arXiv:2305.03328v1 [eess.AS] 5 May 2023

采用GMM估计器进行异常声音检测的时间加权频域音频表示

Jian Guan[1*]、*Youde Liu*[2]、*Qiaoxi Zhu*[3]、*Tieran Zheng*[2]、*Jiqing Han*[2]、*Wenwu Wang*[4]

[1]哈尔滨工程大学计算机科学与技术学院智能信号处理组，中国[2]哈尔滨工业大学计算机科学与技术学院，中国[3]悉尼科技大学音频、声学与振动中心，澳大利亚[4]萨里大学视觉、语音与信号处理中心，英国

摘要

尽管深度学习是无监督异常声音检测中的主流方法，但以统计音频频率表示为输入的高斯混合模型（GMM）也能取得相当的效果，同时显著降低模型复杂度并减少参数数量。现有的统计频率表示（例如对数梅尔频谱图在时间维度上的平均值或最大值）并不总能适用于不同机器。本文提出将时间加权频域表示（TWFR）与GMM方法结合（TWFR-GMM），用于异常声音检测。TWFR是一种广义统计频域表示，可通过对时域进行全局加权排序池化，适应不同机器类型。即使在域偏移条件下，这也能使GMM估计器识别异常，并通过基于马氏距离的度量进行可视化。DCASE 2022挑战赛任务2数据集上的实验表明，我们的方法比近期的深度学习方法具有更好的检测性能。TWFR-GMM是我们提交方案的核心，该方案在DCASE 2022挑战赛任务2中获得了第3名。

索引词— 异常声音检测、音频表示、高斯混合模型、马哈拉诺比斯距离

1. 引言

异常声音检测（ASD）通过目标机器发出的声音判断其是否存在异常。在无监督异常声音检测中，由于异常声音可能十分罕见且种类繁多，实际中难以采集，因此检测器通常仅使用正常声音进行训练。用于机器状态监测的无监督异常声音检测一直是声学场景与事件检测和分类（DCASE）挑战赛中的任务2，自 2020 [1]年起。与 DCASE 2020 的任务2不同，DCASE 2021 和 DCASE 2022 的任务2侧重于域偏移问题，即训练数据与测试数据的声学特征有所不同 [2, 3]。

深度学习方法是异常声音检测的主流方法。基于自编码器（AE）的方法是 DCASE 任务2中无监督异常声音检测的基线方法 [1–3]。该方法使用 AE 模型重建正常声音的对数梅尔频谱图，以学习正常声音的分布，并利用重建误差计算异常分数。随后，通过引入设备信息来学习更好的正常声音特征表示，从而改进了该方法 [4, 5]。例如，MobileNetV2 [6] 使用自监督的机器身份（ID）分类器学习更好的正常声音表示，并将相应 ID 的负对数似然（NLL）作为异常分数 [4]。另一种方法 ResNet-GMM [7] 则不根据 ID 分类的输出概率计算异常分数，而是在由 ResNet 构建的自监督 ID 分类器之后，使用高斯混合模型（GMM）作为异常估计器 [8]。这里，GMM 将有限个高斯分布的混合用于拟合正常声音的分布，从而对正常声音的音频特征进行聚类。

最初有人尝试将 GMM估计器 应用于异常声音检测中的非深度学习方法，输入为沿时间维度对 log-Mel 频谱图进行平均池化或最大池化后的结果 [7], ，即均值高斯混合模型和最大值高斯混合模型。它们的输入是频域音频表示，采用时间维度上的简单统计特征（均值或最大值），而非从神经网络（如 ResNet）中提取。这些模型通常比深度学习模型参数更少、复杂度更低 [4, 9, 10], ，但检测性能相当。不过，简单的平均池化或最大池化表示并非总能在不同机器类型上取得良好效果。例如，在 ToyCar、Fan 和 Gearbox 上，均值高斯混合模型的表现优于最大值高斯混合模型

*通讯作者 本研究部分得到黑龙江省自然科学基金资助，项目编号：YQ2020F010；同时获得英国文化协会Newton机构合作奖资助，项目编号：623805725。为实现开放获取，作者已将由此产生的任何作者接受稿件版本采用知识共享署名（CC BY）许可。

