仅用单一短音频训练的音效生成GAN模型

姜林[1,2]， 苗向阳[3]， 洪紫荆[2]
（1. 湘江实验室，湖南 长沙 410205；
2. 湖南工商大学人工智能与先进计算学院，湖南 长沙 410205；
3. 湖南工商大学智能工程与智能制造学院，湖南 长沙 410205）

摘 要：针对音效生成中音频逼真度低、风格多样性欠缺的问题，提出了一种基于多频带注意力机制的生成对抗网络（generative adversarial network，GAN）模型。首先，采用多频带扩展模式提取不同采样率的音频特征，并引入RaHingeGAN（relativistic average hinge GAN）损失函数提高音频生成稳定性。其次，结合Transformer注意力机制增强谐波信息和频谱结构表达，并引入Alpha Dropout自适应正则层缓解过拟合。最后，设计音频风格迁移模块以增强风格可控性，在特征学习过程中，融合梅尔频率倒谱系数（Mel frequency cepstral coefficient，MFCC）、伽玛通频率倒谱系数（Gammatone frequency cepstrum coefficient，GFCC）及其多阶差分系数捕捉音频信号动态特征。实验表明，基于多频带注意力机制的GAN模型在音效逼真度、风格多样性和生成稳定性方面均优于现有模型，可有效提升音效生成质量。

关键词：音效生成；生成对抗网络；多频带；Transformer
中图分类号：TP391
文献标志码：A
doi: 10.11959/j.issn.2096−6652.202524

A GAN generation method of sound effect using only a single short audio for training

JIANG Lin[1,2], MIAO Xiangyang[3], HONG Zijing[2]
1. Xiangjiang Laboratory, Changsha 410205, China
2. School of Artificial Intelligence and Advanced Computing, Hunan University of Technology and Business, Changsha 410205, China
3. School of Intelligent Engineering and Intelligent Manufacturing, Hunan University of Technology and Business, Changsha 410205, China

Abstract: To address the issues of low audio realism and insufficient style diversity in sound effect generation, a generative adversarial network (GAN) model based on a multi-band attention mechanism was proposed. Firstly, a multi-band expansion mode was adopted to extract audio features at different sampling rates, and a relativistic average hinge GAN (RaHingeGAN) loss function was introduced to improve audio generation stability. Secondly, a Transformer attention mechanism was incorporated to enhance the expression of harmonic information and spectral structure, while an Alpha Dropout adaptive regularization layer was applied to mitigate overfitting. Finally, an audio style transfer module was designed to enhance style controllability. During the feature learning process, Mel frequency cepstral coefficient (MFCC), Gammatone frequency cepstrum coefficient (GFCC), and their higher-order differential coefficients were fused to capture the dynamic characteristics of the audio signal. Experiments demonstrated that the proposed GAN model based on the multi-

收稿日期：2025−03−23；**修回日期**：2025−05−26
通信作者：姜林，jlcdf@163.com
基金项目：湘江实验室重大项目（No.23XJ01003，No.23XJ01009）；湖南省教育厅科学研究重点项目（No.22A0441）；湖南省研究生科研创新项目（No.CX20231163）
Foundation Items: Major Project of Xiangjiang Laboratory (No. 23XJ01003, No. 23XJ01009), Major Scientific Research Project of Hunan Provincial Department of Education (No. 22A0441), Hunan Provincial Innovation Foundation For Postgraduate (No. CX20231163)

band attention mechanism outperformed existing models in terms of sound effect realism, style diversity, and generation stability, effectively improving the quality of sound effect generation.

Key words: sound effect generation, GAN, multi band, Transformer

0　引言

音效广泛应用于虚拟现实、游戏设计、电影制作等领域[1-3]。传统的音效制作多依赖素材剪辑或专业音频团队手工设计，通常耗时费力，成本相对昂贵[4]。近年来，音效生成技术利用人工智能，通过训练音频样本自动生成特定类型的音效[5]，如基于一段给定的雷电声样本，可生成不同时长或风格变化的雷电声。音效生成技术能够辅助完成音效设计，减少重复劳动，从而提高数字创意效率。因此，研究音效生成模型具有重要的应用价值[6-8]。

音效生成的训练样本通常包括音频、文本、视频和类别标签等，音频是音效生成模型学习的基础，而文本、视频、类别标签等作为条件输入，通过约束生成过程提升音效生成的可控性和针对性。Karan 等[9]仅用音频作为训练样本，采用结构化状态空间序列建模[10]，生成高质量音效。为了提高音效生成的可控性，Yang 等[11]在训练时加入文本描述信息，使用扩散模型实现了文本到音频的跨模态生成。为了实现音效生成与视频信息同步，学者们也研究了视频引导的音效生成方法[12-14]。为了生成特定音效，Liu 等[15]引入类别标签，结合离散的时频表达学习方法，建模音频样本长距离依赖关系，可生成单一音效。上述模型均需大量训练样本，且生成音效宽泛，精准性不足，如生成的雷电声中往往含“意外”声音，如爆炸声。若仅用单一短音频信号作为训练样本，既可极大减少训练样本量，又可提高音效生成的精准度。因此，使用单样本训练生成特定音效具有重要的研究价值。

使用单一短音频训练的音效生成模型亟须解决逼真度问题，比如狗叫声，听起来必须是狗叫声。现有的音效生成模型以深度神经网络为主，常用的有自回归模型、变分自编码器（variational autoencoder，VAE）、扩散模型、Transformer 模型和生成对抗网络（generative adversarial network，GAN）等。自回归模型基于时间序列，能有效捕捉音频波形的时域特征，一般从时间序列维度预测生成音频信号，音质较好，但时间复杂度高[16-17]。VAE 通过编码器学习样本潜在表达，在潜在表达空间中利用解码器采样生成相似分布的音频信号[18-19]，可提高生成信号的多样性，但生成信号的真实性难以保证。扩散模型通过前向扩散和逆向生成迭代训练，能够生成高质量音频信号。例如，AudioLDM[20]通过潜空间扩散实现文本到音频生成，Make-an-Audio[21]进一步引入提示增强机制提升多样性，但两者都高度依赖大规模训练数据和外部引导信息。基于 Transformer 的模型，如 MusicLM[22]通过长序列建模实现音乐生成，但参数量大且对短音频样本的适应性不足。GAN 模型由生成器和判别器组成，生成器试图生成逼真的音频样本，判别器努力区分真实音频和生成音频[23-24]的相似度，可无条件生成高质量音频信号。相比扩散模型、基于 Transformer 的模型，GAN 模型凭借其架构相对灵活，可根据任务调整，生成的音频信号接近真实数据分布，成为目前主流的音效生成方法之一。但主流 GAN 模型架构多以全频带信号为整体进行训练和生成[25]，该结构不能充分捕获谐波成分，导致生成的信号频谱过于平滑，音效逼真度不够，音色不饱满。现有音效内容生成方法中，Greshler 等[26]提出了多尺度 GAN 模型，使用该模型生成的音频信号在谐波和韵律上有明显提升。但是，该方法面临单一短音频训练时谐波成分捕获不充分，过拟合导致生成的信号相距预期信号较远、逼真度不够，以及训练过程不稳定等不足，其生成的音频信号仍然存在“鬼影”噪声和音色干涩的问题。

此外，使用单样本训练模型生成的音效在多样性方面同样面临挑战。现有音乐风格迁移和音效内容生成方法多采用改变网络输出长度和简单拼接策略，例如，Yang 等[11]直接将输出音频时长设定为固定长度，Greshler 等[26]通过调整模型改变音频生成长度，以上方法都会导致风格特征丢失、生成信号衔接生硬。Wu 等[27]将 Transformers 与 VAE 结合用于音乐生成，该方法融合了 Transformers 的长序列建模能力和 VAE 的用户控制特性，实现了对音乐风格和内容的高效生成。Koo 等[28]提出的端到端音乐混音风格迁移系统通过对比学习和自监督学习，实现了将输入多轨音频的混音风格转换为参考歌曲的风格。Zhang 等[29]提出的 StyleSinger 采用残

差风格适配器和不确定性建模层归一化，实现了零样本风格迁移。Wu 等[30]提出的 TransPlayer 通过自编码器模型和 Diffwave 模型，实现了一对一和多对多的音色风格迁移等。以上方法在风格迁移中的多样性控制表现优于现有基线模型，在捕捉和保持音频风格特征方面取得了一定进展，但在应对单一短音频样本的风格多样性和生成信号自然衔接时，依然面临挑战。

针对现有音效生成方法的不足，本文提出了一种多频带注意力音效生成 GAN 模型。该模型的主要创新点如下。

（1）多频带注意力谐波增强。使用多频带扩展模式分解不同采样率的音频特征，结合 Transformer 注意力机制动态提取关键谐波成分和频谱上下文结构，解决传统 GAN 模型全频带训练导致的频谱平滑、谐波不足和音色干涩问题。

（2）单样本训练优化策略。在音频重建部分引入 Alpha Dropout 自适应正则层，通过自适应丢弃神经元缓解过拟合。同时，采用 RaHingeGAN（relativistic average hinge GAN）损失函数提升生成器与判别器的对抗平衡性，显著改善训练稳定性和收敛速度。

（3）加入多特征融合的风格迁移模块。利用卷积网络联合学习风格音频与内容音频的混合特征，并在特征学习过程中，使用梅尔频率倒谱系数（Mel frequency cepstral coefficient，MFCC）、伽玛通频率倒谱系数（Gammatone frequency cepstrum coefficient，GFCC）及其多阶差分系数捕捉音频信号动态特征，增强信号抗噪性与风格保持能力，解决生成信号衔接生硬和多样性不足的问题。

1 基于多频带注意力 GAN 的音效生成模型

1.1 总体网络框架

受单图像生成对抗网络模型 SinGAN[31]和基于 GAN 的音频多尺度模型 Catch-A-waveform[26]的启发，设计了一种基于多频带 GAN 的扩展模型，该模型可从低分辨率音频逐一累积生成高分辨率音频。其总体网络框架如图 1 所示，整个模型采用金字塔式的多尺度架构，由若干对生成器和判别器模块构成。每一层负责在对应频带上进行信号的建模与重建，从最低频（500 Hz）逐层向上，逐步增强生成音频的细节和清晰度，最终在最高频带（16 000 Hz）完成整体信号的还原。各频带模块之间通过上采样与跳跃连接协同工作，实现从粗略轮廓到精细纹理的逐级增强。该模型的输入来自稳态音频源的短样本 x，旨在从源分布中学习并生成新的随机样本。在训练阶段前，先对输入信号 x 进行规范化（即其最大绝对值为 1），再单独选择每个信号的初始频带（即频带 N）。

