arXiv:2601.03443v1 [eess.AS] 6 Jan 2026

使用基于嵌入的分类器判别真实与合成的超分辨率音频样本

Mikhail Silaev[1], *Konstantinos Drossos*[2],和 *Tuomas Virtanen*[1]

[1]坦佩雷大学，芬兰坦佩雷[2]诺基亚技术公司，芬兰埃斯波

摘要

生成对抗网络（GAN）和基于扩散的模型近期在音频超分辨率（ADSR）领域取得了最先进的性能，能够根据窄带输入生成感知上令人信服的宽带音频。然而，现有评估主要依赖信号级或感知指标，合成超分辨率音频与真实宽带音频的分布究竟有多接近，仍是一个悬而未决的问题。本文通过分析真实音频与超分辨率音频在各种嵌入空间中的可分性，探讨这一问题。我们针对语音和音乐，考察中频段（$4 \to 16$kHz）和全频段（$16 \to 48$ kHz）两种上采样任务，并训练线性分类器，根据多种类型的音频嵌入区分真实样本与合成样本。与客观指标和主观听音测试的比较表明，即使生成音频具有较高的感知质量，且指标得分达到最先进水平，基于嵌入的分类器仍能实现近乎完美的区分。这一现象在不同数据集和模型中均一致，包括近期的基于扩散的方法，凸显了ADSR模型的感知质量与真实分布保真度之间持续存在的差距。代码和演示可见：https://github.com/msilaev/ADRS 。

索引词— 生成对抗网络、判别器、数据分布、可分性、特征表示、带宽扩展、音频超分辨率

1. 引言

生成对抗网络（GAN） [1]的主要目标是生成与真实数据高度相似的合成（“伪造”）样本，从而有效地从近似真实分布的分布中进行采样。评估 GAN 是否出色地实现这一目标仍是一项根本挑战，因为传统指标往往无法衡量感知质量、多样性以及对训练数据的过拟合程度。可靠的评估对于理解模型行为和指导生成质量的提升至关重要。为评估 GAN 的性能，人们提出了多种方法。例如，可以复用判别器冻结的中间层作为特征提取器，用于图像分类等有监督下游任务，从而评估 GAN 的表征质量

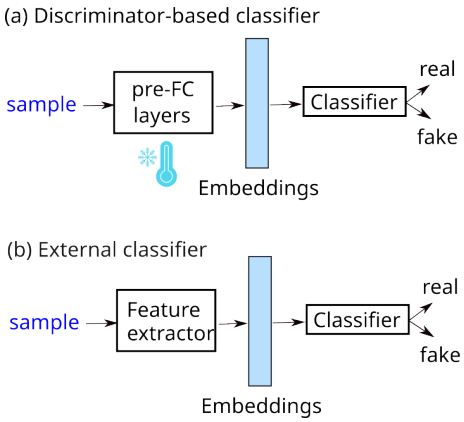


图 1：特征提取与“真实”/“伪造”分类任务。(a) 基于判别器的分类器使用其全连接层之前（pre-FC）各层在权重冻结时生成的内部判别器特征。(b) 外部分类器基于由独立网络提取的特征运行，从而能够分析两个分类器之间潜在的表示偏差。

。基于学习到的表征嵌入训练的分类器，已用于衡量语音和音频领域的隐私保护，包括自动语音识别 [3]和声音事件检测 [4]。

另一个互补方向是在嵌入空间中比较真实样本与生成样本的统计相似性，例如由预训练 Inception 网络 [5,6, 7] 所构建的嵌入空间。这类嵌入可用于计算 Fr´echet Inception 距离 [5], Fr´echet 音频距离 [8],Deep Speech 距离 [9], 以及精确率–召回率指标 [6,7, 10], 等量化指标，用于评估真实数据流形与合成数据流形之间的重叠程度和覆盖范围。

然而，这些评估策略尚未系统地应用于基于 GAN 的超分辨率任务，无论是在图像 [11] 还是音频 [12] 领域。在这项工作中，我们通过探讨一个简单但在实践中很重要的问题，扩展了对生成对抗网络和近期提出的扩散模型 [13, 14] 的评估：在以下任一项上训练的线性分类器能否

——冻结的判别器表征（图 1a）或外部嵌入（图 1b）——可靠地区分真实宽带音频样本与超分辨率音频样本?

