2025 年 第 25 卷 第 13 期
2025, 25(13):05515-12

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2403342
引用格式:秦昊铭, 卜凡亮, 钟方昊. 基于音频的高分辨率人脸画像方法[J]. 科学技术与工程, 2025, 25(13): 5515-5526.
Qin Haoming, Bu Fanliang, Zhong Fanghao. Audio-based high-resolution face portrait method[J]. Science Technology and Engineering, 2025, 25(13): 5515-5526.

基于音频的高分辨率人脸画像方法

秦昊铭, 卜凡亮*, 钟方昊
(中国人民公安大学信息网络安全学院, 北京 100038)

摘 要 现有的语音驱动人脸生成方法在特征提取与生成质量上仍面临挑战,且尚未充分挖掘音频与人脸特征之间的深层关联。为解决这些问题,提出一种结合梅尔频率倒谱系数(Mel frequency cepstral coefficients,MFCC)音频特征提取与第二代样式生成对抗网络(style generative adversarial net-works 2,StyleGAN2)图像生成技术的研究方法。在音频处理方面,采用了梅尔频率倒谱系数作为特征提取方法。为了更有效地从音频中提取和传递特征,设计了一种基于 ResNet18 的残差模块,并融入了 SE(squeeze-and-excitation)注意力机制。同时对原残差块中的激活函数进行了优化改进,采用 Mish 激活函数,旨在减少深层网络中的梯度消失问题,保持特征信息的完整性并提高模型的准确性和泛化能力。采取 StyleGAN2 模型作为人脸图像的生成模型。实验结果表明,结合了设计的音频处理网络和 StyleGAN2 的人脸生成模型,在语音驱动的人脸生成任务中展现出了卓越的性能。通过综合评估 Fréchet 起始距离(Fréchet inception distance,FID)和路径长度等指标,本文方法在语音驱动的人脸生成任务中相较于现有方法,在生成质量上有显著提升,充分证明了所提方法的有效性和优越性。

关键词 语音生成人脸; 梅尔频率倒谱系数; 样式生成对抗网络; 注意力机制
中图法分类号 TP391; **文献标志码** A

Audio-based High-resolution Face Portrait Method

QIN Hao-ming, BU Fan-liang*, ZHONG Fang-hao
(School of Information Network Security, People's Public Security University of China, Beijing 100038, China)

[**Abstract**] Existing voice-driven facial generation methods still face challenges in feature extraction and generation quality, and have yet to fully explore the deep correlation between audio and facial features. To address above mentioned issues, a research approach that combines Mel frequency cepstral coefficients (MFCC) was proposedfor audio feature extraction with the image generation capabilities of the second generation of style generative adversarial networks (StyleGAN2) was proposed. In terms of audio processing, MFCC was employed as the feature extraction method. To more effectively extract and transmit features from the audio, a ResNet18-based residual module was designed and integrated with the squeeze-and-excitation (SE) attention mechanism. Additionally, the activation function in the original residual blocks was optimized and improved by using the Mish activation function, aiming to mitigate the gradient vanishing problem in deep networks, maintain the integrity of feature information, and enhance the accuracy and generalization ability of the model. The StyleGAN2 model was then utilized as the facial image generation model. Experimental results demonstrate that the integration of the designed audio processing network with the StyleGAN2 facial generation model exhibits outstanding performance in the task of voice-driven facial generation. Through comprehensive evaluation using metrics such as Fréchet inception distance (FID) and path length, the proposed method shows a significant improvement in generation quality compared to existing methods, thus fully proving its effectiveness and superiority.

[**Keywords**] voice-to-face generation; Mel frequency cepstral coefficients; style generative adversarial networks; attention mechanism

在社会科学和神经科学的研究领域,对人类感知和沟通机制的探索始终占据着核心位置。其中,声音和视觉信息的作用备受关注。声音不仅是人们沟通交流的基本媒介,还携带着丰富的个体特征信息。研究显示,在先闻其声而未见其面的情形下,人大脑便能绘制出说话者的模糊轮廓[1]。这一现象不仅揭示了声音与外貌之间的密切联系,而且强调了声音特征,如音调、节奏和口音,常常与特定

收稿日期: 2024-05-07 **修订日期**: 2025-01-10
第一作者: 秦昊铭(1998—),男,汉族,湖北宜昌人,硕士研究生。研究方向:计算机视觉、多模态学习。E-mail:2022211483@stu.ppsuc.edu.cn。
*通信作者: 卜凡亮(1965—),男,汉族,江苏徐州人,博士,教授,博士研究生导师。研究方向:计算机控制与信息处理。E-mail:bufanliang@sina.com。

的面部特征相呼应。例如,说话者声音的高低可能会引发听者对其年龄和性别的猜测,口音则可能透露其文化背景或地理来源[2]。这种声音与外貌之间的联系揭示了声音在个人特征标识中的重要作用,并指向了面部特征在声音识别过程中的关键影响。

进一步地,科学研究揭示了声音中蕴含的丰富信息。声音的频率和质地可能与说话者的年龄、性别乃至健康状况相关[3]。同样,面部特征,如骨骼结构、肌肉构造和嘴唇形状,对声音的产生具有直接影响。这些生理和生物学特征受到遗传和环境因素的双重影响,反映了个体之间以及不同群体间的多样性。特定文化或地理区域的群体可能展现出相似的发音习惯和面部特征[4]。因此,声音和面部特征之间的这种复杂关联,不仅体现在个体特征上,也映射在人类社会与文化的广袤图景之中。

在人类社会互动中,语音和面部图像扮演着至关重要的角色。对声音和面部表情的解读能力是社会沟通的基石,它帮助人们理解和回应他人的情感和意图[5]。这种能力的研究不仅对于理解人类行为至关重要,也为人工智能领域提供了重要的启示。随着技术的进步,语音和面部识别技术在人工智能中的应用日益增多,但要准确模拟人类的这一复杂能力仍是一个巨大的挑战。未来的研究需要继续探索声音和面部特征之间的关联,以及这些特征是如何在不同社会和文化背景中变化的[6]。通过深入了解这些机制,人们不仅能更好地理解人类行为和社会交往的复杂性,也能为人工智能的发展提供更加深刻的洞见。

语音驱动人脸生成作为新兴的研究领域,融合了人工智能与计算机视觉的尖端技术。主要挖掘语音与人脸之间的深层关联,以实现跨模态的生成、验证和匹配[7]。通过深入研究语音与人脸之间的内在联系,为语音驱动的人脸生成、伪造视频检测以及说话人身份识别等领域提供了创新解决方案,尤其在公共安全与公安实践中具有广泛的应用潜力。在公共安全领域,这项技术可成为警方调查的有力工具。例如,在处理电信诈骗、绑架勒索或经济犯罪等案件时,若涉及语音证据,警方可以利用语音驱动人脸生成技术,将收集到的语音数据与人脸数据库进行匹配,迅速锁定嫌疑人,从而提升破案效率,预防和打击犯罪行为,保护公民的人身安全与财产安全。同时,在身份验证、门禁安全以及智能家居等领域,语音驱动人脸生成技术同样具有广阔的应用空间。用户可以通过简单的语音指令,结合人脸识别技术,实现快速、安全的身份验证,从而访问受限区域,极大地提升了系统的便捷性与安全性[8]。然而,这项技术在公共安全领域的应用也面临一定的挑战,特别是录音设备在不同环境下的使用限制,对语音数据的收集质量有所影响[9]。

