arXiv:2602.01060v1 [cs.SD] 1 Feb 2026

TLDiffGAN：一种融合时序信息的潜在扩散生成对抗网络框架，用于异常声音检测

马成源[1], 贾鹏[2],郭红月[2],杨文明[1⋆]

[1]清华大学深圳国际研究生院， 中国[2]大连海事大学交通运输研究协同创新中心， 中国

摘要

现有用于无监督异常声音检测的生成模型受限于无法充分捕捉正常声音的复杂特征分布，而强大的扩散模型在该领域的潜力仍未得到充分探索。为应对这一挑战，我们提出了一种新颖的框架 TLDiffGAN，该框架由两个互补分支组成。其中一个分支将潜在扩散模型融入生成对抗网络生成器中进行对抗训练，从而增加判别器的任务难度，并提升生成样本的质量。另一个分支利用预训练音频模型编码器直接从原始音频波形中提取特征，用于辅助判别。该框架能够有效地从原始音频和梅尔频谱图两个方面捕捉正常声音的特征表示。此外，我们引入了一种 TMixup 频谱图增强技术，以增强模型对细微且局部化时序模式的敏感性，而这些模式通常容易被忽略。在 DCASE 2020 挑战赛任务 2 数据集上进行的大量实验表明，TLDiffGAN 具有卓越的检测性能，并且在异常时频定位方面展现出强大的能力。

索引词— 异常声音检测、特征融合、无监督学习、潜在扩散模型、生成对抗网络。

1. 引言

异常声音检测（ASD）旨在仅使用正常声音数据来识别此前未见过的异常事件 [1], ，在工业设备智能监控等应用中具有重要意义。现有主流方法大致可分为两类。第一类是基于分类的自监督模型 [2, 3, 4], ，这类模型在特定条件下表现出色，但通常高度依赖机器ID等元数据。此外，其决策过程缺乏内在可解释性，也无法在时频域中定位异常，这严重限制了其泛化能力。鉴于这些局限，本文聚焦于第二类方法——基于重构的生成模型。这类模型无需元数据，并且能够通过分析重构残差，为异常定位和可解释性提供内在支持。

生成模型通过学习正常声音的分布来识别异常。然而，早期的自编码器（AEs）及其变体 [5, 6, 7] 在学习机制上存在固有局限，因此难以捕捉正常声音特征的完整分布。结果，模型往往会学到相对宽泛的重建映射，不仅适用于正-常声音，还可能部分适用于结构相似的异常声音，导致异常声音的重建误差不足。随后，人们引入了生成对抗网络（GANs）[8, 9, 10] ，但其固有的训练不稳定性和模式崩溃风险限制了其在复杂声学场景中的应用。近年来，去噪扩散概率模型（DDPMs） [11] 展现出卓越的生成能力，但也带来了新的挑战。其强大的分布学习能力可能导致模型将异常特征视为噪声，并在重建过程中将其去除，从而生成与异常输入仅有细微差异的分布，使异常检测更加困难。此外，一个常被忽视的局限是，上述生成模型几乎都只使用梅尔频谱图作为输入。然而，在时频变换过程中，频谱图不可避免地会丢失原始波形中包含的部分关键信息 [12]。这种单模态输入的缺陷从根本上限制了模型性能的上限。此外，现有生成模型在训练期间往往倾向于捕捉正常声音的全局和宏观特征。这种学习范式可能使模型对仅出现在局部时频区域中的微弱瞬态变化不够敏感。

为应对上述挑战，本文提出一种新颖的异常声音检测框架，旨在克服现有生成模型在重建保真度、训练稳定性和异常敏感性之间固有的权衡。该框架通过有效整合来自频谱图和原始波形的互补声学信息，大幅增强生成模型的异常检测能力，最终实现对异常的优越时频定位。本研究的主要贡献总结如下：

（1）我们提出一种双分支声学建模框架，将声谱图重建与原始波形特征提取相结合。该框架通过创新的潜在扩散生成对抗网络（LDGAN）和预训练编码器，有效融合来自对数梅尔频谱图和原始波形的互补声学信息。

（2）我们设计了一种自适应时间混合模块（TMixup），利用注意力机制识别并增强位于正常数据分布边界处的特征区域，从而提升模型在正常声音与异常声音之间的决策边界上的判别能力。

(3) 在 DCASE 2020 任务 2 基准数据集上进行的大量实验表明，我们提出的框架在多项关键指标上显著优于主流生成模型，验证了其设计的有效性和卓越性能

⋆ 通讯作者 本研究部分受中国国家重点研发计划（2023YFB4302200）和深圳市战略性新兴产业发展专项资金（KJZD20231023094700001）资助。