我们聚焦于具有实际应用意义的音频超分辨率（音频超分辨率）问题，其目标是通过生成新的高频内容来增强低带宽音频信号，从而扩展其频谱范围。近年来，这一课题吸引了大量研究关注 [13, 14, 15, 16, 17, 18, 19, 12, 20, 21, 22]。针对音频超分辨率，生成对抗网络和扩散模型的评估主要采用信号级指标，例如信噪比（SNR）、对数谱距离（LSD），以及平均意见分（MOS）和主观听音测试等感知度量 [12, 18]。相比之下，针对生成对抗网络的专门评估尚未得到探索。这一空白对于全频带音频超分辨率尤为重要 [12],；全频带音频超分辨率是本文的主要研究重点，而传统客观指标和人类听众往往难以捕捉真实信号与合成信号之间细微却在感知上十分重要的差异。

本文其余部分的组织结构如下：第 2 节概述基于 GAN 的音频超分辨率及其架构。第 3 节解释“真实”/“伪造”分类器的概念。第 4 节报告评估结果，包括客观指标、针对全频带音频超分辨率的 MUSHRA（多刺激隐藏参考与锚点）听音测试 [23] 以及分类器准确率。第 5 节给出结论。

2. 音频超分辨率模型

给定音频信号 $\boldsymbol{x} = [x(n/f_s)], \quad n = 0, ..., N$ 音频超分辨率任务旨在重建上采样后的信号 $\boldsymbol{y} = [y(m/rf_s)], \quad m = 0, ..., rN$，其中 r 为上采样倍率。因此，输入信号和上采样后的信号长度分别为 N和 rN，但持续时间 N/f_s 相同。提高采样率 $f_s \rightarrow rf_s$ 会将奈奎斯特频率从 $f_s/2$扩展至 $rf_s/2$，从而允许频带 $[f_s/2, rf_s/2]$中出现新的频率分量。原始信号和上采样后的信号分别称为窄带（NB）和宽带（WB）。

用于音频超分辨率的方法发展迅速，从 AudioUNet 等有监督模型 [22] 演进到基于生成对抗网络和基于扩散的生成模型 [18, 19,20, 21, 24]。在本研究中，我们采用了 MU-GAN（多尺度 U-Net 生成对抗网络）架构，该架构最初是为 $4 \rightarrow 16$kHz 音频超分辨率提出的 [19]。我们进一步将 MU-GAN 扩展至全频带 $16 \rightarrow 48$ kHz 音频超分辨率。由于实现细节有限且没有公开可用的代码，我们从头重新实现、训练并评估了该模型。作为有监督基线，我们采用 Audio UNet 模型 [22]，分别用于 $4 \rightarrow 16$ kHz 和 $16 \rightarrow 48$ kHz 的上采样。训练得到的 MU-GAN 判别器特征被用作嵌入，用于在所有所考虑的模型中执行“真实”/“伪造”分类任务。

Both AudioUNet 和 MU-GAN 模型通过 ple- PyTorch[1] 实现，并使用 VCTK 语音数据集 [25] 对 $4 \rightarrow 16$ kHz 和 $16 \rightarrow 48$ kHz 音频超分辨率进行训练，同时使用 FMA-small 音乐数据集 [26] 对 $16 \rightarrow 48$ kHz 音频超分辨率进行训练。对于 FMA 数据集 [26]，我们采用了官方的训练集/验证集/测试集划分；对于 VCTK 数据集，则采用了训练集/测试集划分，遵循以往工作 [12, 19, 22]所使用的协议。训练在单个 NVIDIA A100 GPU 上进行。对于 VCTK 数据集，两个模型均训练了 500 个训练轮次。对于 FMA-small 数据集，AudioUNet 和 MU-GAN 分别训练了 100 和 80 个训练轮次，并采用了提前停止。学习率设置为 10^{-4}，小批量大小设置为 128。生成器和判别器分别使用 Adam 和随机梯度下降（ SGD ）进行优化。为稳定对抗训练，我们采用了定期更新策略，即生成器的更新频率高于判别器。对于已收敛的 MU-GAN 模型，判别器在 VCTK 数据集上的准确率约为 51%，在 FMA 数据集上约为 49%，这表明模型达到了训练良好的生成对抗网络所具有的理想平衡状态，即真实样本与生成样本几乎无法区分。