早期研究主要集中在声音特征与面部特征之间的关联上。例如,Radford 等[10]提出了通过声音识别个人面部特征的可能性。随后,深度学习的发展为这一领域注入了新的动力。生成对抗网络(generative adversarial networks,GAN)和卷积神经网络(convolutional neural networks,CNN)等深度学习模型被广泛应用于语音生成人脸技术。文献[11]展示了利用 GAN 从一段简短的音频生成对应的面部图像的研究。融合多种模态数据是提高语音生成人脸技术准确性和真实感的关键。文献[12]展示了如何利用神经网络从语音中预测人的面部特征,该研究通过分析大量的声音和面部数据,训练了一个模型,能够从一段语音中预测出相应的面部特征。为了提高模型的泛化能力和准确性,多任务学习和注意力机制被引入语音生成人脸的研究中。

GAN 的应用提高了人脸图像生成的质量与多样性。通过对抗性训练方法,生成器能够学会创造逼真的人脸图像,判别器则被训练来区分真实的图像和由生成器创造出的图像[13]。这种方法在提高生成图像的真实感方面取得了显著进展。确定声音特征和面部特征之间的映射关系是语音生成人脸技术的核心。研究者们通过分析声音的频谱特性,尝试找出与面部特征(如嘴唇、面部表情)之间的对应关系[14]。

高质量、多样化的数据集是提高语音生成人脸技术准确性的关键。研究者们通过收集和分析不同性别、年龄、种族和语言的声音和面部数据,不断优化和训练模型。如今的技术发展主要集中在提高生成图像的真实感、增强模型的泛化能力以及提升计算效率等方面[15]。例如,自注意力机制的引入有助于模型更准确地捕捉到语音与面部特征之间的复杂关系。尽管取得了显著进展[16],但语音生成人脸技术仍面临一系列挑战,如确保生成图像的多样性和逼真性、提高声音到面部特征转换的准确性以及处理大规模数据集时的计算效率问题。

针对上面提到的问题,现提出一种基于生成对抗网络的语音驱动人脸生成的方法[17]。在音频处理器中使用梅尔频率倒谱系数进行特征提取,并设计一种基于 ResNet18 结合注意力机制的新型残差模块[18]。通过增强特征传递来提升模型性能,并融合 SE(squeeze-and-excitation)注意力机制,通过全局

平均池化对特征通道进行信息汇总，随后通过压缩与扩张的两个全连接层对特征进行重新校准，以此增强模型对不同特征的敏感度。同时将原 ResNet 中的激活函数换为表现更优异的 Mish 激活函数，其平滑、自正则化和非单调性的特点有助于改善梯度流动，减少深层网络中的梯度消失问题，从而在多层传播中更好地保持和增强特征信息，这有助于模型学习恒等映射，进一步提升模型的泛化能力。因为 StyleGAN2 模型强大的特征表达的能力[19]，所以选择其作为图像生成器以确保生成高分辨率图像的视觉质量，以期为实现更强大的语音生成人脸方法提供强有力的支持。

1　算法原理

1.1　StyleGAN：革新图像生成的关键技术

StyleGAN 模型作为图像生成领域的一个突破，其最突出的贡献是引入了“样式”概念，使得对图像的高层次特征能够进行精确操控，如面部特征、发型和眼睛颜色等。该模型通过将潜在向量映射为多种样式向量，实现了对图像样式的精细控制。在这个过程中，潜在向量主要控制如纹理和颜色等低级特征，而样式向量则主导如面部结构和发型等高级特征的管理。StyleGAN 还引入了一种创新的插值方法——随机插值，它通过在不同的潜在向量之间引入变化，生成新的样式向量，从而在图像生成过程中增添更多的多样性。这一机制极大地扩展了图像生成的可能性，使得生成的图像更具个性化特征。

StyleGAN 的网络结构分为两个主要部分：Mapping network 和 Synthesis network。Mapping network 从隐藏变量 z 转换生成中间隐藏变量 $\boldsymbol{w}$，后者控制生成图像的风格。这个过程中，通过 8 个全连接层，解决了特征纠缠的问题，让生成的图像特征更加独立且清晰。Mapping network 的特点包括非线性映射、多层次映射以及可控的潜在向量，这些特性共同增加了潜在空间向量的多样性和生成图像的丰富度。Synthesis network 作为模型的核心部分，结合了 Mapping network 的潜在向量和随机噪声向量，生成最终的图像。其特点包括可变分辨率的方法、渐进式生长策略、逆卷积操作和高度可控性。这些技术使得生成的图像更加细腻自然，同时也大大提升了生成效率。StyleGAN 生成器结构如图 1 所示。

AdaIN（adaptive instance normalization）是 StyleGAN 中的一个关键模块，它的核心公式可以表述为

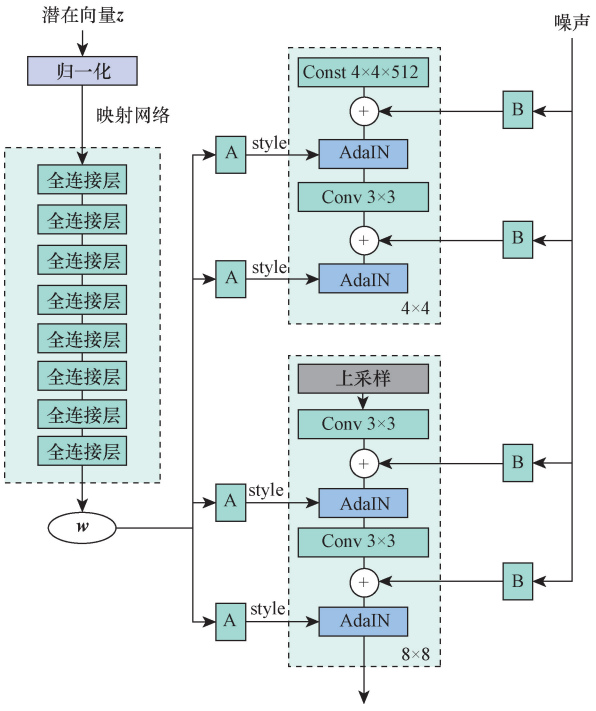

图 1　StyleGAN 生成器结构

Fig. 1　StyleGAN generator structure

$$\mathrm{AdaIN}(x_i,y)=y_{\mathrm{s},i}\frac{x_i-\mu(x_i)}{\sigma(x_i)}+y_{\mathrm{b},i} \tag{1}$$

式(1)中：x_i 为原始卷积输出；y 为由样式向量 $\boldsymbol{w}$ 经过仿射变换得到的放缩因子 $y_{\mathrm{s},i}$ 和偏差因子 $y_{\mathrm{b},i}$；$\sigma(x)$ 和 $\mu(x)$ 分别为特征图 x 的标准差和均值。

AdaIN 通过调整放缩因子，实现对图片样式信息的精确调整。此外，通过对压缩倍数 φ 的调整，可以实现对 ω 的截断，进而改变生成图片的样式，这对于处理低概率密度数据的表达能力尤为关键。

1.2　StyleGAN2：图像生成的深度创新

StyleGAN2 在原始 StyleGAN 的基础上实现了显著的技术创新。这一模型不仅提高了生成图像的质量，同时也解决了多个技术挑战。StyleGAN2 保留了 StyleGAN 的基本框架，即通过操控“样式”来生成图像，同时对其关键部分进行了重要改进。