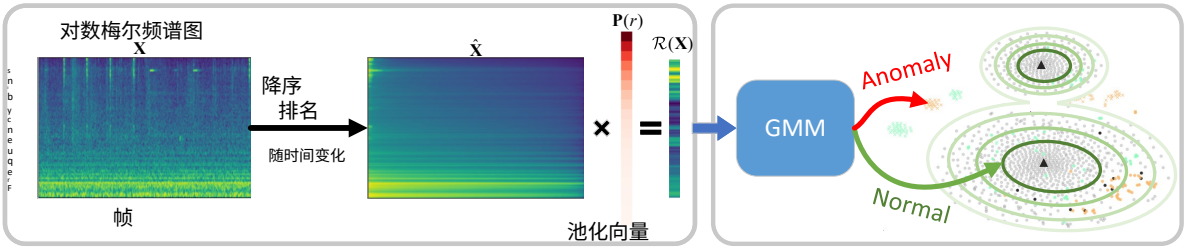


图 1：我们提出的基于高斯混合模型的时间加权频域表示（TWFR-GMM）方法。我们在每个频带上对各时间帧的对数梅尔频谱图进行降序排名，并将提取的 TWFR 与全局加权排序池化$\mathbf{P}(r)$得到的加权向量相乘，从而形成降维后的音频表示 $\mathcal{R}(\mathbf{X})$。

，而在 DCASE 2022 Task2 数据集上进行测试时，最大值高斯混合模型在 ToyTrain 和 Valve 上的表现则远优于均值高斯混合模型。

由于不同机器的音频时频分布各不相同，因此有必要开发一种能够适应不同机器的音频信号时频表示自适应统计模型。此外，在对音频信号的平稳部分和非平稳部分进行适当权衡的同时，还希望保留相较于深度学习方法模型复杂度较低的优势。

为此，本文提出一种用于无监督异常声音检测的基于高斯混合模型的时间加权频域表示（TWFR-GMM）方法。TWFR 是通过对对数梅尔频谱图沿时间维度进行全局加权排序池化（GWRP） [11] 得到的，该方法考虑整个时间范围，并为声能较高的时间帧赋予更大的权重。为考察所提方法的有效性，我们设计了一种基于马哈拉诺比斯距离的度量，用于对 TWFR-GMM 进行 t 分布随机邻域嵌入（t-SNE） [12] 可视化。t-SNE 可视化展示了高斯混合模型对训练和测试声音的聚类效果。这验证了直接从正常机器声音中提取的 TWFR 能够很好地区分异常声音，即使在域偏移条件下也是如此。在 DCASE 2022挑战赛任务2数据集上的实验表明，我们的方法比当前最先进的深度学习方法具有更好的检测性能。TWFR-GMM 是我们提交方案的核心，该方案在 DCASE 2022挑战赛任务2中排名前三。为便于复现我们的工作，我们在以下地址公开了源代码：
https://github.com/liuyoude/TWFR-GMM。

2. 提出的方法

所提出的用于无监督异常声音检测的TWFR-GMM方法如图1所示，音频表示模块、时间加权频域表示（TWFR）的详细信息见第2.1节，GMM估计器的详细信息见第2.2节。

2.1. 时域加权频域表示

现有的统计频域表示仅仅是对数梅尔频谱图在时间上的平均值或最大值，并不总是适用于不同的机器。具体而言，最大值高斯混合模型只考虑声能最大的时间帧，而忽略其他所有时间帧。因此，正常机器声音的一项关键特征——稳态特征——未被计入。相反，均值高斯混合模型对整个时间段内的每个时间帧一视同仁，能够很好地捕捉稳态特征，但由于平均操作，可能无法捕捉短时信号的瞬态特征。然而，正常声音与异常声音的瞬态特征可能存在显著差异。

所提出的 TWFR-GMM 通过采用预先选择的简单时域加权方式，使其能够适应每种机器类型，从而在模型复杂度较低方面保持了相较于深度学习方法的优势。该加权方式用于权衡平稳和非平稳音频信号，以形成音频信号的简单统计频域表示。具体而言，我们在时间维度上对数梅尔频谱图应用全局加权排序池化（GWRP） [11]，从而为声能较高的时间帧赋予更大的权重。考虑一个包含 M 个梅尔频率频带和 N 个时间帧的音频信号的对数梅尔频谱图$\mathbf{X} \in \mathbb{R}^{M\times N}$。我们按照降序排名对其第 i 行向量$\mathbf{X}_i \in \mathbb{R}^{1\times N}$（$i = 1, 2, \cdots, M$）（即，跨越 $\hat{}$时间帧的第 i 个频率频带）中的元素进行排序，如$\mathbf{X}_i$所示。它按照能量递减的顺序，重新排列每个频率频带在各时间帧上的值，并忽略时间序列。所得序列$\hat{}\mathbf{X} = [\hat{}\mathbf{X}_1, \hat{}\mathbf{X}_2, \ldots, \hat{}\mathbf{X}_M]$由$\mathbf{X}$按时间降序导出。由$\mathbf{X}$ 的 GWRP 提取的 TWFR 为