（1）下采样：构建训练信号 X。下采样过程可用数学表达式表示为：

$$\begin{aligned} x_0 &= x\downarrow \\ x_n &= \left(x\cdot h_n\right)\downarrow d_n, n=1,\cdots,N \end{aligned} \tag{1}$$

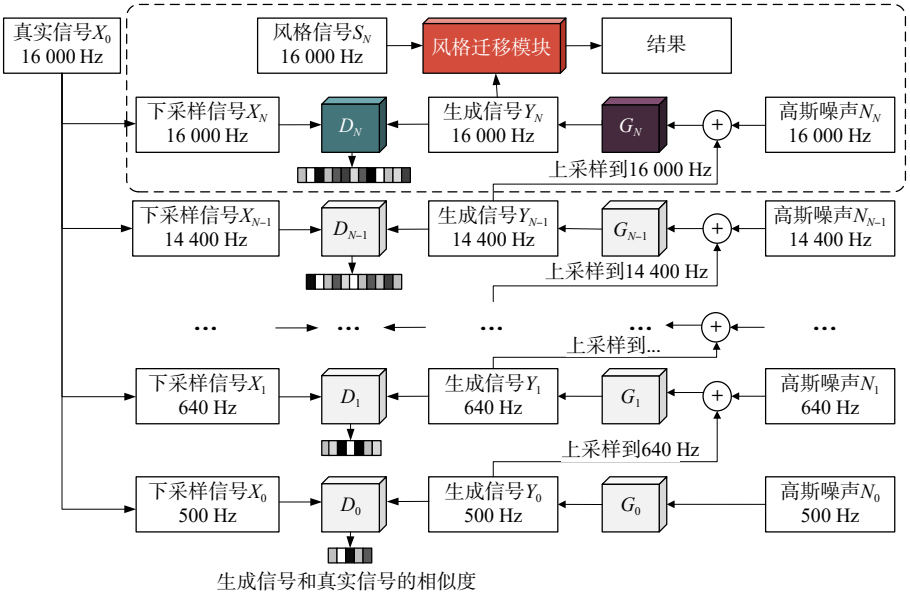


图 1 总体网络框架

其中，d_n是下采样因子，↓表示下采样过程，h_n是相应的抗混叠滤波器。

x的采样频率为f^s（实验中设置为16 000 Hz），第n层采样频率为$f_n^s=f^s/d_n$。类似地，$\tilde{x}_n$表示假信号$\tilde{x}$的第n层频带表示。

（2）上采样：在每个采样率级别上依次生成其前一个采样率级别，按照从低到高的顺序执行生成假样本$\tilde{x}$。上采样过程可用数学表达式表示为：

$$\begin{aligned}&\tilde{x}_N=G_N\left(N_N\right)\\&\tilde{x}_n=G_n\left(N_n,\left(\tilde{x}_{n+1}\right)\uparrow^{\alpha_n}\right),n=N-1,\cdots,0\end{aligned}\tag{2}$$

其中，N_N是高斯噪声；G_n是一个卷积神经网络生成器；上采样因子α_n是频带n+1和频带n之间的采样比率，$\alpha_n=d_{n+1}/d_n$；$(\cdot)\uparrow^{\alpha}$表示通过立方插值上采样α倍；$\tilde{x}_0$是最终的生成样本。

初始频带可以保持音频信号低频的远距离依赖性，每个后续生成器只需要在前一个频带生成信号中添加一个窄频带，便可形成最终完整频带。

1.2　感知生成器、判别器设计

所有频带上的生成器（G_N）和判别器（D_N）都具有相同的全卷积网络框架。生成器、判别器网络结构如图2所示，其中，自注意力特征提取层、韵律恢复层和自适应正则层的关键模块细节相同，以判别器为例，如右侧子图所示。

首先，使用初始膨胀因子为1的膨胀卷积块和注意力编码层组成自注意力特征提取层。其次是韵律恢复层，由6个依次堆叠的膨胀卷积块组成，每个膨胀卷积块叠加批归一化和斜率为0.2的leaky ReLU单元，每层的膨胀因子为前一层的2倍。其中生成器的末尾使用由tanh和sigmoid的逐元素乘积组成的门控激活单元，该1×1卷积块为额外的非膨胀卷积。所有卷积层的卷积核大小均为9，每个频带的总感受野为2 040，并使用权重归一化改善训练稳定性。接着，在训练块的末尾添加自适应正则层，即将Alpha Dropout与SELU激活函数配合使用，并用2个全连接层将其连接。最后，使用一个固定的非线性预加重（pre-emphasis，PE）滤波器

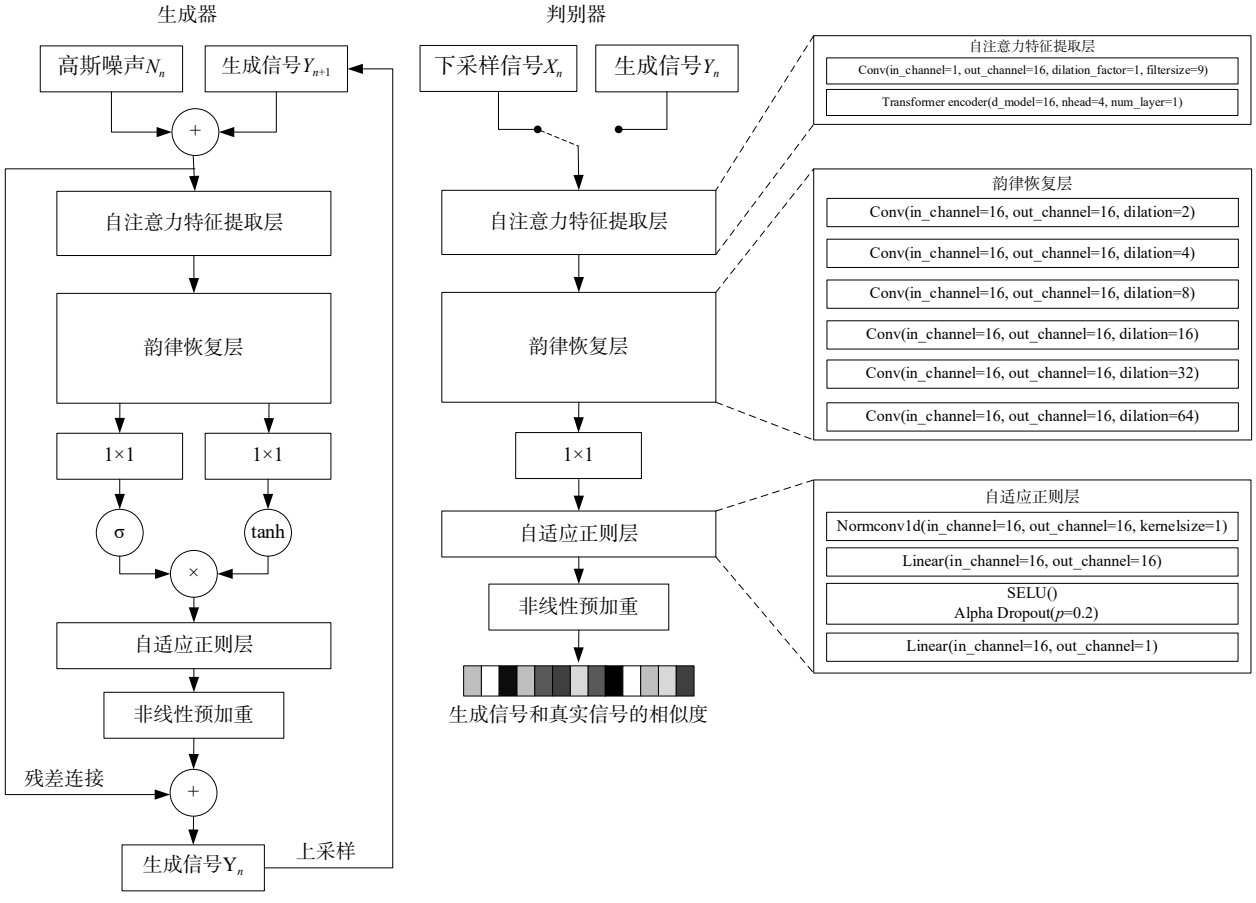


图2　生成器、判别器网络结构

增强高频率。

本节重点介绍自注意力特征提取层、自适应正则层和重建感知损失。

训练第n层生成器的目标是，驱动生成信号$\tilde{x}_n$中长度为T的段的分布尽可能地接近下采样到第n频段的真实信号x_n中长度为T的段的分布。使用一个判别器网络D_n，将其输入中重叠的长度为T的窗口中的每个窗口分类为真或假，使其输出成为与输入相同长度的分类序列。判别器的最终判别分数是该分类序列的平均值。总体而言，模型训练旨在解决以下优化问题：

$$\min_{G_n}\max_{D_n} L_{D_n}^{\text{HingeGAN}}+L_{G_n}^{\text{HingeGAN}}+L_{\text{rec}}\left(G_n\right) \tag{3}$$

其中，$L_{D_n}^{\text{HingeGAN}}$、$L_{G_n}^{\text{HingeGAN}}$、$L_{\text{rec}}$分别表示第$n$频带辨别器的RaHingeGAN损失函数、第$n$频带生成器的RaHingeGAN损失函数、重建感知损失函数，详细见1.2.3节。

1.2.1 自注意力特征提取层

自注意力特征提取层是Transformer架构中的一个关键组件，用于处理自然语言和序列到序列任务。这一标准的编码器层基于Vaswani等[32]的工作，主要包括以下要素：

（1）多头自注意力：计算查询（$\boldsymbol{Q}$）、键（$\boldsymbol{K}$）、值（$\boldsymbol{V}$）的点积，以集中模型在输入序列不同部分的注意力。这一要素允许模型同时处理序列中多个位置的信息，并通过以下表达式进行计算：

$$\text{Attention}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{4}$$

其中，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\in\mathbb{R}^{n\times d_k}$，$d_k$是键的维度。

计算得到每个位置的注意力权重分布，然后将权重分布应用于值序列。多头注意力的输出为拼接每个注意力头的结果，其数学表达式如下：

$$\text{MultiHead}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{Concat}\left(\text{head}_1,\cdots,\text{head}_h\right)W^O \tag{5}$$

每个头的计算为：$\text{Attention}\left(\boldsymbol{Q}W_i^{\boldsymbol{Q}},\boldsymbol{K}W_i^{\boldsymbol{K}},\boldsymbol{V}W_i^{\boldsymbol{V}}\right)$。这种并行计算不同于单注意力头的机制，提高了模型对序列全局关系的建模能力。

（2）前馈神经网络：经过多头自注意力计算后，每个位置的输出进入一个两层全连接神经网络，如下：

$$\text{FFN}\left(Z\right)=\max\left(0,Z\boldsymbol{W}_1+\boldsymbol{b}_1\right)\boldsymbol{W}_2+\boldsymbol{b}_2 \tag{6}$$