在 StyleGAN 中 Adain 模块的运行逻辑，承接来自 ω 空间的风格调制，调制对象是来自于卷积模块的特征图像，并且特征图还加上了噪声。由于原始 StyleGAN 图像中存在伪像的一些问题，因此在 StyleGAN2 中对此做出改进。首先是 Adain 模块拆分为 Norm mean/std 模块和 Mod mean/std 模块，将噪声和偏置加入卷积层之后。但是这里同时又出现一个问题，噪声和偏差会受风格影响很大，于是之后的调整是把噪声和偏差移出来，风格调制的时候只考虑卷积模块输出的特征图，只对特征图做标准化，做完再把这俩加上去，是有改善的。

在 StyleGAN 中求标准差的时候,当特征图像中含有多个区别较大的特征层时,由于标准差的结果可能是一样的,所以特征图中的差异化和多样性很多时候被标准差屏蔽了。StyleGAN2 中改进了风格调制的逻辑,由 A 变换后得到的风格调制系数,不再直接调制主逻辑中通过卷积模块得到的这个特征图像,特征图像素里面每一个值都不会改变。通过以上改进,标准化不针对特征图里面的像素值了,而是针对卷积模块的权重。直接对权重进行调制,因为调制后影响到了其的原有结构,所以进行反调制标准化,让其收敛。调整后再去对得到的特征图进行卷积运算、输出,相当于间接用这个 Adain 模块的逻辑思想对特征图进行了运算。同时 StyleGAN2 中的归一化层经过重新设计,用于降低生成图像中某些特定模式的重复性,从而提升图像的自然度。StyleGAN2 采用了权重解耦技术,以进一步减少生成图像中的伪像。这种技术通过调整卷积层的权重来实现,有助于在不同样式的变化中保持图像质量的一致性。StyleGAN2 生成器结构如图 2 所示。

StyleGAN2 采用了一种新的正则化方法,该方法有助于模型在训练过程中保持稳定,并减少了生成图像中的伪像。这种正则化方法可以用公式表示为

$$\text{Regularization} = \lambda\left(E_z\left[\left\|\boldsymbol{J}^{\mathrm{T}}\boldsymbol{J}-\boldsymbol{I}\right\|_{\mathrm{F}}^2\right]\right) \tag{2}$$

式(2)中:λ 为正则化项的权重;$\boldsymbol{J}$ 为生成网络中某层的雅可比矩阵;$\boldsymbol{I}$ 为单位矩阵;$E_z[\cdot]$ 表示对随机变量 z 的期望;$\|\boldsymbol{J}^{\mathrm{T}}\boldsymbol{J}-\boldsymbol{I}\|_{\mathrm{F}}$ 为 Frobenius 范数,用于衡量矩阵能量和大小。

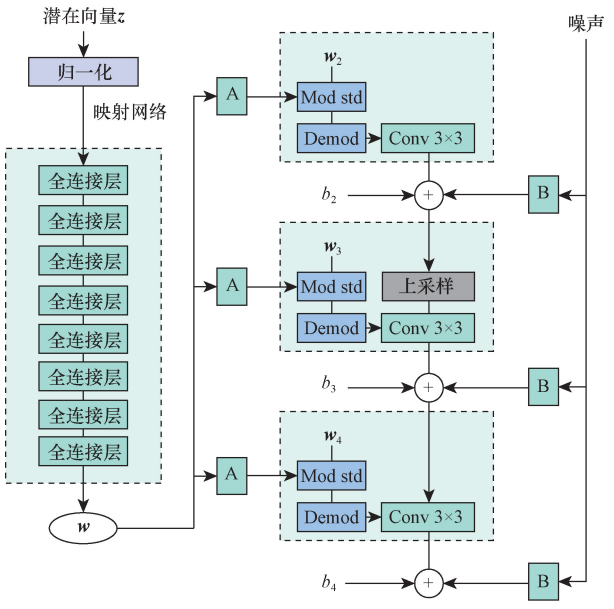

图 2 StyleGAN2 生成器结构

Fig. 2 StyleGAN2 generator structure

StyleGAN2 引入了路径长度正则化,以保证不同的样式输入能够产生可预测的变化。路径长度正则化的公式为

$$\text{Path Length} = E_{z,y}\left[\left\|\boldsymbol{y}-\boldsymbol{y}'\right\|_2\right] \tag{3}$$

式(3)中:$\boldsymbol{y}$ 和 $\boldsymbol{y}'$ 分别为原始样式向量和微调后的样式向量。

1.3 音频处理器

为了实现语音生成人脸的跨模态操作,首先要对这两种模态的特征进行提取,使其能在高维潜空间中进行匹配。开发了一个新的音频处理器模块,专注于有效地处理和转换音频信号[20]。首先,需要对音频信号进行数字化处理,通过读取 WAV 音频文件,得到一维的 NumPy 数组,它代表了音频信号的幅度随时间的变化。MFCC 特征提取如图 3 所示。

为了补偿由麦克风和声带传输引起的高频衰减,MFCC 特征提取过程从预加重开始。预加重的公式为

$$s'(n) = s(n) - \alpha s(n-1) \tag{4}$$

式(4)中:$s(n)$ 为原始信号,将 α 取值为 0.95;$s'(n)$ 为预加重后的信号;n 为当前处理信号中的第 n 个采样点。

接下来,预加重后的信号需要被分帧。信号被划分为 20 ~ 40 ms 的帧。为了防止帧与帧之间的突变引起的频谱失真,帧与帧之间需要有一定的重叠,重叠的长度是帧长的一半。

分帧后,每帧信号需要通过窗函数进一步处理,采用汉明窗(Hamming window)以减少帧两端的信号泄漏,公式为

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \tag{5}$$

处理完之后对每个窗口化的帧应用快速傅里叶变换(fast Fourier transform,FFT)来转换到频率域。FFT 的结果将用于计算 Mel 滤波器组的响应。Mel 滤波器组由多个三角带通滤波器构成,这些滤波器按照 Mel 尺度均匀分布在整个可听频率范围内。Mel 尺度的频率转换公式为

$$m = 2\,595\ln\left(1+\frac{f}{700}\right) \tag{6}$$

每个滤波器的输出是对应频率成分的能量和,之后对每个滤波器的输出取对数,得到对数能量为

$$\ln E_i = \ln\left[\sum_k |X(k)|^2 H_i(k)\right] \tag{7}$$

式(7)中:$X(k)$ 为 FFT 的复数结果;$H_i(k)$ 为第 i 个 Mel 滤波器的增益;k 为 Mel 频谱中的频率分量。

对每帧的对数能量系数应用离散余弦变换(discrete cosine transform,DCT),以提取信号的频域

特征并去相关性，即

$$c(n) = \sum_{i=1}^{N} \ln E_i \cos\left[n(i-0.5)\frac{\pi}{N}\right] \tag{8}$$

式(8)中：N 为 Mel 滤波器的数量；$c(n)$ 为 DCT 后的系数。

一阶差分计算了序列中每个点与其前后点的差异，类似于数学中的一阶导数。对于 MFCC 系数 $c(n)$，一阶差分 $\Delta c(n)$ 的计算公式为

$$\Delta c(n) = \frac{\sum_{k=1}^{K} k[c(n+k) - c(n-k)]}{2\sum_{k=1}^{K} k^2} \tag{9}$$