$$\mathcal{R}(\mathbf{X}) = \hat{\mathbf{X}}\mathbf{P}(r), \tag{1}$$

其中 $\mathcal{R}(\mathbf{X}) \in \mathbb{R}^M$ 以及池化向量

$$\mathbf{P}(r) = [\frac{r^0}{z(r)}, \frac{r^1}{z(r)}, \ldots, \frac{r^{N-1}}{z(r)}]^{\mathrm{T}}, \tag{2}$$

表1：DCASE 2022 Task2 开发数据集上以曲线下面积（AUC）（%）和部分曲线下面积（pAUC）（%）为指标的性能比较。平均值：所有机器类型的 AUC 和 pAUC 数值的平均值。

方法	玩具车		玩具列车		Fan		变速箱		轴承		滑块		阀门		平均值	
	曲线下面积	部分曲线下面积	曲线下面积	部分曲线下面积	曲线下面积	部分曲线下面积	曲线下面积	部分曲线下面积	曲线下面积	部分曲线下面积	曲线下面积	部分曲线下面积	曲线下面积	部分曲线下面积	曲线下面积	部分曲线下面积
深度学习 方法：																
MobileNetV2 [3]	55.54	52.27	51.57	51.51	59.48	56.89	62.70	56.03	60.25	57.14	51.69	54.67	62.14	62.41	57.62	55.85
自编码器 [3]	62.61	52.74	49.83	50.48	62.89	57.52	65.78	58.49	56.40	51.98	62.81	55.78	50.73	50.36	58.72	53.91
STgram-MFN [5]	48.62	49.91	52.10	49.62	64.42	60.95	75.97	64.21	**78.58**	**64.28**	79.46	63.22	70.80	63.32	67.14	59.36
非深度学习方法：																
均值高斯混合模型 [7]	79.33	58.93	58.74	52.74	**71.44**	**62.13**	79.62	65.37	69.41	54.74	78.57	63.62	52.39	50.54	69.93	58.30
最大值高斯混合模型 [7]	60.21	51.19	67.34	55.28	62.51	52.71	71.40	54.68	64.15	52.29	82.65	66.49	91.87	69.86	71.45	57.50
TWFR-GMM	80.65	58.12	67.81	59.07	**71.44**	**62.13**	80.23	64.69	69.41	54.74	87.96	73.33	**92.61**	**70.23**	78.59	63.19
SMOTE-TWFR 高斯混合模型	**84.70**	**60.55**	**68.02**	**59.69**	71.44	62.13	**82.22**	**65.85**	70.80	54.22	**89.91**	**76.74**	**92.61**	**70.23**	**79.96**	**64.20**

使用 $z(r) = \sum_{n=1}^{N} r^{n-1}$ 进行归一化。上标 T 表示转置操作。

平均池化和最大池化是 GWRP 的两种特殊情况。当 $r = 0$时，GWRP 退化为最大池化；当 $r = 1$时，GWRP 则变为平均池化。我们的方法会在训练阶段为每种机器类型从 0 到 1 中找到一个合适的 r，以获得最佳检测性能。因此，TWFR 可以适配不同的机器类型，从而为异常声音检测实现更稳健的音频特征表示。

2.2. 高斯混合模型（GMM）估计器

TWFR-GMM：GMM 用于将正常声音的分布拟合为有限个高斯分布的混合。GMM 使用正常声音进行训练，并根据定义的负对数似然检测测试声音中的异常：

$$\mathcal{A}(\bar{\mathbf{X}}) = -\max_{k\in[1,K]} log\mathcal{N}(\mathcal{R}(\bar{\mathbf{X}})|\mu_k, \mathbf{\Sigma}_k), \qquad (3)$$

其中，$\mathcal{N}(\mathcal{R}(\bar{\mathbf{X}})|\mu_k, \mathbf{\Sigma}_k)$ 是训练所得 GMM 中第 k 个高斯分布，其均值向量为 $\mu_k \in \mathbb{R}^M$，协方差矩阵为 $\mathbf{\Sigma}_k \in \mathbb{R}^{M\times M}$；$\bar{\mathbf{X}}$ 是测试声音的对数梅尔频谱图。

