件注意力权重 $W_s \in \mathbb{R}^{C\times H\times W}$ 和经过空间自适应归一化操作的归一特征图 M_n，得到最终生成的内容特征图，定义如下：

$$\tilde{M}_f = M'_f \otimes W_s + M_n \tag{7}$$

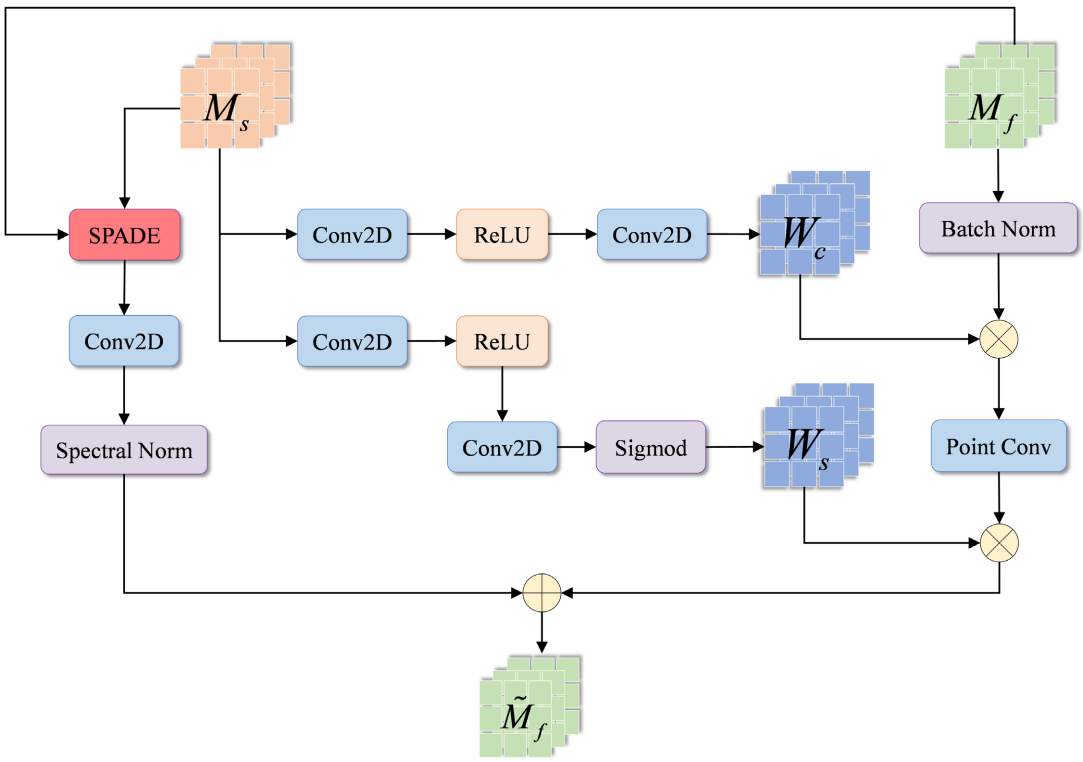

Figure 4. Conditional convolution module
图 4. 条件卷积模块

2.5. 损失函数

关键点生成器的总损失函数可以表示为平均误差损失和连续速度损失，定义如下：

$$L_d = L_m + L_v \tag{8}$$

其中，L_m 为预测连续关键点序列与目标连续关键点序列的 L_1 距离，即绝对误差损失，L_v 为真实序列与预测序列的速度变化，定义如下：

$$L_v = \frac{1}{T-1}\sum_{j=2}^{T}\left\|\left(gt_j - gt_{j-1}\right)-\left(pt_j - pt_{j-1}\right)\right\|_2 \tag{9}$$

其中，T 为连续关键点序列的总条数，gt_j 和 pt_j 分别表示真实序列和预测序列中第 j 条关键点序列，$T = 2k+1$。

人脸生成器中的损失函数分为生成器损失 L_{gen} 和判别器损失 L_{dis}。生成器损失中包含对齐模块的扭

曲感知损失 L_w、翻译模块的生成感知损失 L_t，多尺度均方误差损失 L_g，生成器特征匹配损失 L_f，定义如下：

$$L_{gen} = \lambda_w L_w + \lambda_t L_t + \lambda_g L_g + \lambda_f L_f \tag{10}$$

其中 L_w 和 L_t 为使用预训练网络 VGG19 计算的多层感知损失(Perceptual Loss) [18]，L_g 为真实样本和虚假样本通过判别器生成特征图之间的均方误差损失，L_f 为 Pix2PixHD 网络中的特征匹配损失。