其中，$\boldsymbol{W}_1$、$\boldsymbol{W}_2$是可学习的权重矩阵，$\boldsymbol{b}_1$、$\boldsymbol{b}_2$是偏置向量。

（3）残差连接和层归一化：为了解决梯度消失问题并加速训练过程，每个子层的输出与其输入进行残差连接，并接受层归一化处理。

1.2.2 自适应正则层

在标准Dropout中，每个神经元都有一定的概率被随机丢弃，即在训练期间将其输出设置为0。这样做有助于减少神经网络对特定神经元的过度依赖，提高网络泛化能力。然而，当使用ReLU激活函数时，标准Dropout可能引入一些问题，因为它将激活值直接置为0。

Alpha Dropout是一种经过改进的自适应正则化技术，其设计目的在于保持自归一化的特性。自归一化主要通过选择特定的激活函数和初始化方法来实现保持均值和标准差的稳定性，有助于避免网络在训练过程中出现过拟合或梯度消失等问题。通常，Alpha Dropout与SELU激活函数一起使用，以确保输出具有该特性。

在训练期间，对于输入x，Alpha Dropout通过从伯努利分布中抽取样本来随机屏蔽一些元素。每次前向调用都会随机选择要屏蔽的元素，并进行缩放和偏移，以保持0均值和单位标准差。具体而言，对于每个元素x_i，按照概率p进行屏蔽（置0），如下：

$$\text{AlphaDropout}(x_i)=\begin{cases}0, & \text{以概率 } p\\ \dfrac{x_i}{1-p}, & \text{以概率 } 1-p\end{cases} \tag{7}$$

其中，p是Alpha Dropout的丢弃概率，x_i是输入张量的第i个元素。

在屏蔽的情况下，元素被置0；在不屏蔽的情况下，元素按$1/\left(1-p\right)$比例进行缩放。该缩放比例基于期望不变性。在Dropout操作中，按概率p将神经元输出置0，若不对未被屏蔽的部分进行缩放，会导致整体输出的期望值降低。为了使Dropout后的输出期望与输入相同（即保持激活分布的均值不变），令输出期望$\text{E}\left[x'\right]$为：

$$\text{E}\left[x'\right]=p\cdot 0+\left(1-p\right)\cdot\frac{x}{1-p}=x \tag{8}$$

由此可知，非0部分需要乘以$1/\left(1-p\right)$才能保持期望不变。

1.2.3 重建感知损失

判别器D_n每轮交替执行更新为$\tilde{D}_n$，该步骤包括最小化梯度惩罚项，并在生成器G_n上执行一次更

新步骤。式（3）中前2项（$L_{D_n}^{\text{HingeGAN}}$和$L_{G_n}^{\text{HingeGAN}}$）为RaHingeGAN损失及梯度惩罚，具体的对抗损失定义如下：

$$\begin{aligned}L_D^{\text{HingeGAN}}=&\mathrm{E}_{x_n\sim \mathrm{P}_{x_n}}\left[\max\left(0,1-\tilde{D}_n\left(x_n\right)\right)\right]\\&+\mathrm{E}_{\tilde{x}_n\sim \mathrm{P}_{\tilde{x}_n}}\left[\max\left(0,1-\tilde{D}_n\left(\tilde{x}_n\right)\right)\right]\\&+\lambda\mathrm{E}_{\hat{x}\sim \mathrm{P}_{\hat{x}}}\left[\left(\left\|\nabla_{\hat{x}}C\left(\hat{x}\right)\right\|_2-1\right)^2\right]\end{aligned}\tag{9}$$

$$\begin{aligned}L_G^{\text{HingeGAN}}=&\mathrm{E}_{\tilde{x}_n\sim \mathrm{P}_{x_n}}\left[\max\left(0,1-\tilde{D}_n\left(\tilde{x}_n\right)\right)\right]\\&+\mathrm{E}_{x_n\sim \mathrm{P}_{\tilde{x}_n}}\left[\max\left(0,1-\tilde{D}_n\left(x_n\right)\right)\right]\end{aligned}\tag{10}$$

$$\begin{aligned}\tilde{D}_n\left(x_n\right)&=C\left(x_n\right)-\mathrm{E}_{\tilde{x}_n\sim \mathrm{P}_{\tilde{x}_n}}C\left(\tilde{x}_n\right)\\\tilde{D}_n\left(\tilde{x}_n\right)&=C\left(\tilde{x}_n\right)-\mathrm{E}_{x_n\sim \mathrm{P}_{x_n}}C\left(x_n\right)\end{aligned}\tag{11}$$

其中，x_n、$\tilde{x}_n$分别属于真实音频分布P_{x_n}和生成音频分布$\mathrm{P}_{\tilde{x}_n}$。$\mathrm{P}_{\hat{x}}$是$\hat{x}=\varepsilon x_n+(1-\varepsilon)\tilde{x}_n$的分布，$x_n\sim P_{x_n}\in[0,1]$,$\tilde{x}_n\sim P_{\tilde{x}_n}\in[0,1]$,$\varepsilon\sim U\in[0,1]$。

除对抗损失函数外，表达式（3）中第3项L_{rec}为重建感知损失函数，其使用基于生理和心理声学模型的特征，如感知响度（loudness）、音调（pitch）和音色（timbre），以更好地捕捉人类的听觉感知。其数学表达式如下：

$$L_{\text{rec}}\left(G_n\right)=\alpha_1 Px_n-\tilde{x}_n^r P_2^2+\frac{\alpha_2}{M}\sum_{m=0}^{M-1}\left\|\left|\mathrm{STFT}_m\left(x_n\right)\right|-\left|\mathrm{STFT}_m\left(\tilde{x}_n\right)\right|\right\|_2+\alpha_3 L_{\text{perceptual}}\tag{12}$$

$$L_{\text{perceptual}}=\left\|\mathrm{MFCC}\left(x_n\right)-\mathrm{MFCC}\left(\tilde{x}_n\right)\right\|_2^2+\frac{1}{M}\sum_{m=0}^{M-1}\left\|A_m\left(x_n\right)^2-A_m\left(\tilde{x}_n\right)^2\right\|_2^2+\left\|\mathrm{Pitch}\left(x_n\right)-\mathrm{Pitch}\left(\tilde{x}_n\right)\right\|_2^2\tag{13}$$

其中，使用MFCC表示音色特征，因为MFCC能够很好地捕捉音色的细节；使用加权滤波器$A(x)$计算响度，这种滤波器能更好地反映人耳对不同频率声音的感知；M表示采样点样本个数；音调Pitch(x_n)使用torchaudio库的detect_pitch函数检测，捕捉音高信息。使用上述感知相关特征构建更有效的听音感知损失函数，可有效捕捉音频信号的感知差异，提高模型生成音频的质量。

在实践中，通常只使用表达式（12）中的其中一项（将另一个系数α_i设为0），具体取决于声音类型。选择每个频带的特定输入x_n^r，并确保其对应的重建信号$\tilde{x}_n^r=G_n\left(x_n^r\right)$与该频带的真实信号$x_n$接近，即确保了模型的潜在空间中至少有一点映射到真实信号。通过以上改进得到基于多频带注意力GAN的音效生成模型（multi-band sound effect GAN，MuSEGAN）。

1.3　基于风格迁移的音效多样性调制

在多频带注意力GAN的高效生成模型的基础上加入风格迁移模块，构建基于风格迁移的音效多样性调制模块。基于上述生成器模型得到的信号生成带有新风格的音效信号，音频风格迁移框架如图3所示。音效多样性调制模块接收来自两个主要组成部分的输入：一个是新引入的风格信号（风格表示），另一个是内容信号（内容表示），即第1.2节中GAN最终的生成信号。输出是一个风格转换后的音频信号，它结合了输入内容的语义特征和目标风格的特征。该信号将保持内容音频的结构和语义信息，同时融入目标风格的音效特性。

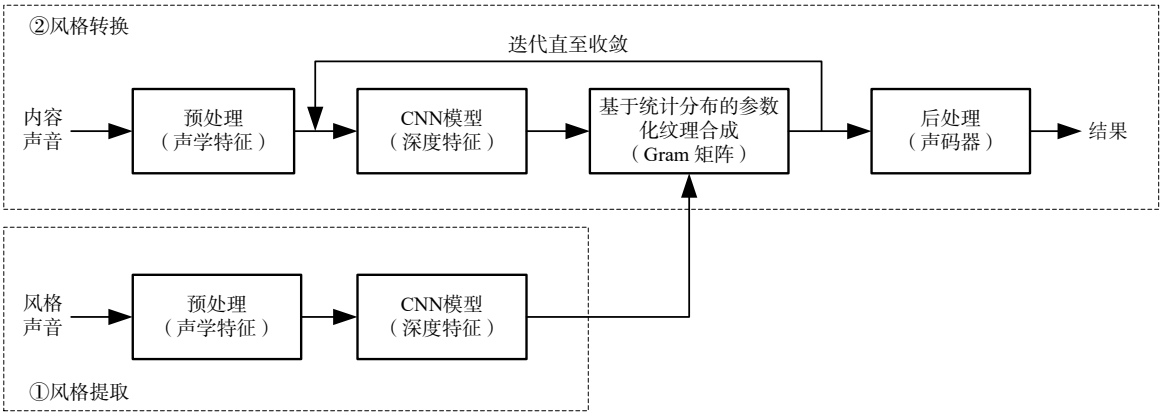


图3　音频风格迁移框架

（1）模型选择：考虑到音频信号在空间维度上与图像存在显著差异，因此采用了一种结构简单但滤波器数量较多的卷积神经网络（convolutional neural network，CNN）模型。其中未使用任何预训练权重，保持编码器的通用性与无偏性。