性能。

2. 所提方法

我们所提方法的流程如图 1 所示。本节将全面介绍相关细节。第 2.1 节介绍用于对数梅尔频谱图特征提取的潜在扩散生成对抗网络骨干模块。第 2.2 节讨论旨在增强对数梅尔频谱图时序特征的 TMixup 模块。第 2.3 节概述本方法采用的预训练音频编码器模块，最后，第 2.4 节介绍本方法的检测器组件。

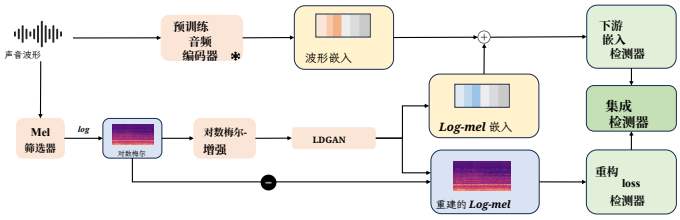


图 1. 我们的 TLDiffGAN 的主要流程。

2.1. LDGAN 主干网络

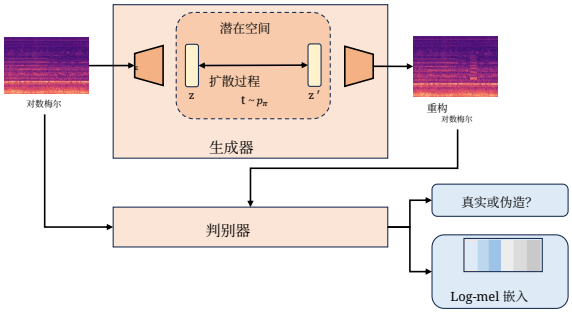


图 2. LDGAN 框架。

我们框架的核心是一种创新的 LDGAN 主干网络（图 2），它将潜在扩散模型（LDM）的逐步去噪能力与生成对抗网络（GANs）的对抗训练机制深度融合，以重建高质量的正常声音梅尔频谱图。与直接对数据分布建模的传统生成对抗网络生成器不同，我们的生成器通过从噪声到数据的渐进式反向去噪过程来学习合成频谱图。该过程在低维潜在空间中进行，并由 GAN 判别器引导，其设计旨在将扩散模型的高生成质量与 GAN 框架的训练稳定性相结合。

在训练过程中，生成器以双重目标进行优化：它既必须准确学习扩散模型的反向去噪轨迹，又必须确保最终生成样本在特征空间中与真实样本无法区分。因此，总体生成器损失 $\mathcal{L}_G$ 被定义为噪声预测损失 $\mathcal{L}_{\text{noise}}$ 与统计匹配损失 $\mathcal{L}_{\text{stat}}$ 的加权组合：

$$\mathcal{L}_G = \mathcal{L}_{\text{noise}} + \lambda_{\text{stat}} \cdot \mathcal{L}_{\text{stat}} \tag{1}$$

具体而言，噪声预测损失 $\mathcal{L}_{\text{noise}}$ 遵循标准的 LDM 目标，通过最小化预测噪声与真实噪声之间的 L_2 误差来实现。相比之下，统计匹配损失 $\mathcal{L}_{\text{stat}}$ 通过最小化判别器提取的真实样本与生成样本深层特征之间的 L_2 距离，使二者的特征分布保持一致。

判别器 D则面临更具挑战性的任务：它不仅必须区分真实频谱图与最终生成的频谱图，还必须评估中间去噪步骤，从而在每个阶段为生成器的细化提供梯度信号。为稳定训练，我们对 D 施加谱归一化，并引入梯度惩罚项 $\mathcal{L}_{\text{GP}}$ [13]。判别器的总损失表述为：

$$\mathcal{L}_D = \mathcal{L}_{\text{adv}} + \lambda_{\text{GP}} \cdot \mathcal{L}_{\text{GP}} \tag{2}$$

其中 $\mathcal{L}_{\text{adv}}$ 表示标准对抗损失。

通过将潜在扩散过程整合到GAN框架中，生成器沿着结构化、渐进式的细化路径学习重建，而对抗目标则驱动合成感知上更清晰、更真实的频谱细节。这有效克服了基于自编码器的方法中常见的重建模糊问题。此外，扩散的结构化特性为对抗训练提供了强有力的正则化效果，从而缓解了GAN固有的模式崩溃风险。