SMOTE-TWFR 高斯混合模型：针对存在域偏移问题的 DCASE 2022挑战赛任务2，我们针对每种机器类型的每个区段分别训练高斯混合模型，其中不同区段对应不同的域偏移条件。此外，SMOTE [13] 通过对部分机器类型目标域中的样本进行过采样，来解决样本不足问题。

3. 实验与结果

3.1. 实验设置

数据集：我们在 DCASE 2022 挑战赛任务2 [3], 的开发数据集上进行了实验，该数据集包含七种机器类型（ToyCar、ToyTrain、Fan、Gearbox、Bearing、Slider 和 Valve）。每种机器类型的数据集都分为三个部分（Section 00、01 和 02），不同部分表示不同的域偏移条件。每个音频文件时长为 10 秒，采样率为 16 kHz，包含机器声音和环境噪声。

评估指标：遵循 DCASE 2022挑战赛任务2 [3], ，我们使用所有机器类型、区段和域上的受试者工作特征（ROC）曲线下面积（曲线下面积）以及部分曲线下面积（pAUC）作为性能指标。此外，分别采用参数数量和浮点运算次数（浮点运算次数）评估模型规模和复杂度。

实现：从音频中提取对数梅尔频谱图，窗口大小为 1024 个采样点，重叠率为 50%，梅尔滤波器组数为 128。高斯混合模型和 t-SNE 可视化通过 *scikit-learn* 库 [14],实现，高斯混合模型的混合成分数根据机器类型设为 1 或 2。SMOTE 通过*imbalanced-learn*库 [15] 实现，并应用于部分机器类型。权重参数 r 在公式（1）中，对于 Toycar、ToyTrain、Fan、Gearbox、Bearing Slider 和 Valve，分别为 0.99、0.81、1.00、0.99、1.00、0.88 和 0.45。

3.2. 性能比较

表1 比较了所提出的 TWFR-GMM 和 SMOTE-TWFR-GMM 与均值高斯混合模型、最大值高斯混合模型及基于深度学习的方法（自编码器 [3], MobileNetV2 [3] 和 STgram-MFN [5]）在 DCASE 2022 Task2 开发数据集上的表现，其中自编码器和 MobileNetV2 是任务2的基线方法。与基于深度学习的方法和非深度学习的高斯混合模型方法相比，TWFR-GMM 显著提升了平均曲线下面积和部分曲线下面积表现。此外，SMOTE-TWFR-GMM 通过在目标域中采用过采样策略进一步提升了性能，证明了所提出的方法在域偏移情况下用于异常声音检测的有效性。本文提出的 SMOTE-TWFR-GMM 是提交至 DCASE 2022 挑战赛任务 2 的方法的核心，并在比赛中获得第 3 名 [16]。

我们还比较了

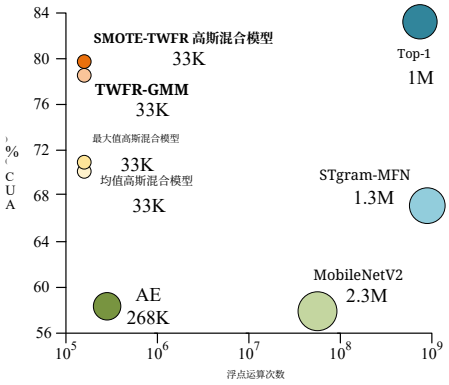


图 2：模型规模与复杂度（参数数量和浮点运算次数）的比较，其中模型规模以圆圈大小表示，浮点运算次数以对数刻度表示。

上述方法与 DCASE 2022挑战赛任务2 中排名 Top-1 的方法（自监督深度学习方法）的模型大小和复杂度 [17]，如图2所示。所提出的 TWFR-GMM 方法显著优于上述基于深度学习的方法（自编码器、MobileNetV2 和 STgram-MFN），且模型大小和复杂度都小得多。与排名 Top-1 的方法（83.28%）相比，我们的 SMOTE-TWFR 高斯混合模型（79.96%）表现略差。我们的方法模型更小，模型复杂度也显著降低。具体而言，排名 Top-1 的方法模型大小为 4.1 M，浮点运算次数为 700 M，而我们提出的方法模型大小仅为 33 k，浮点运算次数仅为 164 k。这表明，与基线方法相比，我们提出的方法在模型大小和复杂度大幅降低的同时，检测性能更优或相当。