式(9)中：K 取值为 2，计算当前点与前后两个点的差异。式(9)确保了差分计算对时间序列的平滑变化敏感。

二阶差分是对一阶差分再次应用同样的差分计算，类似于数学中的二阶导数，反映了特征变化的加速度。二阶差分 $\Delta^2 c(n)$ 的计算公式与一阶差分类似，即

$$\Delta^2 c(n) = \frac{\sum_{k=1}^{K} k[\Delta c(n+k) - \Delta c(n-k)]}{2\sum_{k=1}^{K} k^2} \tag{10}$$

一旦计算得到 MFCC 的基本系数 $c(n)$，以及相应的一阶和二阶差分，这些特征会被组合成最终的特征向量。这个特征向量不仅包含了关于信号在频率域的静态信息，还包含了关于信号变化的动态信息。

但是通过 MFCC 所得到的语音特征向量维度比较低，为了能顺利作为 Stylegan2 生成器的输入，设计了一个基于 ResNet18 的结合自注意力机制的残差模块。使用残差模块来增强特征传递，并在其中插入自注意力机制。在残差模块中加入 SE(squeeze-and-excitation)注意力机制，这种机制通过重新校准通道的特征响应来增强模型的表现力。SE 模块首先对特征通道进行全局平均池化，然后通过两个全连接层重新校准。在残差块中使用 Mish 是为了在深层网络结构中保持和增强模型在多层传播中的特征信息，减少梯度消失。最后通过 sigmoid 激活函数得到每个通道的重要性权重。每个残差块的输出通过激活函数后与输入相加，形成残差连接。整个网络的前向传播通过连续调用各个模块的方法来实现。数据流从输入层通过展开层、残差块、注意力模块，最终通过输出层得到 512 维的输出特征。改进后 ResNet18 残差块如图 4 所示。

其中激活函数选择 Mish，因为它在各种任务中表现优异，有助于改善梯度的流动。与其他激活函数相比，Mish 激活函数的优势主要体现在其平滑性、自正则化和非单调性等方面，并且在训练过程中能够传递更多的信息，从而提高模型的准确性和泛化能力。

Mish 激活函数的数学表达式为

$$\mathrm{Mish}(x) = x\tanh[\ln(1+\mathrm{e}^x)] \tag{11}$$

式(11)中：x 为激活函数的输入；tanh 为双曲正切函数，提供了输出值的非线性范围；$\ln(1+\mathrm{e}^x)$ 为软加性(softplus)函数的输出，它提供了一个平滑的、非线性的阈值。

SE 模块包含 Squeeze 和 Excitation 两部分。在 Squeeze 部分中，特征会先进行空间维度的聚合，生成描述每个特征映射的全局描述符。这种聚合方式能沿着每个特性映射进行，进而获得全局信息的嵌入，使得网络能够利用并让所有层级都能访问到全局接收域的信息。接下来的 Excitation 部分，会利用这些嵌入为每个特征映射计算出相应的调制权值。这些权值被应用到特征映射上，形成加权特征映射，并为每个特征通道赋予相应的权重。最后，Excitation 输出的权重通过逐通道乘法的方式加权到先前的特征上，实现对原通道上特征的重标定。

SE 模块的具体结构如图 5 所示。输入的 $X \in \mathbf{R}^{H\times W\times C}$ 首先通过全局池化函数进行处理，输出一个维数为 $1\times1\times C$ 的向量。接着，使用全连接层对这个全局信息进行压缩，其中 $H\times W\times C$ 分别代表输入特征图的高度、宽度和通道数。实验表明，比例

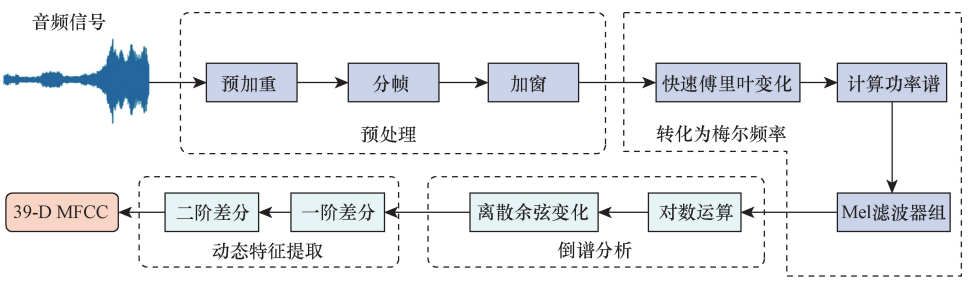

图 3　MFCC 特征提取

Fig. 3　MFCC feature extraction

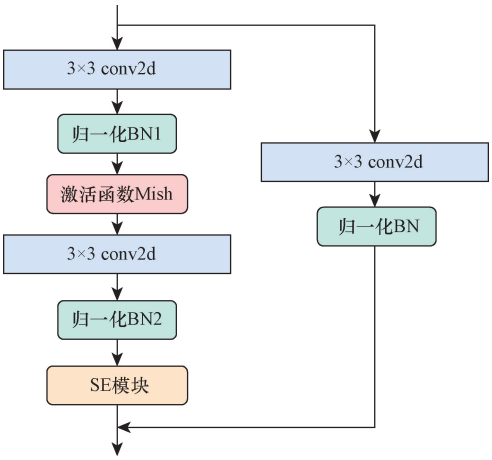

图 4 改进后的 ResNet18 残差块

Fig. 4 Improved ResNet18 residual block

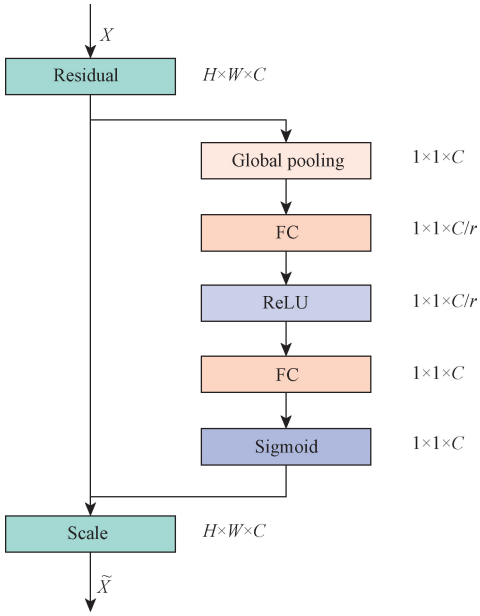

图 5 SE 模块的结构

Fig. 5 Structure of the SE module

$r=16$ 为最合适的压缩比例。之后,使用 ReLU 函数进行激活。在 Excitation 操作中,另一个全连接层用于将特征上采样到 $1\times1\times C$ 的维数,并通过 Sigmoid 函数激活,以给出每个通道的权值。输入 X 随后使用 Excitation 的输出进行重缩放,以获得加权特征映射,其输出为 $\tilde{X}\in\mathbf{R}^{H\times W\times C}$。

2 网络模型构建

2.1 改进后的网络框架

提出的网络模型是基于 StyleGAN2 框架进行的改进,用于从音频数据生成逼真的人脸图像。该模型的结构设计考虑到了 StyleGAN2 强大的特征表达的能力,以确保生成高分辨率图像的视觉质量。试验结果表明,与传统的图像处理方法相比,基于声音的方法支持更多样化和详细的信息[21]。