2.2. 对数梅尔频谱图增强

为增强对数梅尔频谱图中的时序特征表示，我们在 [14] 的基础上引入幂平均池化和可训练权重，对模块进行了改进。具体而言，我们定义了三种池化操作：最大池化、平均池化和幂平均池化 [15],，并使用可训练权重 $\hat{}w_{\max}$、$\hat{}w_{\text{avg}}$ 和 $\hat{}w_{\text{pow}}$ 对其结果进行加权求和。

$$\begin{aligned}\text{Pool}(x_{\text{mel}}) = \hat{w}_{\max} \cdot \text{MaxPool}(x_{\text{mel}}^T) + \hat{w}_{\text{avg}} \cdot \text{AvgPool}(x_{\text{mel}}^T) \\ + \hat{w}_{\text{pow}} \cdot \text{PowerAvgPool}(x_{\text{mel}}^T)\end{aligned} \tag{3}$$

其中，$x_{\text{mel}} \in \mathbb{R}^{F \times T}$ 表示对数梅尔频谱图，F 和 T 分别表示频率维度和时间维度；$\hat{}w_{\max}$、$\hat{}w_{\text{avg}}$ 和 $\hat{}w_{\text{pow}}$ 为可训练权重参数，并通过 Softmax 进行归一化，以确保其总和为 1。这些权重在训练过程中通过反向传播自动调整，以优化时序特征的提取。为进一步优化池化结果，我们首先应用 Sigmoid 激活函数，将加权池化输出转换为软时间注意力图 $x_{TA} \in \mathbb{R}^T$，其取值范围为 (0，1)，用于量化每个时间帧偏离正常模式的置信度。随后，为准确定位并选择最可疑、需要增强的区域，我们使用在每个训练步骤中随机采样的阈值 τ，将该软注意力图转换为硬掩码 $M_{\text{mask}} = \mathbb{I}(x_{TA} > \tau)$，其中 $\mathbb{I}(\cdot)$ 表示指示函数，用于明确指定后续局部 Mixup 操作所针对的区域。随后，我们在这些高注意力区域执行 Mixup 局部混合操作：

$$x_{\text{mixed}} = \lambda \cdot x_{\text{mel}} + (1 - \lambda) \cdot [x_{\text{mel}} \otimes M_{\text{mask}}] \tag{4}$$

其中，混合系数 $\lambda \sim$ 贝塔分布(α， α)控制原始频谱图与局部时间注意力特征之间的加权混合比例，而 $\otimes$ 表示逐元素乘法。通过在 TMixup 模块中使用对数梅尔频谱图，模型能够捕获多个时间尺度上的特征，从而有效检测偏离正常模式的区域。我们

通过对正常数据分布边界上的区域应用局部增强，我们刻意增强模型对细微特征变化的敏感度。这促使模型在训练期间学习关注正常范围的边界，从而能够在推理时更有效地检测真正的异常。

2.3. 预训练音频编码器

为了弥补梅尔频谱图中可能的信息损失，我们设计了一个并行分支，利用预训练的自监督学习（SSL）模型直接从原始音频波形中提取深层特征。工业设备声音的声学特性往往与语音信号存在显著差异，通常表现为长期平稳模式和稀疏的瞬态事件。因此，通用音频预训练模型比语音专用模型更适合该任务。

在本研究中，我们主要探讨强大的基于Transformer的架构。例如，AST [16] 及其自监督变体ATST [17] 将视觉Transformer架构应用于音频频谱图，从而能够高效建模声音事件的全局上下文信息。另一代表性模型BEATs[18], 采用带有学习得到的音频分词器的迭代预训练框架，以提炼语义更丰富的高层声学表示。此外，EAT[19] 等高效音频变换器侧重于捕捉长程依赖关系，这对于分析稳定运行状态与异常状态之间细微的声学变化至关重要。通过集成这些编码器，我们的框架从原始波形中获得稳健的嵌入，从而保留在转换为梅尔频谱图过程中可能被平滑掉或忽略的关键声学细节。

2.4. 检测器

我们的检测框架由两个并行组件构成：基于重建的检测器和基于嵌入的检测器。首先，基于重建的检测器旨在量化给定样本与模型所学习的正常声音潜在流形的吻合程度。当输入为异常样本时，其特征会偏离所学习的正态分布，使得准确重建变得困难，因而产生较大的差异。异常分数 s_r 通过真实样本与重构样本的潜在表征之间的欧几里得距离平方计算得出：