3.3. 基于马哈拉诺比斯距离的可视化分析

我们通过使用 t 分布随机邻居嵌入（t-SNE）进行可视化分析，来评估所提出的 TWFR-GMM 的有效性 [12]，并采用基于马哈拉诺比斯距离（MD）的度量函数，定义如下

$$\mathcal{M}(\mathbf{y}_1, \mathbf{y}_2) = \min_{k \in [1,K]} \sqrt{(\mathbf{y}_1 - \mathbf{y}_2)^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_1 - \mathbf{y}_2)}, \quad (4)$$

其中，$\boldsymbol{\Sigma}_k$ 是训练好的 GMM 中第 k 个混合分量的协方差矩阵，$\mathbf{y}_1$ 和$\mathbf{y}_2$ 是分别来自训练集和测试集的两个特征向量，涵盖源域和目标域。请注意，我们使用 MD 度量函数，而非欧氏距离（ED），因为 ED 的表现不佳；具体而言，我们观察到，即使曲线下面积较高，异常特征和正常特征仍无法在可视化中分离。

图 3展示了 DCASE 2022 任务2开发数据集中 Valve 第 02 节的 TWFR-GMM、均值高斯混合模型和最大值高斯混合模型的 t-SNE 可视化结果，以及 MD

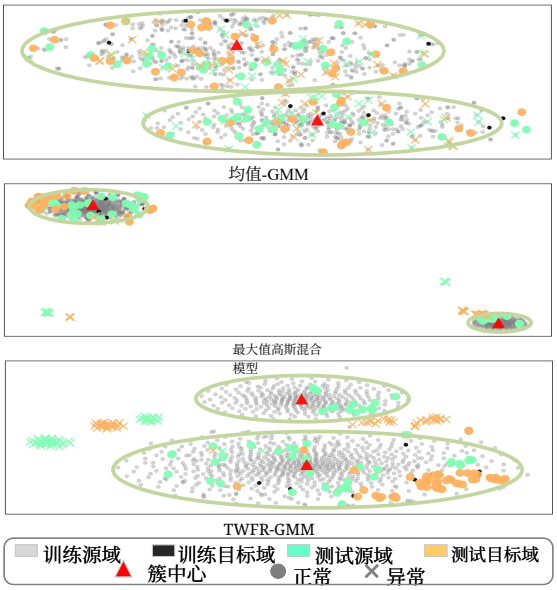


图 3：Valve 的第 02 节中，基于马哈拉诺比斯距离的 Mean-GMM、Max-GMM 和 TWFR-GMM 度量函数的 t-SNE 可视化结果。此处，采用训练所得 GMM 的均值向量作为簇中心，“•”和“×”分别表示正常样本和异常样本。

指标。这里展示了源域（训练源域）和目标域（训练目标域）的正常训练声音分布，以说明使用不同方法检测到的源域（测试源域）和目标域（测试目标域）声音如何符合正常声音的分布。

在这种情况下（即 Valve 的第 02 节），Mean-GMM 无法区分正常特征和异常特征，而 Max-GMM 可以做到。然而，在 Max-GMM 中，一些异常特征与正常特征非常接近（即仅位于簇的外部）。相比之下，尽管存在域偏移且目标域中的训练数据较为稀疏，所提出的 TWFR-GMM 仍能在马哈拉诺比斯距离空间中较好地区分异常声音。

4. 结论

本文提出了一种新的无监督异常声音检测方法，即采用高斯混合模型的时间加权频域表示（TWFR-GMM）。该方法复杂度较低，因此适用于计算资源有限的应用场景。在 DCASE 2022 Task2 数据集上的实验表明，该方法优于若干对比方法。此外，本文通过基于马哈拉诺比斯距离的度量函数进行 t-SNE 可视化，分析了该方法的有效性。尽管该方法在 DCASE 2022 Task2 上未取得最佳性能，但它为通过改进音频表示来优化模型提供了一个有前景的方向。