在音频处理器方面,采用了梅尔频率倒谱系数作为特征提取方法。为了更有效地从音频中提取和传递特征,设计了一种基于 ResNet18 的残差模块,并融入了 SE 注意力机制。同时对原残差块中的激活函数进行了优化改进,采用 Mish 激活函数,旨在减少深层网络中的梯度消失问题,保持特征信息的完整性并提高模型的准确性和泛化能力。整个音频处理网络的前向传播通过连续调用设计的模块来实现,确保了特征的有效传递和处理。

图像生成器网络由 16 个卷积层组成,分布在不同的阶段,以逐步构建和细化图像。每一层都专注于特定的特征细节,从基本的轮廓到复杂的纹理和色彩。这些层通过渐进式增长的方式逐步增加图像分辨率,从而实现从粗糙到精细的图像生成过程。在每一层都引入了样式控制,允许在不同的分辨率级别上独立控制图像的样式。这是通过将每层的输入与特定的样式向量结合来实现的。这种样式混合技术使得模型能够在不同的层级上生成多样化的图像特征,并使用了特定的网络设计来提高纹理和细节的表现。在较高层级的卷积层中,使用了 3×3 的卷积核来捕捉细微的纹理变化,而在较低层级,使用了的 4×4 或 5×5 卷积核来捕捉更为宏观的图像特征。

图像生成模型是生成逼真人脸图像的关键组件。它采用高维潜在向量来控制图像生成过程,每个维度代表图像的不同特征,如颜色、纹理和形状。通过随机采样获取这些潜在向量,并输入生成器的深度神经网络中。网络将潜在向量映射到图像空间,生成高分辨率图像。通过将音频处理器输出的特征与 StyleGAN2 的图像生成过程结合,以生成与原始音频数据匹配的人脸图像。这一过程包括使用梯度下降等优化算法来调整潜在向量,以使生成的图像尽可能接近音频处理器输出的视觉特征。

在训练过程中采用了自适应学习率调整和批归一化(batch normalization)以确保网络的稳定学习和收敛[22]。同时为了量化音频驱动的人脸生成系统的性能,采用了 Fréchet 起始距离(Fréchet inception distance,FID)作为主要评估指标。通过计算真实样本集和生成样本集之间的 Fréchet 距离,来量化它们的统计差异。整体框架如图 6 所示。

2.2 损失函数

采用的损失函数旨在优化生成器(G)和判别器(D),以生成高质量且多样化的图像。损失函数的实现在多个关键方面展开。

生成器损失包含两个主要部分:主要损失(Gmain)和路径长度正则化(Gpl)。

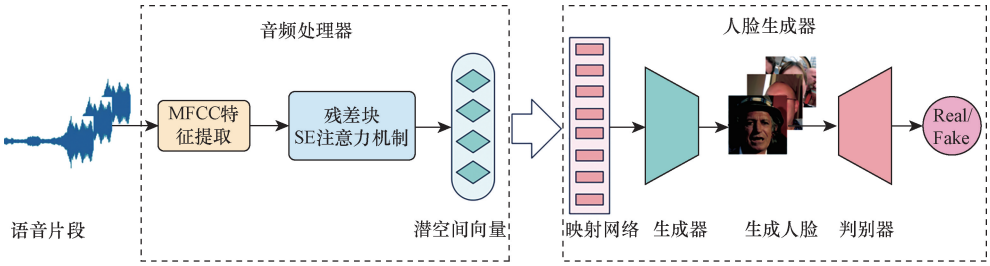

图 6　改进后的整体框架

Fig. 6　Improved overall framework

(1)主要损失(Gmain)：该损失鼓励生成器生成更逼真的图像。数学表达式为

$$L_{\text{Gmain}} = -\ln\{\sigma[G(z)]\} \tag{12}$$

式(12)中：σ 为 sigmoid 激活函数；$G(z)$ 为生成器对潜在空间向量 z 的输出。

(2)路径长度正则化(Gpl)：此项旨在使生成图像对于潜在空间的微小变化更加敏感。它计算生成图像与潜在空间的 Jacobian 矩阵的 Frobenius 范数，并将其与历史平均值进行比较。数学上，路径长度损失为

$$L_{\text{Gpl}} = (\|\boldsymbol{J}_{G(z)}\|_{\text{F}} - a)^2 \tag{13}$$

式(13)中：$\boldsymbol{J}_{G(z)}$ 为生成器输出关于 z 的 Jacobian 矩阵；$\|\cdot\|_{\text{F}}$ 为 Frobenius 范数；a 为历史平均路径长度。

判别器的目标是区分生成的图像与真实图像，同时应用 R1 正则化以稳定训练。判别器损失同样包含两个部分：主要损失(Dmain)和 R1 正则化(Dr1)。

(1)主要损失(Dmain)：判别器的主要损失由两部分组成：一部分是最小化生成图像的对数几率，另一部分是最大化真实图像的对数几率。数学表达式为

$$L_{\text{Dmain}} = -\ln\{\sigma[D(x)]\} - \ln(1-\sigma\{D[G(z)]\}) \tag{14}$$

式(14)中：$D(x)$ 为判别器对真实图像 x 的输出；$G(z)$ 为生成器的输出。

(2)R1 正则化(Dr1)：R1 正则化通过惩罚判别器输出关于真实图像的梯度的平方，来提高训练的稳定性。数学表达式为

$$L_{\text{Dr1}} = [\nabla D(x)]^2 \tag{15}$$

式(15)中：$\nabla D(x)$ 为判别器输出关于真实图像 x 的梯度。

3　实验结果与分析

3.1　实验环境搭建

模型的开发语言为 Python，实验过程在 Win11 操作系统下进行，处理器型号为 i9-12900K，显卡型号为 RTX4090(24GB)，编程平台为 Pycharm，深度学习框架为 PyTorch1.11.0。网络模型训练参数设置中，学习率为 0.05，训练轮数为 50 000。

3.2　数据集介绍

在实验中，语音记录来自 Voxceleb 数据集，人脸图像来自 VGGFace 数据集的手动过滤版本，这两个数据集都有身份标签。使用两个具有共同身份的数据集的交集，得到 1 225 名受试者的 149 354 段语音记录和 139 572 张正脸图像。为了确保实验的有效性，将数据集划分为训练集、验证集和测试集。

音频片段和人脸图像采用分离的数据预处理流水线。对于音频片段，使用语音活动检测器接口来隔离录音的语音承载区域[23]。随后使用 25 ms 的分析窗口，在帧之间跳跃 10 ms 的情况下，提取 64 维对数 MEL 谱图。实验中对每个 MEL 频段执行均值和方差归一化。同时随机剪裁 3～8 s 的音频片段进行训练，但使用整个录音进行测试。对于人脸数据，使用检测所有图像中的人脸地标。通过相似变换得到大小为 3×64×64 的 RGB 人脸裁剪图像。RGB 图像中的每个像素通过减去 127.5 然后除以 127.5 来归一化。

在实验设置中，采用了 Voxceleb 数据集的语音记录和 VGGFace 数据集的手动过滤版本的人脸图像。这两个数据集均具备身份标识信息。通过筛选两个数据集中具有相同身份的样本，共有 1 225 名受试者的 149 354 条语音记录和 139 572 张正面人脸图像。实验过程中具体的数据划分细节如表 1 所示。