$$s_r = ||z_{\text{real}} - z_{\text{rec}}||_2^2 \tag{5}$$

其次，基于嵌入的检测器在高维联合特征空间（$\mathcal{Z}$）中运行，该空间由梅尔频谱图特征 $\mathcal{Z}_{\text{mel}}$ 与原始波形特征 $\mathcal{Z}_{\text{wave}}$拼接而成。为全面捕捉不同类型的异常模式，我们采用了一组互补的经典算法：K近邻（KNN） [20]用于衡量样本与正常特征流形之间的距离；局部离群因子（LOF） [21] 用于识别局部密度异常的样本；高斯混合模型（GMM） [22] 用于估计样本在正常声音复杂分布下的似然度；子空间离群分数（SOS） [23] 用于检测仅在特定特征子空间中才显著的细微异常。我们的方法采用基于验证的度量选择策略。具体而言，对于每种机器类型，我们仅依据验证集性能独立选择表现最佳的检测器；该配置在推理过程中保持固定，确保整个过程不依赖机器ID。最终的异常分数按如下方式确定：

$$s_{\text{final}}^{(k)} = \begin{cases} s_r & \text{if } \mathcal{M}_k^{(r)} = \max(\{\mathcal{M}_k^{(m)}\}) \\ s_{e_1} & \text{if } \mathcal{M}_k^{(e_1)} = \max(\{\mathcal{M}_k^{(m)}\}) \\ \vdots & \vdots \\ s_{e_4} & \text{if } \mathcal{M}_k^{(e_4)} = \max(\{\mathcal{M}_k^{(m)}\}) \end{cases} \tag{6}$$

其中，$\mathcal{M}_k^{(m)}$ 表示第 m个检测器在第 k种机器类型上的整体性能指标，定义为曲线下面积和部分曲线下面积的算术平均值。

3. 实验与结果

3.1. 实验设置

数据集： 我们在 DCASE 2020 挑战赛任务 2 数据集 [24, 25], 上开展实验。该数据集包含两个子集 MIMII 和 ToyADMOS, 涵盖六种机器类型：风扇、泵、滑块、阀门、玩具汽车和玩具传送带。每个音频样本均为单声道录制，采样率为 16 kHz, 时长约为 10 秒。我们没有采用后续版本的 DCASE 数据集，因为这些版本主要侧重于领域偏移相关挑战。相比之下，本研究的核心重点是在单一领域内对正常声音的分布进行建模。

评估指标： 我们采用 ROC 曲线下面积（AUC）和在受控假阳性率下的部分曲线下面积（pAUC）作为评估指标。遵循官方挑战赛协议[1], , pAUC 在低假阳性率区间 $[0, p]$内计算，其中 p 设为 0.1。对于所有指标，数值越高表示性能越好。

实现细节： 对于输入特征，原始音频波形被转换为维度为 128×313的对数梅尔频谱图。在 TMixup 模块中，时间注意力图的阈值 τ 从均匀分布 U(0.2,0.5) 中随机采样。关于网络配置，LDGAN 主干网络中的判别器与编码器结构相似，但在最后一层采用分组卷积，以提取深层语义特征。损失加权系数 λ_{stat} 和 λ_{GP} 分别设为 1.0 和 10。模型使用 Adam 优化器进行训练，学习率为 0.0001，批量大小为 512，总计训练 150 个周期。我们对判别器施加梯度惩罚，以进一步提升训练稳定性。

3.2. 性能比较

如表 1 所示，我们将 TLDiffGAN 与其他竞争方法进行了性能比较。由于生成模型不依赖元数据且能够定位异常区域，我们将 TLDiffGAN 与其他竞争生成模型进行了比较，尤其是在滑块和泵数据集上，AUC 和 pAUC 分别提升了 2.94%、5.76% 和 3.23%、1.13%。尽管 AEGAN-AD 在玩具传送带数据集上的 pAUC 表现略优，但在绝大多数机器类型上，我们提出的方法在 AUC 和 pAUC 方面均展现出更优的检测性能。此外，TLDiffGAN 的平均 AUC 和 pAUC 得分分别达到 88.60% 和 74.35%，为最佳成绩，进一步证明了我们模型的优越性。

3.3. 预训练音频编码器比较

不同预训练模型因架构设计各异，在 ASD任务上的表现水平也有所不同。如表2所示，仅替换预训练音频编码器的实验结果表明，EAT表现最佳，其次是BEATs，而其他模型的表现相对

表 1：在 DCASE 2020 任务 2 数据集上与其他方法的性能比较。官方基线：DCASE 提供的自编码器基线。平均值：所有机器类型的曲线下面积和部分曲线下面积的平均值。