判别器损失中包含对真实样本的损失 L_a 和对虚假样本的损失 L_c，如式(11)所示。

$$L_{dis} = \lambda_d \left(L_a + L_c \right) \tag{11}$$

L_a 和 L_c 都为均方误差损失。

3. 实验和结果

3.1. 定量评估

实验训练所采用的数据集使用公开大规模数据集 LRS2 [19]，LRS2 数据集由 48164 个来自 BBC 电视台户外节目的视频片段组成，视频的长度在 1 秒到 6 秒之间，其训练集、验证集和测试集根据广播日期进行拆分，分别包括 45,839、1082 和 1243 个视频。本文将数据集视频通过开源 Python 库 MediaPipe 检测提取人脸图片和对应的人脸轮廓图，LRS2 数据集中提取的人脸图片分辨率统一缩放为 128 × 128，采样率处理为每秒 25 帧(FPS)，提取的音频采样率为 16 kHz。设置四种算法模型：(1) Base (IP_LAP)；(2) Base/Wav2vec Audio Encoder；(3) Base/Wav2vec Audio Encoder + Cross Attention；(4) LTFG-GAN (Base/Wav2vec Audio Encoder + Cross Attention + Conditional Convolution)分别计算评价指标 PSNR，FID，LipLMD，其中 LipLMD 为预测图像中唇部和下颚关键点与真实图像中对应关键点之间的距离误差，对比结果如表 1 所示。

Table 1. Comparison of the results of the four algorithms in the LRS2 dataset
表 1. 四种算法在 LRS2 数据集的结果对比

算法模型	PSNR↑	FID↓	LipLMD↓
Base	33.05	16.65	0.01292
Base/Wav2vec Audio Encoder	32.97	15.87	0.01103
Base/Wav2vec Audio Encoder + Cross Attention	33.89	15.16	0.01103
LTFG-GAN	34.07	14.42	0.01103

表 1 数据表明，模型(2)将基于梅尔频谱和普通卷积的音频特征提取器替换为基于 Wav2vec 和 Conformer 的音频特征提取器后，可以显著降低预测关键点和真实关键点之间的距离误差，显著提高预测关键点的准确度；模型(3)中的交叉注意力策略可以使得模型更好的融合两个模块间的信息，有助于捕捉全局鲁棒特征，避免了模型过度关注单一特征。模型(4)中使用的条件卷积使得卷积核权重可以被预测，更明确和有效地控制图像的生成过程，提升了模型的生成性能。

3.2. 定性分析

图 5 中，可以直观的对比四种方法人脸生成的效果。Wav2lip [4]模型生成的人脸口齿不清，时好时坏，嘴唇内部形变较为严重；TalkLip [20]模型生成的人脸面部较为模糊，色泽与纹理较差，口齿可见但

变为一团，清晰度较差；IP_LAP [13]的牙齿缝隙较为还原，但仍不够清晰，嘴部闭合也较差；而 LTFG-GAN 模型生成的人脸其脸部细节与色泽光彩较为还原，张嘴幅度与真实人脸几乎一致，对比其它模型生成的人脸整体质量最高，从人物还原度、图像质量、嘴唇同步来说，LTFG-GAN 模型的生成结果与先前的方法相比均有一定程度的提升。

Figure 5. Continuous facial images generated by each model
图 5. 各模型生成的连续人脸图像

4. 结论

本文通过引入基于 Wav2vec 和 Conformer 的音频编码器，交叉注意力机制和条件卷积，得到基于 IP_LAP 模型改进的 LTFG-GAN，并将模型用于基于语音驱动的人脸生成。LTFG-GAN 对于人脸视频生成有着良好的效果：一方面，音频编码器中的 Confomer 模块，有利于捕捉图像的全局和局部信息，Wav2vec 模型有着更强的音频特征提取能力；另一方面，交叉注意力机制和条件卷积充分融合原始图像信息与人脸关键点信息，使得模型生成的人脸与音频有着更高的同步性。实验结果表明，LTFG-GAN 模型生成的人脸视频音画同步性高，人脸纹理细节逼近真实图像，人脸还原度得到提高。

参考文献

[1] 年福东, 王文涛, 王妍, 等. 基于关键点表示的语音驱动说话人脸视频生成[J]. 模式识别与人工智能, 2021, 34(6): 572-580.

[2] Chung, J.S., Jamaludin, A. and Zisserman, A. (2017) You Said That? arXiv: 1705.02966.