对于音频和人脸图像的处理，采用了不同的预处理流程[24]。对于音频数据，借助语音活动检测器，精准地定位了录音中的语音部分[25]。随后采用

表 1　实验中使用的数据集的统计

Table 1　Statistics of the datasets used in experiments

统计项目	训练集	验证集	测试集	合计
语音段数	113 322	14 182	21 850	149 354
人脸图片	106 584	12 533	20 455	139 572
受试者数量	924	112	189	1 225

25 ms 的分析窗口,并以 10 ms 的步长提取 64 维对数 MEL 谱图。为确保数据的稳定性,对每个 MEL 频段进行了均值和方差的归一化处理。在训练阶段,随机选择 3 ~ 8 s 的音频片段进行训练,而测试时则使用完整的录音。对于人脸图像,检测并标记了所有人脸的关键点[26]。通过相似变换,得到了大小为 3 × 64 × 64 的 RGB 人脸图像。为确保数据的一致性,对 RGB 图像中的每个像素进行了归一化处理,具体操作为减去 127.5 并除以 127.5。

3.3 训练细节

在实验中,鉴别器和分类器卷积层的参数是共享的。实验中使用 ADAM 优化器,学习率为 2×10^{-4}。一节动量参数 β_1 和二阶动量参数 β_2 分别为 0.5 和 0.999。训练整个网络之前,利用 StyleGAN 2 的预训练生成器,根据学习图像的分辨率设置潜在代码的大小[27]。

为 1 024 × 1 024 的图像设置 18 × 512,为 256 × 256 设置 14 × 512。随后使用随机梯度下降(stochastic gradient descent,SGD)和余弦循环学习率调度器训练模型 10 000 次。将学习率设置为 10^{-3},动量为 0.9,权重衰减为 10^{-4}。批量大小设置为 384。具体训练细节如图 7 ~ 图 10 所示,其中 Epochs 表示训练过程中数据集被完整地遍历一遍的次数。

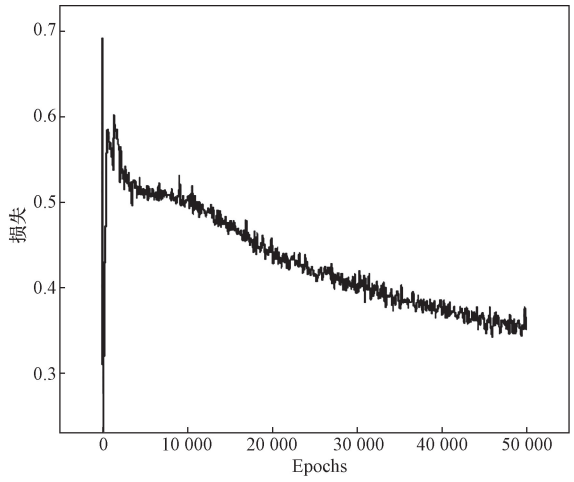

图 7 生成数据的判别器损失

Fig. 7 Loss of discriminator of generated data

3.4 实验结果的量化分析

FID 被用作量化语音驱动人脸生成系统性能的指标。FID 指标通过利用 Inception 网络提取的特征向量来量化真实样本集与生成样本集之间的统计差异。具体而言,FID 计算了两个特征分布的 FID:真实样本分布与生成样本分布,其表达式为

$$\mathrm{FID}(x,g) = \|\boldsymbol{\mu}_x - \boldsymbol{\mu}_g\|^2 + \mathrm{Tr}[\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_g - 2(\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_g)^{1/2}] \quad (16)$$

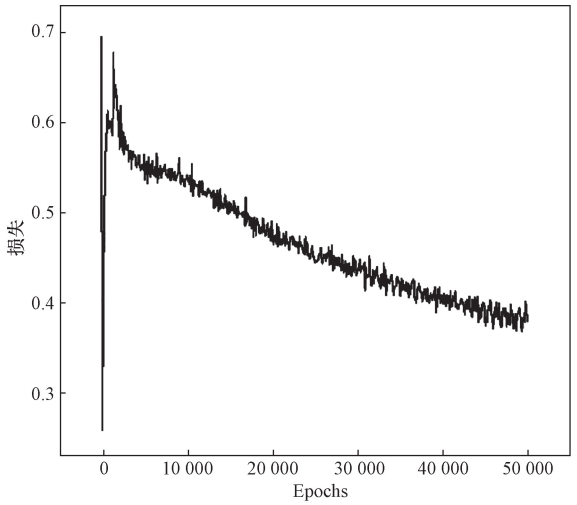

图 8 真实数据的判别器损失

Fig. 8 Discriminator loss of real data

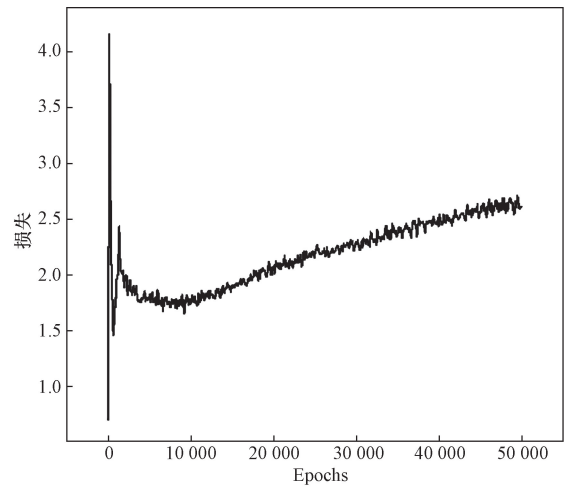

图 9 生成数据通过判别器的生成器损失

Fig. 9 The generated data is lost through the generator of the discriminator

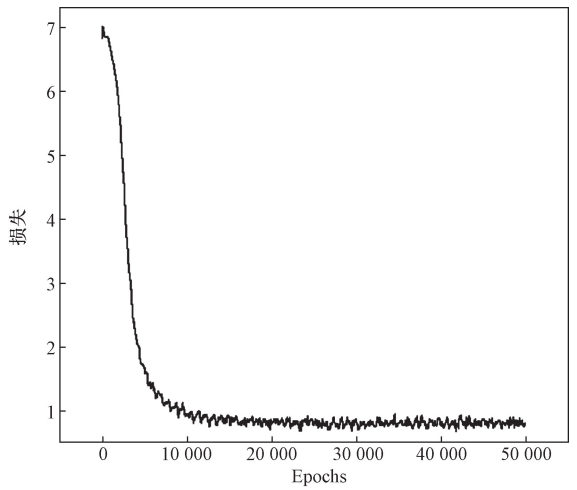

图 10 生成数据通过分类器的生成器损失

Fig. 10 The generated data is lost through the generator of the classifier

式(16)中：$\boldsymbol{\mu}_x$ 和 $\boldsymbol{\mu}_g$ 分别为真实样本和生成样本特征的均值向量；$\boldsymbol{\Sigma}_x$ 和 $\boldsymbol{\Sigma}_g$ 分别为对应的协方差矩阵。理论上，若生成样本与真实样本无差异，则 FID 为零。因此，较低的 FID 指示着生成样本与真实样本在统计特征上的高度相似性。