5. 参考文献

[1] Y. Koizumi、Y. Kawaguchi、K. Imoto、T. Nakamura、Y. Nikaido、R. Tanabe、H. Purohit、K. Suefusa、T. Endo、M. Yasuda 和 N. Harada，“DCASE2020 挑战赛任务2：面向机器状态监测的无监督异常声音检测的描述与讨论”，载于《声学场景与事件检测和分类研讨会（*DCASE*）论文集》，2020年，第81–85页。[2] Y. Kawaguchi、K. Imoto、Y. Koizumi、N. Harada、D. Ni- izumi、K. Dohi、R. Tanabe、H. Purohit 和 T. Endo，“DCASE 2021 挑战赛任务2：域偏移条件下机器状态监测的无监督异常声音检测的描述与讨论”，*arXiv* 预印本 *arXiv: 2106.04492*，2021年。[3] K. Dohi、K. Imoto、N. Harada、D. Niizumi、Y. Koizumi、T. Nishida、H. Purohit、T. Endo、M. Yamamoto 和 Y. Kawaguchi，“DCASE 2022 挑战赛任务2：应用域泛化技术进行机器状态监测的无监督异常声音检测的描述与讨论”，*arXiv* 预印本 *arXiv: 2206.05876*，2022年。[4] R. Giri、S. V. Tenneti、F. Cheng、K. Helwani、U. Isik 和 A. Krishnaswamy，“用于检测异常声音的自监督分类”，载于《声学场景与事件检测和分类研讨会（*DCASE*）论文集》，2020年，第46–50页。[5] Y. Liu、J. Guan、Q. Zhu 和 W. Wang，“利用频谱—时间信息融合进行异常声音检测”，载于《国际声学、语音与信号处理会议（*ICASSP*）论文集》，2022年，第816–820页。[6] M. Sandler、A. Howard、M. Zhu、A. Zhmoginov 和 L.-C. Chen，“MobileNetV2：倒残差与线性瓶颈”，载于《计算机视觉与模式识别会议（*CVPR*）论文集》，2018年，第4510–4520页。[7] K. Wilkinghoff，“子簇 AdaCos：学习用于异常声音检测的表示”，载于《国际神经网络联合会议（*IJCNN*）论文集》，2021年，第1–8页。[8] K. He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，载于《计算机视觉与模式识别会议（*CVPR*）论文集》，2016年，第770–778页。[9] K. Suefusa、T. Nishida、H. Purohit、R. Tanabe、T. Endo 和 Y. Kawaguchi，“基于异常声音检测基于插值的深度神经网络，”收录于声学、语音与信号处理国际会议（*ICASSP*）论文集，2020年，第271–275页。[10]K. Dohi、T. Endo、H. Purohit、R. Tanabe 和 Y. Kawaguchi，“基于流的自监督密度估计用于异常声音检测”，收录于声学、语音与信号处理国际会议（*ICASSP*）论文集，2021年，第336–340页。[11] A. Kolesnikov 和 C. H. Lampert，“种子、扩展与约束：弱监督图像分割的三项原则”，收录于欧洲计算机视觉会议（*ECCV*）论文集，2016年，第695–711页。[12] L. Van der Maaten 和 G. Hinton，“使用 t-SNE 可视化数据。”机器学习研究期刊（*JMLR*），第9卷，第11期，2008年。[13]N. V. Chawla、K. W. Bowyer、L. O. Hall 和 W. P. Kegelmeyer，“SMOTE：合成少数类过采样技术”，人工智能研究期刊（*JAIR*），第16卷，第321–357页，2002年。[14] F. Pedregosa、G. Varoquaux、A. Gramfort、V. Michel、B. Thirion、O. Grisel、M. Blondel、P. Prettenhofer、R. Weiss、V. Dubourg、J. Vanderplas、A. Passos、D. Cournapeau、M. Brucher、M. Perrot 和 E. Duschesnay，“Scikit-learn：Python 机器学习工具”，机器学习研究期刊（*JMLR*），第12卷，第2825–2830页，2011年。[15]G. Lemaˆıtre、F. Nogueira 和 C. K. Aridas，“Imbalanced-learn：用于应对机器学习中数据集不平衡问题的 Python 工具箱”，机器学习研究期刊（*JMLR*），第18卷，第17期，第1–5页，2017年。[16] F. Xiao、Y. Liu、Y. Wei、J. Guan、Q. Zhu、T. Zheng 和 J. Han，“DCASE2022 挑战赛任务2系统：基于自监督属性分类和 GMM 聚类的异常声音检测”，DCASE2022 挑战赛，技术报告，2022年7月。[17] Y. Zeng、H. Liu、L. Xu、Y. Zhou 和 L. Gan，“用于机器状态监测的稳健异常声音检测框架”，DCASE2022 挑战赛，技术报告，2022年7月。