[3] Mukhopadhyay, R., Philip, J., *et al*. (2019) Towards Automatic Face-to-Face Translation. *Proceedings of the* 27*th ACM International Conference on Multimedia*, Nice, 21-25 October 2019, 1428-1436.

[4] Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P. and Jawahar, C.V. (2020) A Lip Sync Expert Is All You Need for

Speech to Lip Generation in the Wild. *Proceedings of the* 28*th ACM International Conference on Multimedia*, Seattle, 12-16 October 2020, 484-492. https://doi.org/10.1145/3394171.3413532

[5] Chung, J.S. and Zisserman, A. (2017) Out of Time: Automated Lip Sync in the Wild. In: Chen, C.S., Lu, J. and Ma, K.K., Eds., *Computer Vision—ACCV* 2016 *Workshops*, Springer, 251-263. https://doi.org/10.1007/978-3-319-54427-4_19

[6] Cheng, K., Cun, X., Zhang, Y., Xia, M., Yin, F., Zhu, M., *et al*. (2022) Videoretalking: Audio-Based Lip Synchronization for Talking Head Video Editing in the Wild. *SIGGRAPH Asia* 2022 *Conference Papers*, Daegu, 6-9 December 2022, 1-9. https://doi.org/10.1145/3550469.3555399

[7] Suwajanakorn, S., Seitz, S.M. and Kemelmacher-Shlizerman, I. (2017) Synthesizing Obama: Learning Lip Sync from Audio. *ACM Transactions on Graphics*, **36**, 1-13. https://doi.org/10.1145/3072959.3073640

[8] Zhang, X. and Weng, L. (2020) Realistic Speech-Driven Talking Video Generation with Personalized Pose. *Complexity*, **2020**, Article ID: 6629634. https://doi.org/10.1155/2020/6629634

[9] Guo, Y., Chen, K., Liang, S., Liu, Y., Bao, H. and Zhang, J. (2021) Ad-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 5764-5774. https://doi.org/10.1109/iccv48922.2021.00573

[10] Zhang, Z., Hu, Z., Deng, W., Fan, C., Lv, T. and Ding, Y. (2023) Dinet: Deformation Inpainting Network for Realistic Face Visually Dubbing on High Resolution Video. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 3543-3551. https://doi.org/10.1609/aaai.v37i3.25464

[11] Baevski, A., Zhou, Y., Mohamed, A., *et al*. (2020) Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Proceedings of the* 34*th International Conference on Neural Information Processing System*, Vancouver, 6-12 December 2020, 12449-12460.

[12] Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., *et al*. (2021) Conformer: Local Features Coupling Global Representations for Visual Recognition. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 357-366. https://doi.org/10.1109/iccv48922.2021.00042

[13] Zhong, W., Fang, C., Cai, Y., Wei, P., Zhao, G., Lin, L., *et al*. (2023) Identity-Preserving Talking Face Generation with Landmark and Appearance Priors. 2023 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 9729-9738. https://doi.org/10.1109/cvpr52729.2023.00938

[14] Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J. and Catanzaro, B. (2018) High-Resolution Image Synthesis and Semantic Manipulation with Conditional Gans. 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 8798-8807. https://doi.org/10.1109/cvpr.2018.00917

[15] Li, J., Tu, W. and Xiao, L. (2023) Freevc: Towards High-Quality Text-Free One-Shot Voice Conversion. *ICASSP* 2023—2023 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Rhodes Island, 4-10 June 2023, 1-5. https://doi.org/10.1109/icassp49357.2023.10095191

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[17] Liu, X., Yin, G., Shao, J., *et al*. (2019) Learning to Predict Layout-to-Image Conditional Convolutions for Semantic Image Synthesis. *Proceedings of the* 33*rd International Conference on Neural Information Processing Systems*, Vancouver, 8-14 December 2019, 570-580.

[18] Johnson, J., Alahi, A. and Fei-Fei, L. (2016) Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In: Leibe, B., Matas, J., Sebe, N. and Welling, M., Eds., *Computer Vision—ECCV* 2016, Springer, 694-711. https://doi.org/10.1007/978-3-319-46475-6_43

[19] Afouras, T., Chung, J.S., Senior, A., Vinyals, O. and Zisserman, A. (208) Deep Audio-Visual Speech Recognition. arXiv: 1809.02108.

[20] Wang, J., Qian, X., Zhang, M., Tan, R.T. and Li, H. (2023) Seeing What You Said: Talking Face Generation Guided by a Lip Reading Expert. 2023 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 14653-14662. https://doi.org/10.1109/cvpr52729.2023.01408