3.5　训练动态的详细评估

实验动态分析依据 FID 以及路径长度两个关键指标，以量化模型训练进程中生成图像质量的变化趋势。

如图 11 所示，StyleGAN、StyleGAN2 以及 DCGAN[28] 在训练过程中的 FID 趋势图提供了模型性能的直观比较。在训练的初始阶段，StyleGAN2 展现出显著的性能优势，其 FID 指标呈现出急剧下降的趋势，这表明在短时间内模型便能有效地学习到高质量的数据表示，更快速地逼近真实样本的分布。相对而言，StyleGAN 虽然也表现出 FID 的下降，但其下降速率和幅度均不及 StyleGAN2，这可能意味着 StyleGAN 在捕获数据分布的关键特征方面存在较大的局限。而 DCGAN 的下降速率和幅度更为平缓，表现不如 StyleGAN2 和 StyleGAN。

随着训练时间的延长，3 种模型的 FID 均逐渐稳定，但 StyleGAN2 的 FID 稳定值显著高于 StyleGAN，表明其在图像质量的持续优化方面具有更为明显的长期优势。此外，StyleGAN2 的 FID 在训练后期的波动性较小，这进一步证实了其在生成过程中的稳定性和可靠性。

路径长度的测量结果进一步印证了这一发现。路径长度指标反映了生成模型在潜在空间中的插值行为，是评估生成样本多样性和连续性的重要工具。在图 12 中，StyleGAN2 表现出相对平稳的路径长度(Path Length)曲线，这表明其在潜在空间中进行样本生成时能够保持高度的一致性和连续性。相比之下，StyleGAN 和 DCGAN 的路径长度在训练过程中出现了更多的波动，这可能指示了模型在潜

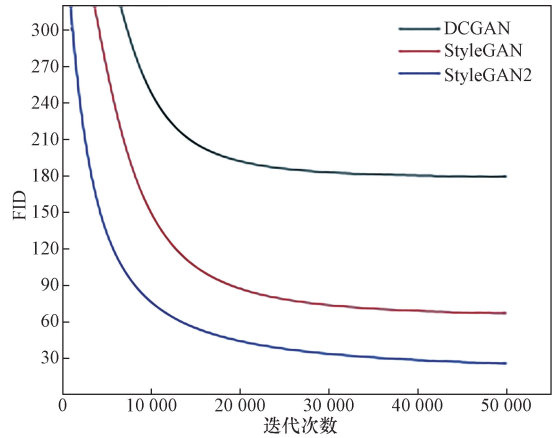

图 11　训练过程中的 FID 变化趋势

Fig. 11　Trends in FID during training

在空间的探索中存在一定程度的不稳定性，这种不稳定性可能导致生成图像在视觉上的不连贯性。

不同模型的评价指标结果如表 2 所示。

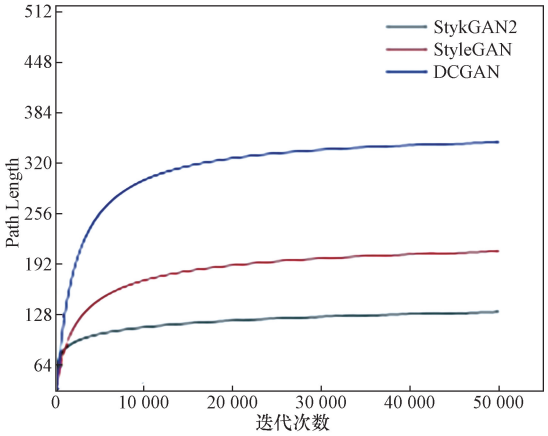

图 12　训练过程中的 Path Length 变化趋势

Fig. 12　Trends in Path Length during training

表 2　不同模型的评价指标结果

Table 2　Results of evaluation indicators for different models

模型	FID	Path Length
DCGAN	179.5	343
StyleGAN	68.2	207
StyleGAN2	28.9	131

3.6　结论性评述

各模型方法在 Voxceleb 数据集和 VGGFace 数据集上所生成的样本如图 13 所示。从实验结果可看出，使用 StyleGAN2 模型所生成的样本具有更高的视觉质量并且与原图具有更高的相似度。在评估图像生成模型的性能时，虽然通过人眼直接观察生成图像的质量是一种直观且便捷的方法，但这种方法可能因各种因素而缺乏客观性。为了确保评估的公正性，采用了多个定量评价指标，以更准确地衡量模型生成图像的质量以及与目标属性的对齐精确度。

综合 FID 及路径长度的评估结果，可以得出结论，StyleGAN2 在语音驱动的人脸生成任务中具备优越的性能。该模型不仅能够快速降低 FID，而且在生成样本的质量、多样性及连贯性方面均展现出显著的优势。此外，相比于 StyleGAN 和 DCGAN，StyleGAN2 在整个训练周期内表现出更高的稳定性和一致性。这些实验结果强有力地支持了使用 StyleGAN2 来提高语音生成人脸系统性能的决策。

3.7　消融实验研究

为了阐明模型中特定组件的贡献，进行了一系列消融研究，结果如表 3、图 14、图 15 所示。这些实

(a)StyleGAN2的生成样本

(b)StyleGAN的生成样本

(c)DCGAN的生成样本

(d)真实的样本

图 13 不同模型的生成结果和真实样本

Fig. 13 The results of the generation of different models and real samples

表 3 消融实验的结果

Table 3 Results of ablation experiments

消融实验	FID	Path Length
BaseModel-NoSE	34	154
BaseModel-ReLU	42	178
BaseModel	28.9	131

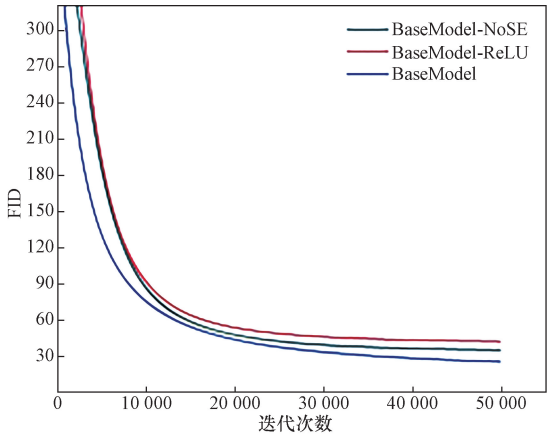

图 14 训练过程中的 FID 变化趋势

Fig. 14 Trends in FID during training

验旨在孤立关键功能和机制的效果,从而提供它们的功能重要性及对模型性能的影响力[29]。在第一个消融实验 BaseModel-NoSE 中,移除了残差块中的 Squeeze-and-Excitation(SE)注意力机制。此变体旨在评估通道级注意力对生成图像质量的影响。在第二个研究 BaseModel-ReLU 中,将所有层中的 Mish