为便于与全频带音频超分辨率领域的先进方法进行比较，我们采用了最初用于去噪和带宽扩展的 HiFi-GAN [12], ，以及两个近期的基于扩散的模型 FlowHigh [13]和 FlashSR [14],，二者均能够将任意输入采样率的音频上采样至 48 kHz。实验中，我们使用了公开可用的非官方 HiFi-GAN 实现及其预训练权重进行推理 [24], ，并使用了为 FlowHigh[2] 和 FlashSR[3]发布的官方推理代码。

3. “真实”/“伪造”音频分类器

利用多个带标签的嵌入数据集，对能够区分真实与合成音频片段的分类器进行训练和评估。这些数据集通过使用不同的特征提取器，将真实与合成音频信号转换为定长嵌入而构建。本文考虑了多种类型的嵌入。首先，为了研究学习到的表征的质量，我们使用从 MU-GAN 判别器的固定预 FC 层中提取的特征（图 1a）。在我们的实现中，该判别器层会将包含 8192 个采样点的输入映射为一个 32 维嵌入向量。

此外，还采用了两种类型的外部嵌入（图 1b）。OpenL3 [27] 模型可从 1 秒的音频片段生成一个 512 维嵌入向量。该模型适用于 $4 \rightarrow 16$ kHz 音频超分辨率任务，但由于输入带宽有限，无法捕获 48 kHz 音频的完整频率内容，因此不能可靠地应用于 $16 \rightarrow 48$ kHz 场景。

[1]https://github.com/msilaev/ADRS[2] https://github.com/resemble-ai/flowhigh[3] https://github.com/jakeoneijk/FlashSR_推理

为解决 OpenL3 的带宽限制，我们采用对数梅尔频谱能量作为替代特征表示。我们使用 256 个梅尔频率频点、4096 点快速傅里叶变换和 256 的帧移。上限频率由奈奎斯特频率决定；对于 16 kHz 和 48 kHz 的输入采样率，该频率分别为 8 kHz 和 24 kHz。为生成与输入时长无关的定长嵌入，我们沿时间维度应用自适应平均池化。

表 1：不同模型的对数谱距离和信噪比指标

模型	VCTK				FMA	
	LSD 4→16	SNR 4→16	LSD 16→48	SNR 16→48	LSD 16→48	SNR 16→48
AudioUnet	4.5	15.4	4.2	22	9.2	24.5
MU-GAN	3.9	14.6	4.2	20.8	6.7	27.3
HiFi-GAN	–	–	2.1	17.5	–	–
FlowHigh	–	–	3	-6.8	3.6	-3
FlashSR	–	–	3.9	16	8.4	18.2

4. 评估与结果

为了评估 ADSR模型的性能和感知真实感，我们结合客观信号比较指标、信噪比（SNR）和对数谱距离（LSD），并进行 MUSHRA 听测。

4.1. 信号级性能评估

表1汇总了信噪比和对数谱距离的结果，主要用于说明未观察到异常偏低或偏高的数值。 $4 \rightarrow 16$ kHz ADSR 的结果与先前研究 [19, 22], 一致；与这些研究一样，对数谱距离采用自然对数计算。

值得注意的是，FlowHigh模型的信噪比为负值，但其对数谱距离低于AudioUNet和MU-GAN。信噪比为负并不表示感知质量下降，而是源于全局幅度缩放因子，该因子可能因样本而异。

HiFi-GAN、AudioUNet 和 MU-GAN 的评分之间未观察到显著差异。然而，如下所示，信噪比和对数谱距离指标与人类听测结果的相关性较差，而人类听测结果表明，不同模型输出之间存在明显的感知差异。

4.2. 多刺激隐藏参考与锚点测试听测

评估时，从 VCTK 测试集中随机选取了 12 段录音，确保男性说话人和女性说话人的数量相同。原始宽带录音以 48 kHz 采样，先降采样至 16 kHz，再升采样回 48 kHz，分别使用三种不同的音频超分辨率模型：AudioUnet、MU-GAN 和 HiFi-GAN。每段录音还配有两个锚定信号：一个截止频率为 3.5 kHz 的低通滤波版本，以及一个截止频率为 7 kHz 的中通滤波版本。此外，还纳入了原始窄带录音，共形成六种实验条件。听测采用 MUSHRA 听测界面进行，听者可以设置循环播放，以便专注聆听音频信号中的特定短片段。