（2）预处理：音频信号需先从时域转换至频域，以更好地提取与音频感知相关的频谱信息。传统的静态特征在特征提取过程中对低频信号具有较强的分辨能力，但在高频部分的区分能力较弱，难以全面捕捉音频的动态变化。针对音频风格迁移模型中鲁棒性与抗噪性不足等问题，李牧等[33]提出了一种名为H-MGFCC的算法，该算法提取到的动态特征参数提高了特征提取的准确性。为进一步增强模型对动态信息的感知能力，本文在特征提取阶段进行了如下改进：首先，对原始音频信号进行短时傅里叶变换（short-time Fourier transform，STFT）以获得频域表示，即STFT_i；随后，分别提取MFCC与GFCC，即MF_i和GF_i，并对其进行一阶、二阶等多阶差分运算，获得动态变化特征，即$\mathrm{D^1MF}_i$、$\mathrm{D^2MF}_i$、$\mathrm{D^1GF}_i$、$\mathrm{D^2GF}_i$；最终，将STFT频谱特征、MFCC、GFCC及其多阶动态差分参数组合构成最终的输入特征向量，即$\mathrm{D^{12}}$-MGFCC算法。参数如下：

$$\boldsymbol{O}_{\mathrm{mix}}=\begin{bmatrix}\mathrm{STFT}_1 & \mathrm{STFT}_2 & \cdots & \mathrm{STFT}_m\\ \mathrm{MF}_1 & \mathrm{MF}_2 & \cdots & \mathrm{MF}_m\\ \mathrm{D^1MF}_1 & \mathrm{D^1MF}_2 & \cdots & \mathrm{D^1MF}_m\\ \mathrm{D^2MF}_1 & \mathrm{D^2MF}_2 & \cdots & \mathrm{D^2MF}_m\\ \mathrm{GF}_1 & \mathrm{GF}_2 & \cdots & \mathrm{GF}_m\\ \mathrm{D^1GF}_1 & \mathrm{D^1GF}_2 & \cdots & \mathrm{D^1GF}_m\\ \mathrm{D^2GF}_1 & \mathrm{D^2GF}_2 & \cdots & \mathrm{D^2GF}_m\end{bmatrix} \tag{14}$$

提取到的粗糙混合特征向量经过维度处理后，被送入神经网络进行训练，以提取更高级的音频向量特征。所得向量表示如下：

$$\boldsymbol{H}_v=\mathrm{CNN}_v\left(x_2\right) \tag{15}$$

（3）风格损失：内容音频主要用于辅助生成风格化的新音频样本，其具体内容信息并不作为优化目标，因此可在损失函数中忽略内容保持约束，仅以风格匹配为优化目标，即总损失为风格损失。对于风格样式提取，使用Gram矩阵$\boldsymbol{G}\in\mathbb{R}^{N\times N}$：

$$\boldsymbol{G}_{ij}=\sum_k \boldsymbol{F}_{ik}\boldsymbol{F}_{jk} \tag{16}$$

其中，$\boldsymbol{G}_{ij}$是特征映射i和j的内积，N是特征映射的数量积。设$\vec{a}$是风格音频，$\vec{x}$由最先的白噪声音频转变成最终的音频（即生成音频）。A和X为风格音频与白噪声音频的风格表达。总损失即风格损失：

$$L_{\mathrm{total}}=L_{\mathrm{style}}\left(\vec{a},\vec{x}\right)=\frac{\sum_{i,j}\left(X_{ij}-A_{ij}\right)^2}{4N^2M^2} \tag{17}$$

（4）后处理：将音频从频域转换回时域，使用声码器（Griffin-Lim算法）[34]重构。通过以上改进，得到引入风格迁移多样性的多频带GAN（multi-band sound effect GAN-style transfer，MuSEGAN-ST）。

2 实验

2.1 实验设置

（1）实验平台：实验采用PyTorch深度学习框架，Ubuntu 18.04 LTS操作系统。硬件使用含15个虚拟CPU的Intel(R) Xeon(R) Platinum 8358P处理器和Nvidia A40显卡，具有80 GB的内存和48 GB的显存。

（2）数据集：训练样本包含常见的10种音效类别，分别为小提琴声、钢琴曲、犬吠声、男性人声、女性笑声、蝉鸣声、风声、欢呼笑声、装弹及子弹发射声、子弹连续发射及人呼喊声。所有训练信号的采样率为16 kHz，时长为5～25 s。样本来源于Freesound网站[35]。

（3）训练设置：对于模型训练，实验使用AdamW优化器[36]，参数（β_1，β_2）=（0.5，0.999），初始学习率为0.0015，训练到2/3阶段时将学习率缩小为原来的10%。训练过程共运行3 000个epochs。在所有数据集上使用相同的参数设置，以确保对比实验的一致性和可比性。本文在基线的基础上，在以下几个应用中进行测试，并在定性和定量方面进行评估。在乐器音乐、自然环境声无条件生成中，训练信号长度为25 s，并在测试时加入适当长度的噪声信号以生成各种长度的信号；在动物声、非自然环境声无条件生成中，训练信号长度为5 s，以加入噪声的方式生成10～20 s的随机样本信号；在人声（即语音信号）无条件生成中，训练信号长度为20 s，生成长度在20～60 s的随机样本信号。

（4）模型评估：由于用单一音频训练的音效生成仍然是一个较为小众的研究领域，存在比较对象不足的问题，也缺少广泛的比较基准。为了弥补这一不足，参考已有的基线研究并结合音效的主观评测方法进行实验设计[26,37]。主观测试邀请了30名年龄在18岁以上的大学生和研究生进行试听测试。

参与者听取音效生成样本，并就样本的可信度和变化度给出评分。主观测试对声音的可信性和变化性进行评估。可信性指声音样本的真实性，即在相应的情境下，声音是否听起来像是真实发生的；变化性则指的是声音样本在播放过程中呈现的多样性，即声音是否有足够的变化来模拟真实世界的声音动态。

为了验证本研究提出的模型性能，对比了原始录音（Real）、基线系统（CAW[26]，NeurIPS 2021）、AudioLM[13]（TASLP 2023），以及本文提出的MuSE-GAN、MuSEGAN-ST这5个系统的表现。Real类别的声音来自Freesound训练示例，直接使用或根据需要进行适当处理（如裁剪）而得。评估过程中，每个类别的5种声音被组成一组，参与者使用1（完全不可信/不变化）至7（完全可信/适当变化）的等级来对声音进行评分[37]。为了避免评分重叠，个体评分采用抖动处理，并且评分范围限定为整数值。

2.2　实验结果与分析

本文通过主观视（展示频谱图）听（音频示例）感知，以及开展听觉实验，对模型效果进行综合分析，包括：（1）展示10种音效类别的原始训练信号、基线模型生成信号以及MuSEGAN生成信号的频谱图对比；（2）进行消融实验对比；（3）开展听觉测试实验，含两类音效与AudioLM模型[18]的对比分析，即3类结果分析论证本文所提方法的优势。频谱图中的横轴代表时间，纵轴代表频率，像素点的值对应音频信号的能量值。其中，MuSEGAN-ST模型（带风格迁移模块）生成的音频在频谱特征上已发生风格化转变，其频谱图与原始音频及基线模型存在本质差异，故于消融实验中展示其在不同特征学习方法上的差异。

2.2.1　10种音效类别训练信号结果展示

本文对比了原始录音（Real）、CAW及MuSE-GAN在处理10种音效类别时的效果，具体为小提琴声、钢琴曲、犬吠、男性人声、女性笑声、蝉鸣声、风声、欢呼笑声、装弹及子弹发射声、子弹连续发射及人呼喊声。10种音效类别生成信号频谱图对比如图4所示。

由图4可知，原始录音（Real）展示了丰富的频谱细节和谐波结构。CAW系统在各类音频中出现不连续性和能量分布不均，导致音质不自然。MuSEGAN在保留频谱结构和减少噪声方面表现较好，但在某些频率上仍有细节损失。总体而言，MuSEGAN在保持声音自然性和细节方面优于CAW，但仍需改进高频细节处理以提升音频质量。

2.2.2　消融实验结果对比

（1）依次加入RaHIinEGAN对抗损失函数、注意力层模块和自适应层模块的重建信号和生成信号对比

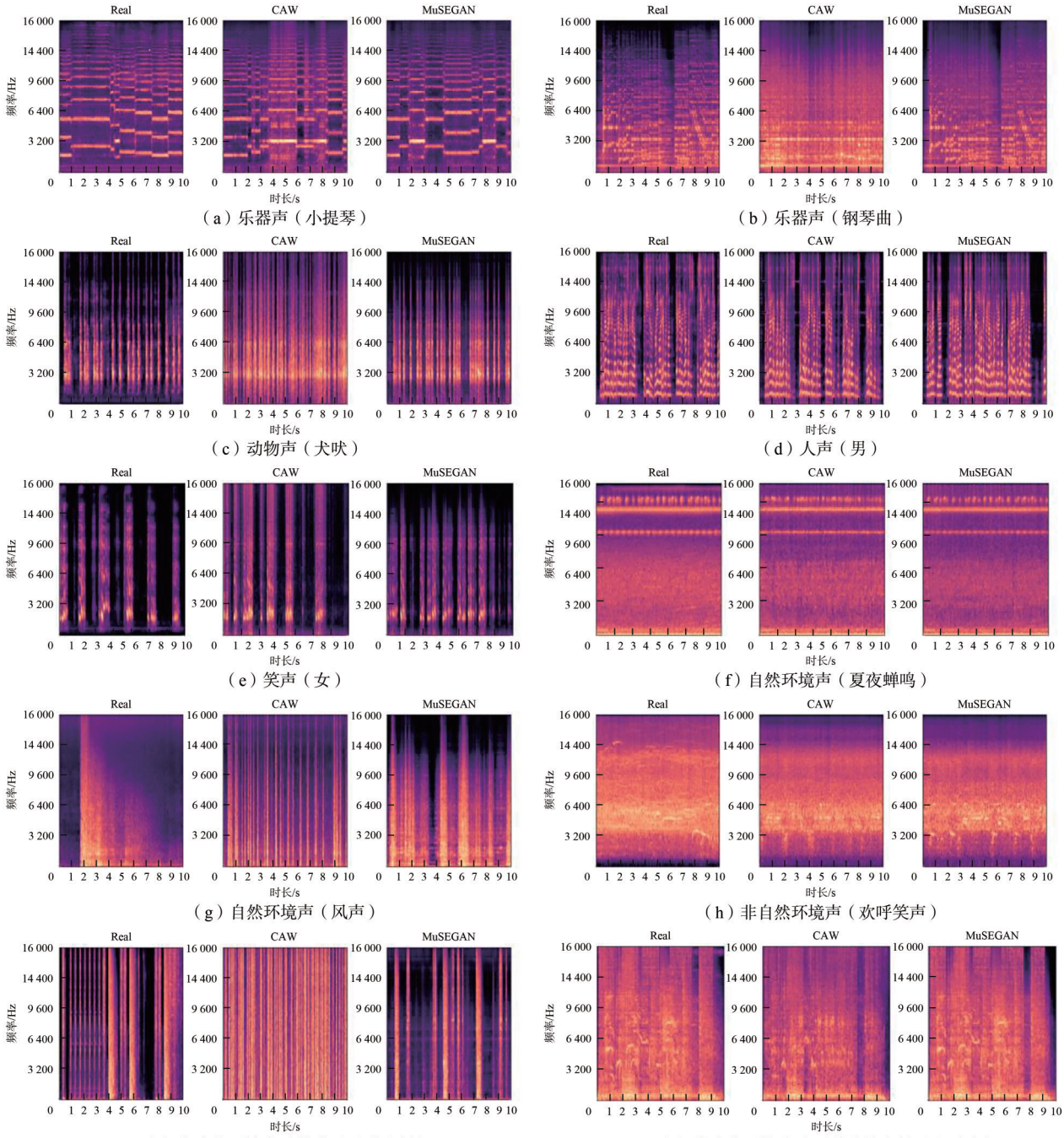
实验中引入重建感知损失函数，即信号在时域和频域重建时做优化迭代。重建信号是与原始真实信号比对，利用重建损失函数让判别器更精确地学到原始整段音频的主要结构，并通过生成器一比一复现出来。不同模块下重建信号频谱图对比如图5所示。实验中的生成信号是多频带GAN最终的生成结果，具有初步的多样性变化，与原始真实信号不同。在此基础上，实验信号采样率均为16 kHz，分别以乐器音乐和人声（即语音信号）无条件生成，对比不同模块下生成信号频谱图，分别如图6和图7所示。其中，乐器音乐无条件生成：以一段25 s单音小提琴音乐为实验示例，生成的信号听起来像是对原始作品的变异或天真的即兴演奏；人声（即语音信号）无条件生成：以一段20 s男性演讲为实验示例，生成的信号保留了说话者的声音，但展现了音节、单词、语调和静默间隙的新组合。