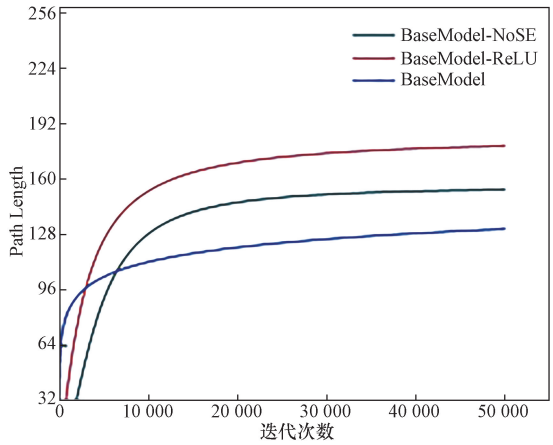

图 15 过程中的 Path Length 变化趋势

Fig. 15 Trends in Path Length during training

激活函数替换为 ReLU,以评估激活函数对网络训练动态和图像保真度的影响[30]。

实验结果显示,BaseModel-NoSE 的 FID 得分为 34,相较于 BaseModel 的 28.9 有所上升,而 Path Length 也从 131 增加至 154。这表明在缺乏 SE 注意力机制的情况下,生成图像的质量有所下降,同时生成过程的复杂性也有所提升。因此,可以推断,SE 注意力机制对于提高生成图像的质量以及优化生成路径具有重要作用。

另一方面,BaseModel-ReLU 的实验结果进一步证实了激活函数对模型性能的影响。与 BaseModel 相比,BaseModel-ReLU 的 FID 得分上升至 42,Path

Length 也增至 178。这一数据表明,ReLU 激活函数在维持图像质量和生成路径效率方面不如 Mish 函数。ReLU 函数的使用可能导致了图像质量的显著降低和生成过程的复杂化。

综上所述,实验结果强调了 SE 注意力机制和 Mish 激活函数在生成对抗网络中的重要性。SE 机制通过提高模型对重要特征的关注度,有助于提升生成图像的质量并简化生成路径。而 Mish 激活函数则以其平滑、非单调的特性,有助于模型在训练过程中更好地学习和适应数据的复杂分布。实验结果表明,SE 机制和 Mish 激活函数对于提升生成图像的质量、降低生成路径的复杂性具有重要作用。

4　结论

针对于现有的语音驱动人脸生成方法在特征提取与生成质量上仍面临挑战,尚未充分挖掘音频与人脸特征之间的深层关联等问题,提出了一种结合梅尔频率倒谱系数音频特征提取和第二代样式生成对抗网络的人脸图像生成技术的研究方法。在音频特征提取方面,通过引入改进的残差模块,优化了特征提取与传递过程,有效减少了梯度消失问题,提升了模型准确性和泛化能力。在人脸图像生成部分,利用 StyleGAN2 模型,结合高质量的数据集进行训练与验证。实验结果显示,本文方法在 FID 和路径长度等指标上的表现均优于现有方法,显著提升了语音驱动人脸生成的质量。

参　考　文　献

[1] Lewicki M S. Efficient coding of natural sounds[J]. Nature Neuroscience, 2002, 5(4): 356-363.

[2] Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning [C]//Proceedings of the 28th International Conference on Machine Learning (ICML-11). Washington: International Conference on Machine Learning, 2011: 689-696.

[3] Owens A, Efros A A. Audio-visual scene analysis with self-supervised multisensory features[J]. ArXiv, 2018: 1804.03641.

[4] Arandjelovic R, Zisserman A. Look, listen and learn[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017: 609-617.

[5] Zhou Y, Wang Z, Fang C, et al. Talking face generation by adversarially disentangled audio-visual representation [J]. ArXiv, 2019: 1807.07860.

[6] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems. Montreal: ACM, 2014: 2672-2680.

[7] Chung J S, Zisserman A. Lip reading in the wild[C]//Asian Conference on Computer Vision. Berlin: Springer, 2016: 87-103.

[8] 李俊屿, 卜凡亮, 谭林, 等. 基于多模态共享网络的自监督语音-人脸跨模态关联学习方法[J]. 科学技术与工程, 2024, 24(7): 2804-2812.
Li Junyu, Bu Fanliang, Tan Lin, et al. Self-supervised voice-face cross-modal association learning method *via* multi-modal shared network [J]. Science Technology and Engineering, 2024, 24(7): 2804-2812.

[9] Smith E R, Zaidel D W. Facial and vocal cues in perception of trustworthiness[J]. Journal of Nonverbal Behavior, 2004, 28(4): 239-262.

[10] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks [J]. ArXiv Preprint ArXiv, 2015: 1511.06434.

[11] Nagrani A, Zisserman A. Seeing voices and hearing faces: cross-modal biometric matching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018: 8427-8436.

[12] Duarte A, Hidalgo G, Lopes A T, et al. Wav2Pix: speech-conditioned face generation using generative adversarial networks[J]. ArXiv Preprint ArXiv, 2019: 1901.03396.

[13] Wen Y, Zhang K, Li Z, et al. A discriminative feature learning approach for deep face recognition[C]//European Conference on Computer Vision (ECCV). Berlin: Springer, Cham, 2016: 499-515.

[14] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 2242-2251.

[15] Suwajanakorn S, Seitz S M, Kemelmacher-Shlizerman I. Synthesizing obama: learning lip sync from audio[J]. ACM Transactions on Graphics, 2017, 36: 95.

[16] 张珂, 侯捷. 基于改进的卷积神经网络图像识别方法[J]. 科学技术与工程, 2020, 20(1): 252-257.
Zhang Ke, Hou Jie. Research on image recognition method based on improved convolution neural network [J]. Science Technology and Engineering, 2020, 20(1): 252-257.

[17] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(12): 4401-4410.

[18] 刘遵雄, 蒋中慧, 任行乐. 多尺度生成对抗网络的图像超分辨率算法[J]. 科学技术与工程, 2020, 20(13): 5217-5223.
Liu Zunxiong, Jiang Zhonghui, Ren Xingle. Image super-resolution algorithm *via* multi-scale generative adversarial networks [J]. Science Technology and Engineering, 2020, 20(13): 5217-5223.

[19] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of StyleGAN[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, IEEE, 2020: 8110-8119.

[20] Kraus N, Chandrasekaran B. Music training for the development of auditory skills[J]. Nature Reviews Neuroscience, 2010, 11(8): 599-605.

[21] Kamachi M, Hill H, Lander K, et al. Putting the face to the voice: matching identity across modality[J]. Current Biology, 2003, 13(19): 1709-1714.

[22] Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: a survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2): 423-443.

[23] Kamachi M, Hill H, Johnston A. This face sounds familiar: auditory face recognition [J]. Cognitive Science, 2008, 32(3):

496-512.

[24] Kim Y, Morikawa C, Hori T. DeepVoice: a new deep learning approach for voice-based emotion recognition[J]. IEEE Access, 2019, 7: 115141-115150.

[25] Smith E, Zhang S, Johnson S, et al. Cross-modal perceptionist: can face geometry be gleaned from voices[J]. ArXiv, 2019: 1909.04315.

[26] Abdel-Hamid O, Mohamed A R, Jiang H, et al. Convolutional neural networks for speech recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(10): 1533-1545.

[27] 冯陈定,李少波, 姚勇, 等. 基于改进卷积神经网络与动态衰减学习率的环境声音识别算法[J]. 科学技术与工程, 2019, 19(1): 177-182.
Feng Chending, Li Shaobo, Yao Yong, et al. Environmental sound recognition with improving convolutional neural networks and learning rate decay [J]. Science Technology and Engineering, 2019, 19(1): 177-182.

[28] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks [J]. ArXiv, 2015: 1511.06434.

[29] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver: IEEE, 2013: 6645-6649.

[30] Chan W, Jaitly N, Le Q, et al. Listen, attend and spell[J]. ArXiv, 2016: 1508.01211.