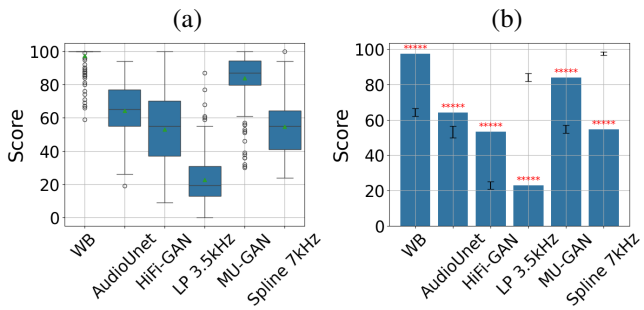


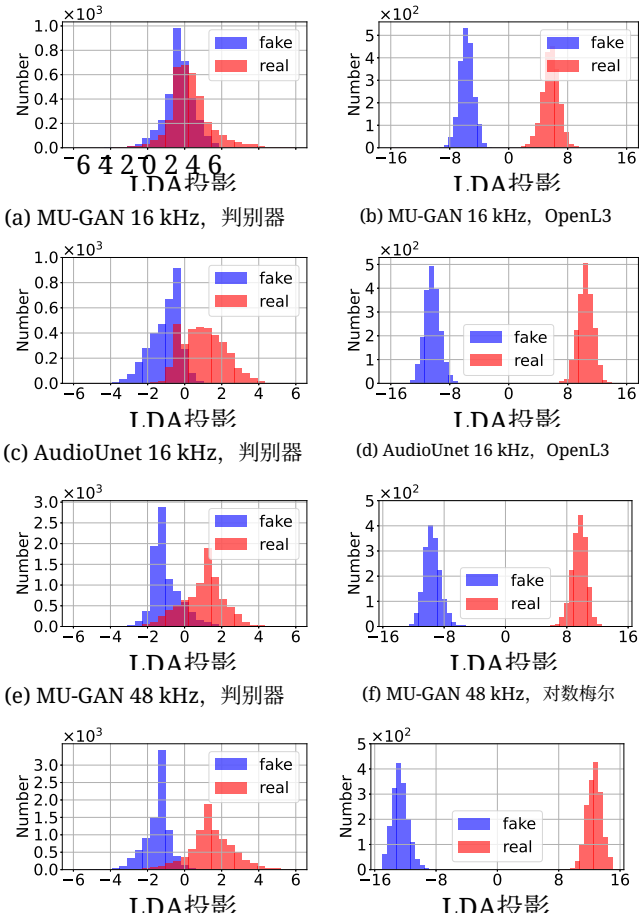
图 2：不同音频超分辨率方法的听者评分。不同条件下听测测试的 MUSHRA 评分：宽带、MU-GAN、AudioUnet、HiFi-GAN、低通 3.5 kHz、样条上采样 7 kHz。(a) 四分位距（IQR）、中位数，以及以绿色三角形标示的均值。(b) 均值及表示 95% 置信区间的误差线。

图 2 展示了按照 MUSHRA 建议 [23] 得出的听测结果。图 2a 的箱线图显示了六种条件下的 MUSHRA 评分分布：宽带、MU-GAN、AudioUnet、HiFi-GAN、低通 3.5 kHz 和样条上采样 7 kHz，并标出了中位数、四分位距、均值和离群值。图 2b 的柱状图显示了均值及其 95% 置信区间。MU-GAN 的评分最高，与宽带参考音频十分接近；HiFi-GAN 的表现最差，与 7 kHz 锚点十分相似。AudioUnet 的表现略好一些。图 2b 中不重叠的置信区间表明，听者能够可靠地区分真实音频和合成音频，尽管 MU-GAN 输出的感知质量仍与目标宽带录音相当。下一节将展示，二元分类器也能捕捉到这种差异。

4.3. “真实”/“伪造”分类器准确率

嵌入数据集由 VCTK（语音）和 FMA-small（音乐）数据集测试子集中的音频片段构建而成，这些片段未用于任何所研究模型的训练。每个嵌入数据集都经过随机打乱，并划分为训练子集（80%）和测试子集（20%）。分类前，将特征向量标准化，使其均值为零、方差为一。随后使用归一化后的训练

数据训练线性判别分析（LDA）分类器，并在测试数据上进行评估，以衡量其区分真实样本与伪造样本的能力。此外，将测试嵌入投影到 LDA 学得的判别子空间中，在该空间中类别分离程度达到最大。该投影可沿单一判别轴直观呈现类别重叠情况及可分性。