图5表明，优化对抗损失函数后的重建信号（如图5（c））明显比基线模型（如图5（b））更接近于原始真实信号（如图5（a））。从图6和图7看出，基线CAW模块的生成信号存在大量噪声，优化对抗损失函数后的生成信号（如图6（c）、图7（c））在频谱结构、能量分布和纹理细节方面明显比基线模型（如图6（b）、图7（b））具有更高的完整性和保真度。加入自注意力特征提取层和自适应正则层（如图6（d）、图7（d））模块的重建及生成信号均在高频部分更清晰。

（2）加入自注意力特征提取层后模型的感受野对比

由于模型没有语言概念，生成的信号不一定是可解释的。生成信号的时间相干性通过改变模型的感受野进行控制。例如，将感受野从4 s减小到2 s会使结构变得不一致，并使生成的语音听起来像是喃喃细语；将感受野增加到8 s，则会保留来自训练信号的短时序列信息。

加入自注意力特征提取层后模型生成信号的感知质量如图8所示，其中fake指生成器生成的假信



（a）乐器声（小提琴）　（b）乐器声（钢琴曲）

（c）动物声（犬吠）　（d）人声（男）

（e）笑声（女）　（f）自然环境声（夏夜蝉鸣）

（g）自然环境声（风声）　（h）非自然环境声（欢呼笑声）

（i）非自然环境声（装弹及子弹发射）　（j）非自然环境声（子弹连续发射及人呼喊）

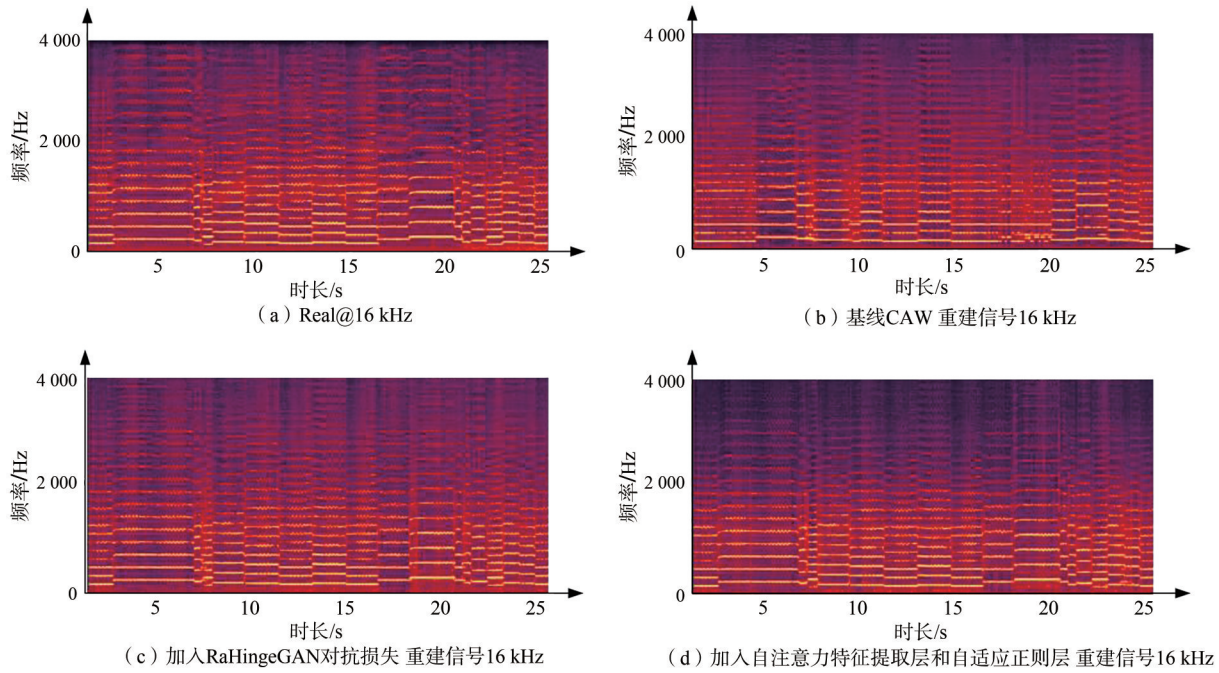
图4　10种音效类别生成信号频谱图对比

号，convnum指韵律恢复层的卷积数量，fake@our指MuSEGAN生成器生成的虚假信号，fake@CAW指CAW基线模型生成器生成的虚假信号。加入自注意力特征提取层后，对比观察图8（c）～图8（i）与图8（a）和图8（b），膨胀卷积层数保证在6～7层都可以达到介于喃喃细语和保留短时序列信息的较为适当的状态，即自注意力特征提取层在前期特征提取中起到了重要作用。

（3）加入Alpha Dropout后模型的训练收敛对比

模型训练在不同时长下的变化，如图9所示。其中D（Real）和D（fake）为判别器判别该样本为真和假的损失，D（loss）和G（loss）分别为对抗损失函数的判别器损失和生成器损失。

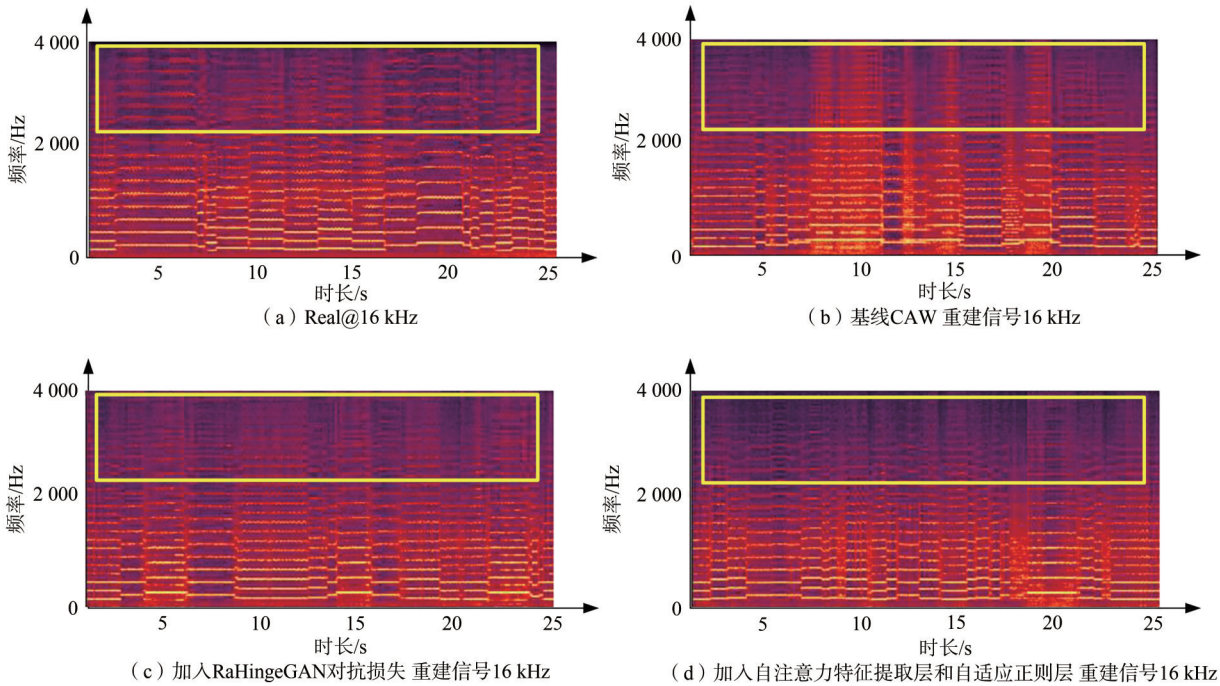
从图9可以看出，在未优化对抗损失函数的情况下，基线CAW的训练损失在8 000 s后趋于稳定，而加入Transformer编码层之后，判别器损失



（a）Real@16 kHz　（b）基线CAW 重建信号16 kHz

（c）加入RaHingeGAN对抗损失 重建信号16 kHz　（d）加入自注意力特征提取层和自适应正则层 重建信号16 kHz

图5　不同模块下重建信号频谱图对比



（a）Real@16 kHz　（b）基线CAW 重建信号16 kHz

（c）加入RaHingeGAN对抗损失 重建信号16 kHz　（d）加入自注意力特征提取层和自适应正则层 重建信号16 kHz

图6　不同模块下生成信号频谱图对比（单个乐器）

D（Real）在13 000 s损失下降，但训练时长有所增加。在优化对抗损失函数后，损失函数的对比方式发生变化，在loss值为1.0标准线范围内浮动。仅优化对抗损失后，模型的训练损失随不同时间变化在2 500 s趋于稳定，而加入Alpha Dropout之后在2 000 s趋于稳定，但后期仍有小幅度波动。在引入Alpha Dropout后，模型在训练初期可能会表现出不稳定性。Dropout的随机性使模型在每个训练步骤中都会随机丢弃一部分神经元的激活值，这可能导致训练过程中的梯度更新波动较大，进而使模型在收敛初期难以找到稳定的优化路径。虽然Alpha Dropout可以缓解过拟合，但如果选择不当，过高的Dropout率也可能导致模型欠拟合。