(a) MU-GAN 16 kHz，判别器　(b) MU-GAN 16 kHz，OpenL3

(c) AudioUnet 16 kHz，判别器　(d) AudioUnet 16 kHz，OpenL3

(e) MU-GAN 48 kHz，判别器　(f) MU-GAN 48 kHz，对数梅尔

(g) AudioUnet 48 kHz，判别器特征 tor　(h) AudioUnet 48 kHz，对数梅尔

图 3：使用 MU-GAN 判别器特征（左列）、OpenL3（右列 b、d）和对数梅尔（右列 f、h）嵌入计算的 VCTK 测试数据集中音频片段的 LDA 投影。(a-d) $4 \rightarrow 16$ kHz 和 (e-h)$16 \rightarrow 48$ kHz音频超分辨率。

图 3 展示了不同嵌入空间中线性判别分析的分布示例。与（a、c、e）对应的面板展示了判别器全连接层前特征嵌入的分布：$4 \rightarrow 16$ kHz MU-GAN 的分布完全重叠，$16 \rightarrow 48$ kHz MU-GAN 的分布部分重叠，而 AudioUNet 的分布几乎完全分离。一方面，这表明判别器学到了能够区分真实与合成音频片段的有效特征。另一方面，MU-GAN 与 AudioUNet 的分布重叠程度存在明显差异，这表明生成器经过训练后能够使这些分布更加相似。

MU-GAN 和 AudioUNet 之间的差异表明，生成器经过训练，旨在使这些分布更加相似。

相比之下，(b、d) 和 (f、h) 面板显示，OpenL3 和对数梅尔嵌入的线性判别分析投影实现了完全的「真实」/「伪造」类别分离，分类准确率达到 100%。请注意，(e、f) 和 (g、h) 面板分别对应于 MU-GAN 和 AudioUnet 模型；这些模型针对 $16 \rightarrow 48$ kHz 上采样进行了研究，并采用了第 4.2 节报告的 MUSHRA 测试。

表 2：基于 MU-GAN 判别器嵌入的二元「伪造」/「真实」分类器准确率（%）。各列对应于 VCTK 上 $4 \rightarrow 16$ kHz、VCTK 上 $16 \rightarrow 48$ kHz 以及 FMA-small 数据集上的 ADSR $16 \rightarrow 48$kHz。

模型	VCTK 4→16	VCTK 16→48	FMA 16→48
AudioUnet	80%	95%	78%
MU-GAN	56%	83%	70%
HiFi-GAN	–	93%	–
FlowHigh	–	85%	74%
FlashSR	–	88%	66%

这些结果在定性上适用于所有考察的模型。如表 2 所示，在 VCTK 数据集上，学习得到的 MU-GAN 判别器嵌入在 AudioUNet、FlashSR 和 FlowHigh 模型上的准确率较高（约 90%）；在 FMA 数据集上则略低（约 70%–80%）。与此同时，对数梅尔和 OpenL3 嵌入在所有考察的任务和模型上均实现了完美的分类准确率（100%）。

5. 结论

本研究揭示了传统信号指标、感知质量，以及生成式 ADSR 模型所产生的真实音频与超分辨率音频分布的可分性之间存在差距。在生成对抗网络训练稳定期间，判别器准确率收敛至约 50%，但其学习到的表征仍能在监督基准测试中区分“真实”和“伪造”音频，这表明生成对抗网络捕获了全面的数据特征。多刺激隐藏参考与锚点测试中的听者评分与目标宽带音频高度吻合。

基于外部嵌入（如 OpenL3 或对数梅尔频谱图）训练的分类器，能够几乎完美地区分真实与生成的音频片段。这种现象在不同领域、不同采样率（$4 \rightarrow 16$ kHz 和 $16 \rightarrow 48$ kHz）下均一致存在，也适用于最先进的扩散模型 [13, 14]。这些结果表明，高感知质量并不一定意味着准确的分布建模，揭示了感知真实感与表征保真度之间持续存在的差距——这是未来 ADSR 研究仍需应对的一项开放挑战。

6. 参考文献

[1] I. Goodfellow 等，“生成对抗网络”，神经信息处理系统进展，第 27 卷，2014 年。[2] A. Radford、L. Metz 和 S. Chintala，“使用深度卷积生成对抗网络进行无监督表示学习”，*arXiv:1511.06434*。[3] B. Srivastava 等，“自动语音识别中的隐私保护对抗表示学习：现实还是幻象？”，发表于 *Interspeech*，2019 年，第 3700–3704 页。[4] S. Gharib 等，“用于音频鲁棒隐私保护的对抗表示学习”，*IEEE*开放信号处理期刊，第 5 卷，第 294–302 页，2024 年。[5] M. Heusel 等，“通过两时间尺度更新规则训练的生成对抗网络收敛于局部纳什均衡”，神经信息处理系统进展，第 30 卷，2017 年。

[6] M. Lucic 等，“生成对抗网络是否生而平等？一项大规模研究”，神经信息处理系统进展，第 31 卷，2018 年。[7] M. Sajjadi 等，“通过精确率和召回率评估生成模型”，神经信息处理系统进展，第 31 卷，2018 年。

[8]K. Kilgour 等，“Frechet 音频距离：评估音乐增强算法的度量指标”，*arXiv:1812.08466*。

[9] M. Binkowski 等人，“利用对抗网络进行高保真语音合成”，*arXiv:1909.11646*。[10] T. Kynk¨a¨anniemi 等人，“用于评估生成模型的改进型精确率和召回率指标”，神经信息处理系统进展，第 32 卷，2019 年。[11] C. Ledig 等人，“利用生成对抗网络实现照片级真实感的单幅图像超分辨率”，发表于计算机视觉与模式识别会议 - *IEEE* 计算机视觉与模式识别会议，2017 年，第 4681–4690 页。[12] J. Su 等人，“带宽扩展就是你所需的一切”，发表于*ICASSP - IEEE* 国际声学、语音与信号处理会议IEEE，2021 年，第 696–700 页。[13] J. Im 和 J. Nam，“Flashsr：通过扩散蒸馏实现一步式通用音频超分辨率”，发表于 *ICASSP - IEEE* 国际声学、语音与信号处理会议IEEE，2025 年，第 1–5 页。[14] J. Yun、S. Kim 和 S. Lee，“Flowhigh：通过单步流匹配实现高效、高质量音频超分辨率”，发表于 *ICASSP - IEEE* 国际声学、语音与信号处理会议 IEEE，2025 年，第 1–5 页。[15] D. Gupta 和 H. Shekhawat，“利用 H∞优化、深度神经网络和语音产生模型进行人工带宽扩展”，发表于 *SPCOM - IEEE* 国际信号处理与通信会议，2022 年，第 1–5 页。

[16] H. Liu 等人，“Audiosr：大规模通用音频超分辨率”，发表于*ICASSP - IEEE* 国际声学、语音与信号处理会议，2024 年，第 1076–1080 页。[17] X. Li 等人，“面向语音识别的语音音频超分辨率”，发表于 *Interspeech*，2019 年，第 3416–3420 页。[18] E. Moliner 和 V. V¨alim¨aki，“Behm-gan：利用生成对抗网络扩展历史音乐的带宽”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 31 卷，第 943–956 页，2022 年。[19]S. Kim 和 V. Sathe，“通过生成对抗网络扩展原始音频的带宽”，*arXiv:1903.09027*。[20] J. Kong、J. Kim 和 J. Bae，“HiFi-GAN：用于高效高保真语音合成的生成对抗网络”，神经信息处理系统进展，第 33 卷，第 17022–17033 页，2020 年。[21] Y. Li 等人，“实时语音频率带宽扩展”，发表于 *ICASSP - IEEE* 国际声学、语音与信号处理会议，2021 年，第 691–695 页。[22] V. Kuleshov、S. Z. Enam 和 S. Ermon，“利用神经网络进行音频超分辨率”，*arXiv:1708.00853*。[23] ITU-R，“音频系统中等质量水平的主观评估方法”，技术报告 BS.1534-3，国际电信联盟，瑞士日内瓦，2015 年 10 月。[24]B. Spell，“Hifi-gan-bwe：高保真带宽扩展”，https://pypi.org/project/ hifi-gan-bwe/，2025 年。[25] J. Yamagishi 等人，“CSTR VCTK 语料库：适用于 CSTR 声音克隆工具包的英语多说话人语料库（版本 0.92）”，爱丁堡大学 *CSTR*，第 271–350 页，2019 年。[26] M. Defferrard 等人，“FMA：音乐分析数据集”，发表于 *ISMIR* - 国际音乐信息检索学会会议，2017 年，第 316–323 页。[27] A. Cramer 等人，“看、听、学更多：深度音频嵌入的设计选择”，发表于 *ICASSP - IEEE* 国际声学、语音与信号处理会议，2019 年，第 3852–3856 页。