（4）加入风格迁移模块后再处理的音频信号频

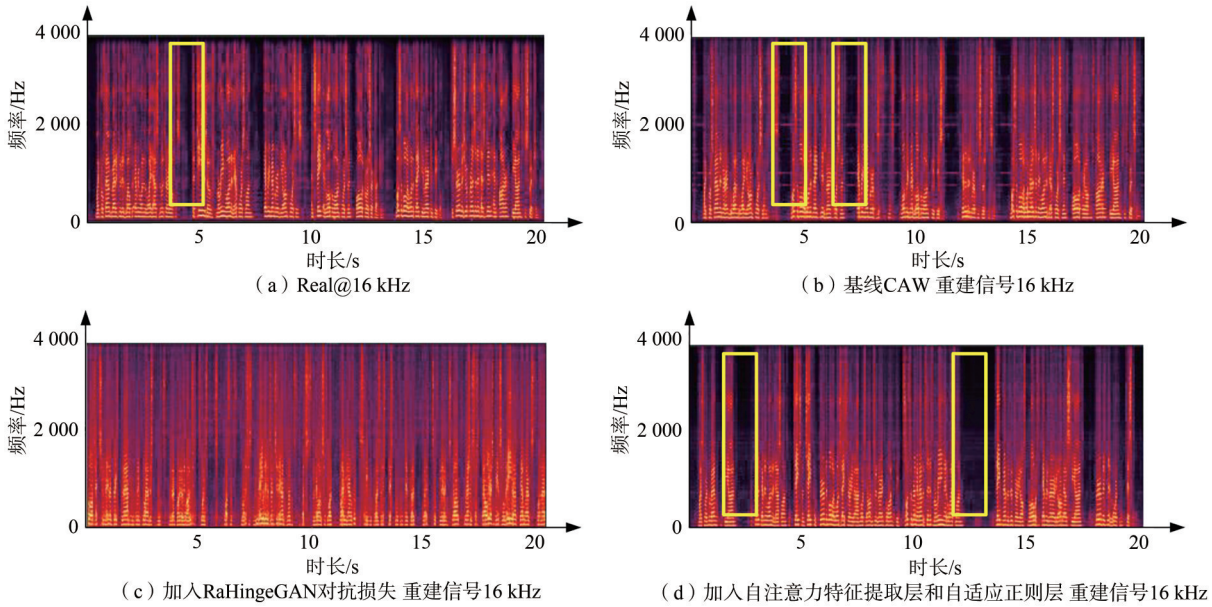


图7 不同模块下生成信号频谱图对比（语音信号）

谱对比

加入风格迁移模块后，不同预处理方法新风格信号频谱图对比如图10所示。在图10中，从左至右，分别是内容音频、风格音频以及通过两种不同预处理方式（STFT 预处理和 D^{12}-MFCC 预处理）得到的生成音频频谱图。具体来说，图10（a）展示了将小提琴音效作为内容音频，并将男性演讲的语音作为风格音频，对内容音频进行风格迁移后的结果。图10（b）则是将男性演讲的语音作为内容音频，将小提琴音效作为风格音频，进行风格迁移的实验结果。

从图10可以看出，模型仅使用STFT算法提取内容音频和风格音频的语音特征时，特征混合混乱，音频信息丢失严重。D^{12}-MGFCC 算法对音频动态特征进行补充，使融合到新风格的音频信号更为完整、平滑。

2.2.3 听觉研究测试实验

试听参与者对“声音可信度”与“声音变化度”的评分分布情况如图11所示。图11中，每个图表由5个独立的小提琴图组成，分别对应不同的条件或技术：Real、CAW、AudioLM、MuSEGAN，以及MuSEGAN-ST。小提琴图利用密度估计来描绘数据分布，其宽度反映了评分的聚集程度，更宽的部分指示着该得分区间有更多的数据点。图中的白点标记为中位数，黑线则表示数据的四分位数范围。

从声音可信度看，MuSEGAN在5类音效中评分均优于CAW方法，尤其在自然环境音中，其分布更贴近真实样本，反映出良好的音频自然度建模能力。AudioLM在人声与乐器声中取得了相对更高的可信度评分，体现出其语义建模机制在语言类与结构化音频生成中的优势，但在环境音效上表现不如MuSEGAN；从声音变化度看，MuSEGAN-TS在多个类别中表现出更广泛的变化度分布，特别是在自然环境声音上，其分布宽度与集中程度优于AudioLM与CAW，凸显了风格迁移模块在提升音频多样性方面的有效性。MuSEGAN在乐器与动物类声音中的评分也优于AudioLM，尤其是在动物声中，MuSEGAN-TS的表现明显更佳，进一步验证了其在非语言类声音生成中的建模能力。

综上所述，MuSEGAN及其风格迁移扩展版本MuSEGAN-TS在音频自然度与多样性两个维度上均展现出较优性能，尤其在环境音与非语言类声音生成任务中表现较好，验证了所提出模型在多类型音频生成任务中的广泛适用性与高主观质量。

由于评分数据不满足正态分布假设，因此本文采用非参数贝叶斯因子（Bayes factor，BF）分析来比较系统在听力研究中的表现。相较于传统的频率统计检验，BF分析不仅可以评估系统间的差异，还可以在数据支持时提供“系统间无显著差异”的证据支持[38]。具体实现中，在JASP软件中采用贝叶斯T检验，使用实验贝叶斯因子 BF_{10} 来衡量对比

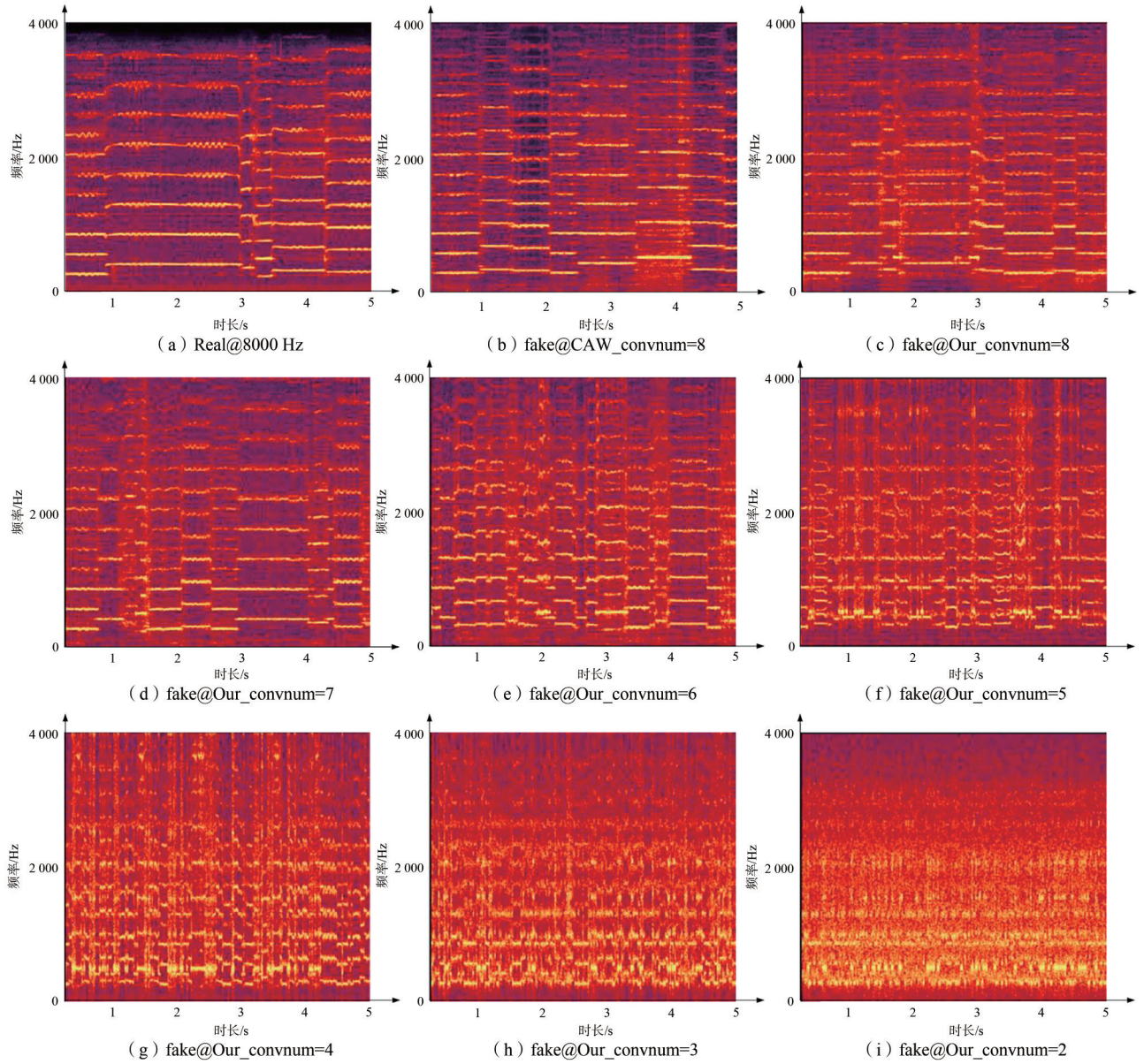


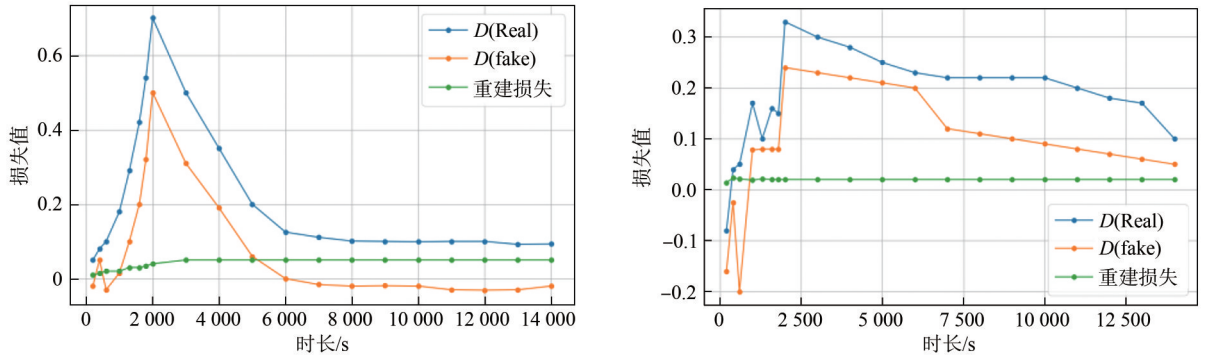
图8　生成信号的感知质量

组之间的差异强度，即差异（备选假设）与不存在差异（零假设）[39]。

假设Real在可信度和变化度方面都优于任何其他系统，BF分析发现了这方面的极端证据（BF_{10}>100）；关于可信度，假设MuSEGAN在乐器声、人声、自然环境声方面的可信度稍高于MuSEGAN-ST，对此BF分析分别提供了轶事性证据（BF_{10}=0.018、0.323、0.187）。假设MuSEGAN在动物声方面的可信度高于MuSEGAN-ST，BF分析发现了这方面的极端强有力证据（BF_{10}>100）。假设MuSEGAN和CAW在人声的可信度上相似，但BFA否定了这一点（BF_{10}>100）。数据表明，MuSEGAN在人声方面的可信度高于CAW。假设MuSEGAN在非自然环境声方面的可信度低于MuSEGAN-ST，BF分析发现了这方面的强有力证据（BF_{10}=28.446）；关于变化度，假设MuSEGAN和MuSEGAN-ST的值不相似，BF分析确认了这一点（BF_{10}=10.609）。数据表明，MuSEGAN、MuSEGAN-ST的变化性值均高于CAW（BF_{10}>100）。

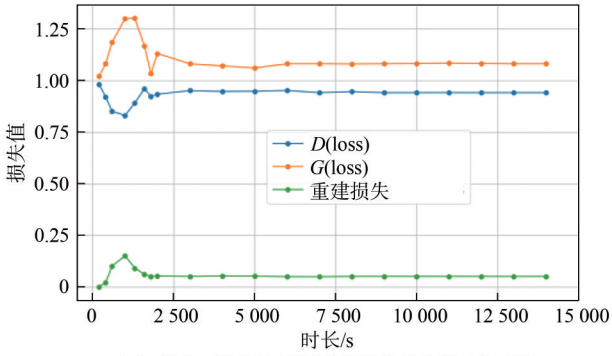
2.2.4　训练效率分析

在训练效率方面，尽管AudioLM在音频生成质量上表现优异，但其模型规模庞大，训练依赖极大规模的数据与算力资源。根据其官方实现，AudioLM使用约60 000 h的Libri-Light数据，在16个

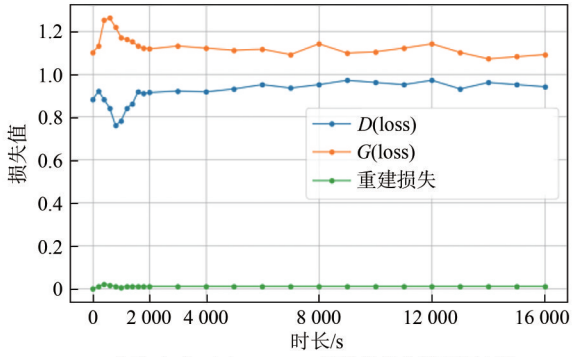


（a）基线CAW的训练损失随时间变化

（b）加入Transformer后训练损失随时间变化

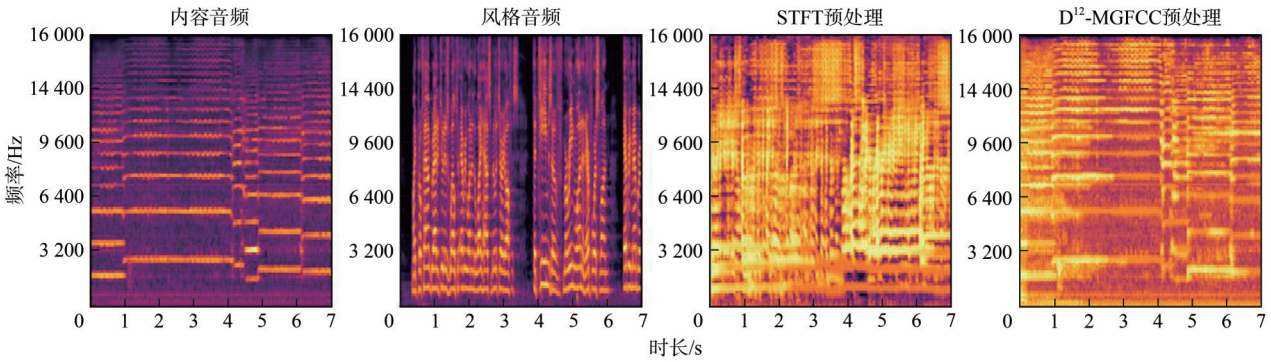


（c）优化对抗损失后模型的训练损失随时间变化

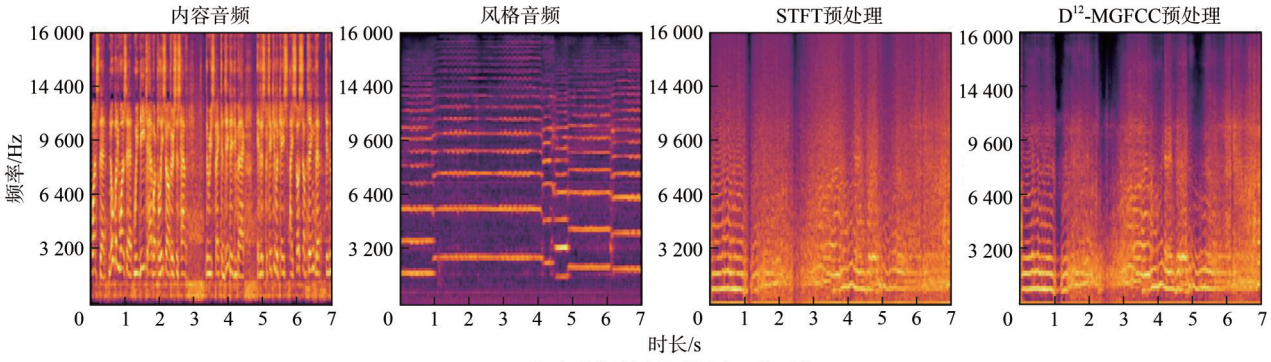


（d）加入Alpha Dropout后训练损失随时间变化

图9　模型训练在不同时长下的变化



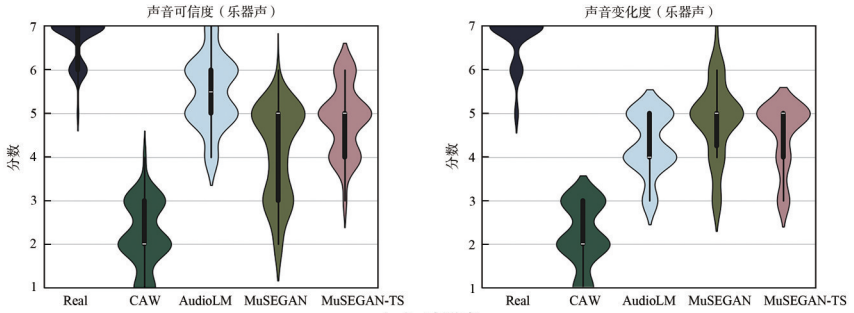
（a）语音到小提琴乐器风格迁移



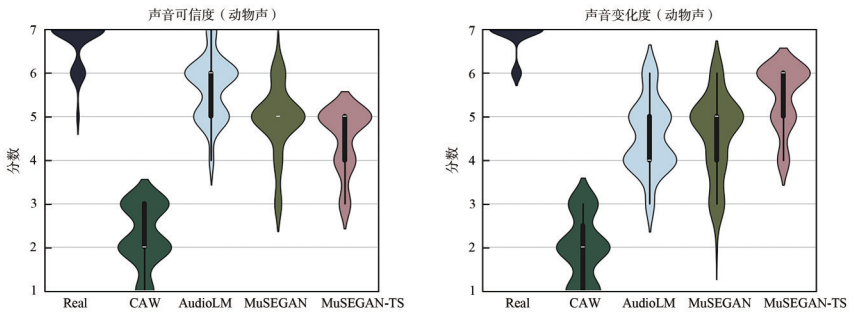
（b）小提琴乐器到语音风格迁移

图10　不同预处理方法新风格信号频谱图对比

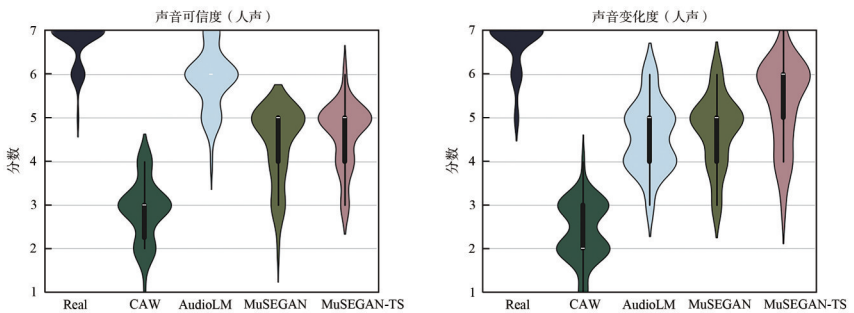
TPUv4上进行训练，每阶段需约100万步，训练和部署成本极高。相比之下，本文提出的MuSEGAN模型在计算资源和训练效率方面更具实际可行性。模型基于单卡A40（48 GB显存）环境。在无须预



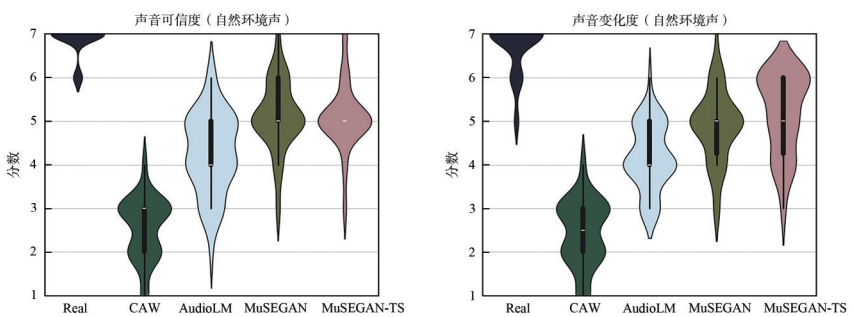
（a）乐器声



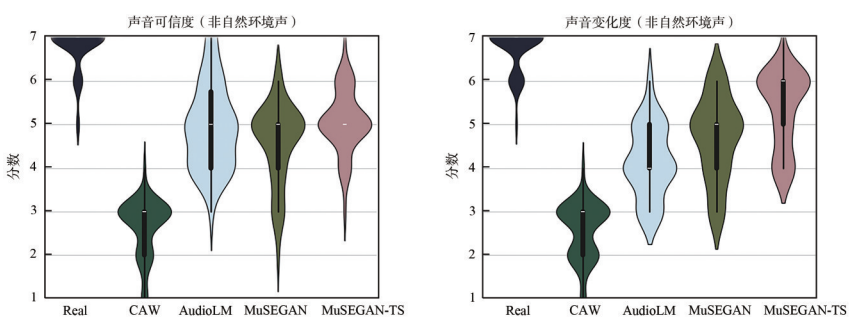
（b）动物声



（c）人声



（d）自然环境声



（e）非自然环境声

图 11　声音可信度和变化度的用户评分数据分布情况

表 1 训练效率分析

模型名称	训练数据规模	计算资源要求	样本训练耗时	是否依赖预训练模型	模型复杂度
CAW(baseline)	单条音频（音乐、语音等）	单卡 RTX2080 GPU	约 10 h/25 s 样本	否	中等
AudioLM	约 60 000 h（Libri-Light）	16×TPUv4，训练百万步级别	每阶段约需数天	是	较高
MuSEGAN(ours)	单条音频（自建数据）	单卡 A40（48 GB 显存）	约 9 h/20 s 样本	否	中等

训练语义模型的前提下，训练单条 20 s 音频样本仅需约 9 h，整体训练效率显著优于 AudioLM。此外，CAW 方法在 RTX2080 单卡上训练一条 25 s 音频样本约需 10 h。考虑到 A40 相较于 RTX2080 具有更高的计算性能，MuSEGAN 在单位算力下展现出更优的训练效率。模型 CAW(baseline)、AudioLM 和 MuSEGAN(ours)的训练效率分析见表 1。

综合来看，相较于 AudioLM 的高资源依赖，MuSEGAN 体现出更强的轻量化、实用性和部署灵活性，在资源受限环境下仍保持了较强的变化度生成能力，展现出良好的扩展潜力。

3 结束语

本文探讨了通过深度神经网络从单一音频示例中学习内部统计信息的方法，提出了一种基于多频带注意力的 GAN 模型。通过使用多频带特征提取和 Transformer 注意力机制，该方法有效提升了音效的谐波表达能力和生成稳定性；同时引入风格迁移模块，实现了音效风格的可控生成。实验结果表明，该方法能够在降低训练样本需求的同时，提高音效的逼真度和多样性。

尽管本研究取得了一定进展，但仍存在优化空间。未来的研究将进一步改进模型结构，从提高生成质量、表现力以及模型轻量化 3 个关键方面进一步提升计算效率，为用户带来更加丰富和真实的听觉体验，推动音效生成技术在各个领域的应用。

参考文献：

[1] NATSIOU A, O'LEARY S. Audio representations for deep learning in sound synthesis: a review[C]//Proceedings of the 2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA). Piscataway: IEEE Press, 2021: 1-8.

[2] 黄峻, 林飞, 杨静, 等. 生成式 AI 的大模型提示工程: 方法、现状与展望[J]. 智能科学与技术学报, 2024, 6(2): 115-133.
HUANG J, LIN F, YANG J, et al. From prompt engineering to generative artificial intelligence for large models: the state of the art and perspective[J]. Chinese Journal of Intelligent Science and Technology, 2024, 6(2): 115-133.

[3] 任维俊 . AI 音频生成在广播电视领域的应用研究[J]. 电声技术, 2024, 48(10): 110-112.
REN W J. Research on the application of AI audio generation in the field of radio and television[J]. Audio Engineering, 2024, 48(10): 110-112.

[4] 林义超, 张泽 . 非渲染技术语境下沉浸式音频空间构建方案的探索与思考[J]. 现代电影技术, 2024(12): 42-49.
LIN Y C, ZHANG Z. An exploration and reflection on immersive audio space configuration schemes in non-rendering technology context[J]. Advanced Motion Picture Technology, 2024(12): 42-49.

[5] ZHAO Y J, XIA X J, TOGNERI R. Applications of deep learning to audio generation[J]. IEEE Circuits and Systems Magazine, 2019, 19(4): 19-38.

[6] 倪清桦, 鲁越, 林飞, 等 . 平行音乐: 大模型时代的人机混合音乐创演[J]. 智能科学与技术学报, 2024, 6(2): 150-163.
NI Q H, LU Y, LIN F, et al. Parallel music: human-machine hybrid music creation and performance in the era of large models[J]. Chinese Journal of Intelligent Science and Technology, 2024, 6(2): 150-163.

[7] 王健宗, 张旭龙, 姜桂林, 等 . 基于分层联邦框架的音频模型生成技术研究[J]. 智能系统学报, 2024, 19(5): 1331-1339.
WANG J Z, ZHANG X L, JIANG G L, et al. Research on audio model generation technology based on a hierarchical federated framework[J]. CAAI Transactions on Intelligent Systems, 2024, 19(5): 1331-1339.

[8] 靳聪, 王洁, 郭子淳, 等 . 融合音画同步的唇形合成研究[J]. 智能科学与技术学报, 2023, 5(3): 397-405.
JIN C, WANG J, GUO Z C, et al. Lipsynthesis incorporating audio-visual synchronisation[J]. Chinese Journal of Intelligent Science and Technology, 2023, 5(3): 397-405.

[9] GOEL K, GU A, DONAHUE C, et al. It's raw! audio generation with state-space models[C]//Proceedings of the 39th International Conference on Machine Learning. New York: PMLR, 2022: 7616-7633.

[10] GU A, GOEL K, RÉ C. Efficiently modeling long sequences with structured state spaces[J]. arXiv preprint, 2021, arXiv: 2111.00396.

[11] YANG D C, YU J W, WANG H L, et al. Diffsound: discrete diffusion model for text-to-sound generation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 1720-1733.

[12] ZHOU Y P, WANG Z W, FANG C, et al. Visual to sound: generating natural sound for videos in the wild[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 3550-3558.

[13] 谢志峰, 孙络祎, 孙郁洲, 等 . 时序对齐视觉特征映射的音效生成方法[J]. 计算机辅助设计与图形学学报, 2022, 34(10): 1506-1514.
XIE Z F, SUN L Y, SUN Y Z, et al. Sound generation method with timing-aligned visual feature mapping[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(10): 1506-1514.

[14] XIE Z Y, LI B H, XU X N, et al. Enhancing audio generation diversity with visual information[C]//Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2024: 866-870.

[15] LIU X B, IQBAL T, ZHAO J Z, et al. Conditional sound generation using neural discrete time-frequency representation learning[C]//Proceedings of the 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP). Piscataway: IEEE Press, 2021: 1-6.

[16] KREUK F, SYNNAEVE G, POLYAK A, et al. AudioGen: textually guided audio generation[J]. arXiv preprint, 2022, arXiv: 2209.15352.
[17] SHEFFER R, ADI Y. I hear your true colors: image guided audio generation[C]//Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2023: 1-5.
[18] BORSOS Z, MARINIER R, VINCENT D, et al. AudioLM: a language modeling approach to audio generation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 2523-2533.
[19] GUO Z F, MAO J G, TAO R, et al. Audio generation with multiple conditional diffusion model[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(16): 18153-18161.
[20] LIU H H, CHEN Z H, YUAN Y, et al. AudioLDM: text-to-audio generation with latent diffusion models[J]. arXiv preprint, 2023: arXiv: 2301.12503.
[21] HUANG R J, HUANG J W, YANG D C, et al. Make-an-audio: text-to-audio generation with prompt-enhanced diffusion models[C]//Proceedings of the 40th International Conference on Machine Learning. New York: PMLR, 2023: 13916-13932.
[22] AGOSTINELLI A, DENK T I, BORSOS Z, et al. MusicLM: generating music from text[J]. arXiv preprint, 2023: arXiv: 2301.11325.
[23] LIU S G, LI S J, CHENG H N. Towards an end-to-end visual-to-raw-audio generation with GAN[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(3): 1299-1312.
[24] GHOSE S, PREVOST J J. FoleyGAN: visually guided generative adversarial network-based synchronous sound generation in silent videos[J]. IEEE Transactions on Multimedia, 2022, 25: 4508-4519.
[25] BAOUEB T, LIU H C, FONTAINE M, et al. SpecDiff-GAN: a spectrally-shaped noise diffusion GAN for speech and music synthesis [C]//Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2024: 986-990.
[26] GRESHLER G, SHAHAM T R, MICHAELI T. Catch-A-waveform: learning to generate audio from a single short example[C]//Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021). New York: Curran Associates Inc., 2021: 1-13.
[27] WU S L, YANG Y H. MuseMorphose: full-song and fine-grained piano music style transfer with one transformer VAE[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 1953-1967.
[28] KOO J, MARTÍNEZ-RAMÍREZ M A, LIAO W H, et al. Music mixing style transfer: a contrastive learning approach to disentangle audio effects[C]//Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2023: 1-5.
[29] ZHANG Y, HUANG R J, LI R Q, et al. StyleSinger: style transfer for out-of-domain singing voice synthesis[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(17): 19597-19605.
[30] WU Y X, HE Y F, LIU X L, et al. Transplayer: timbre style transfer with flexible timbre control[C]//Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2023: 1-5.
[31] SHAHAM T R, DEKEL T, MICHAELI T. SinGAN: learning a generative model from a single natural image[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 4570-4580.
[32] VASWANI A , SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2017: 6000-6010.
[33] 李牧, 杨宇恒, 柯熙政. 基于混合特征提取与跨模态特征预测融合的情感识别模型[J]. 计算机应用, 2024, 44(1): 86-93.
LI M, YANG Y H, KE X Z. Emotion recognition model based on hybrid-mel gama frequency cross-attention transformer modal[J]. Journal of Computer Applications, 2024, 44(1): 86-93.
[34] WEBBER J J, VALENTINI-BOTINHAO C, WILLIAMS E, et al. Autovocoder: fast waveform generation from a learned speech representation using differentiable digital signal processing[C]//Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2023: 1-5.
[35] FONT F, ROMA G, SERRA X. Freesound technical demo[C]//Proceedings of the 21st ACM International Conference on Multimedia. New York: ACM, 2013: 411-412.
[36] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[J]. arXiv preprint, 2017, arXiv: 1711.05101.
[37] BARAHONA-RÍOS A, COLLINS T. SpecSinGAN: sound effect variation synthesis using single-image GANs[J]. arXiv preprint, 2021, arXiv: 2110.07311.
[38] GRAY K L H, HAFFEY A, MIHAYLOVA H L, et al. Lack of privileged access to awareness for rewarding social scenes in autism spectrum disorder[J]. Journal of Autism and Developmental Disorders, 2018, 48(10): 3311-3318.
[39] ROUDER J N, SPECKMAN P L, SUN D C, et al. Bayesian t tests for accepting and rejecting the null hypothesis[J]. Psychonomic Bulletin & Review, 2009, 16(2): 225-237.

[作者简介]

姜林（1977– ），男，博士，湖南工商大学人工智能与先进计算学院教授，主要研究方向为智能语音处理、机器人应用。

苗向阳（1998– ），男，湖南工商大学智能工程与智能制造学院硕士生，主要研究方向为智能语音处理、音效生成。

洪紫荆（1997– ），女，湖南工商大学人工智能与先进计算学院硕士生，主要研究方向为智能语音处理、音效生成。