方法	Fan		Pump		滑块		玩具汽车		玩具传送带		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		
官方基线 [1]	65.91	52.74	71.36	60.02	83.86	66.42	78.23	67.38	71.01	59.78	65.28	49.98	72.61	59.39
ANP [7]	69.20	54.40	72.80	61.80	90.70	74.20	86.90	70.70	72.50	67.30	67.00	54.50	76.52	63.82
GANomaly [9]	79.37	63.48	72.65	61.48	84.21	72.84	85.12	72.23	74.59	61.24	79.30	57.74	79.21	64.84
ASD-Diffusion [26]	83.64	71.92	82.78	74.92	88.51	75.24	92.30	81.48	78.65	63.12	87.78	61.55	85.61	71.37
AEGAN-AD [10]	83.12	71.86	84.37	75.42	91.84	78.18	91.70	80.40	79.00	**65.86**	84.37	60.75	86.08	72.08
Ours	**85.88**	**73.15**	**87.60**	**76.55**	**94.78**	**83.94**	**93.35**	**82.97**	**80.29**	65.21	**89.67**	**64.26**	**88.60**	**74.35**

较差。EAT的大块掩码解除策略进一步增强了其捕捉长期依赖关系的能力，因此特别适合检测机器声音中常见的细微异常特征。

表2：不同预训练音频编码器在六种机器类型上的平均曲线下面积（%）和部分曲线下面积（%）比较。

模型	Size	平均值	
		曲线下面积	部分曲线下面积
AST [16]	86M	85.47	71.40
ATST [17]	85M	85.85	71.24
BEATs [18]	90M	86.92	73.98
EAT [19]	88M	**88.60**	**74.35**

3.4. 消融研究

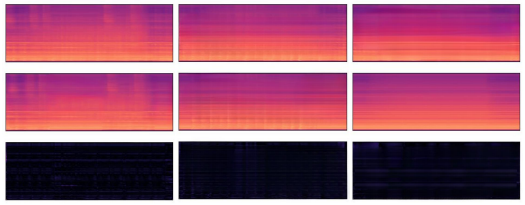
为评估框架中各核心组件的贡献，我们进行了消融研究，结果汇总于表3。平均性能结果清楚地表明，移除潜在扩散、EAT 编码器或对数梅尔频谱图增强模块中的任一组件，都会导致模型整体性能下降。对于风扇这类声学模式高度平稳且简单的声源，加入额外的原始波形特征可能会引入冗余信息。相比之下，对于玩具传送带这类正常运行本身就包含复杂时序模式的声源，TMixup 模块的强增强效果可能会轻微干扰正态分布的学习。尽管如此，完整的所提模型在所有机器类型上始终取得最佳的整体平均性能，从而验证了我们方法的有效性。

表3：TLDiffGAN 核心组件的消融研究，结果以曲线下面积（%）表示。

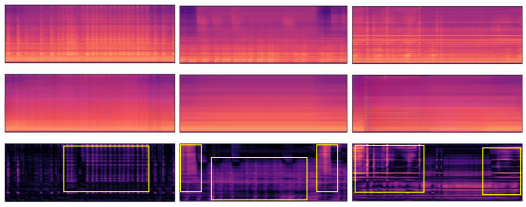
模型	TLDiffGAN	w/o 潜在扩散	w/o EAT	w/o 对数梅尔增强
Fan	85.88	83.27	**86.35**	85.14
Pump	**87.60**	85.95	87.28	86.55
滑块	**94.78**	92.52	90.94	91.69
玩具汽车	**93.35**	92.14	88.51	93.10
玩具传送带	80.29	79.85	78.32	**80.86**
阀门	**89.67**	84.68	86.73	88.50
平均值	**88.60**	86.40	86.36	87.64

3.5. 频谱图区域定位

与其他模型相比，生成式重建模型在工业声学异常检测中展现出显著的时频定位优势。如图 3 所示，对于玩具汽车设备的正常样本（a），其重构频谱图与训练集平均频谱图具有较高的空间一致性，差异频谱图中仅呈现微弱且随机分布的响应。这证实了模型能够精准捕捉设备的稳态运行特征，同时对背景噪声干扰保持较强的鲁棒性。相比之下，对于异常样本（b），差异频谱图呈现出明显的结构化异常响应，在时频域中既有连续的高强度带状图案，也有离散的脉冲状亮点。这充分展示了模型区分声学信号中稳态分量与瞬态分量的能力，也进一步验证了我们所提方法的有效性。



(a) 玩具汽车正常样本



（b）玩具汽车异常样本

图 3：测试集中玩具汽车设备正常样本和异常样本的异常定位结果。第一行表示模型重构的频谱图，第二行显示训练集的平均频谱图，第三行展示两者之间的差异。颜色越亮，表示异常程度越高。

4. 结论

本文提出了一种新颖的生成框架 TLD- iffGAN，用于机器声音的无监督异常声音检测。该框架以潜在扩散生成对抗网络（LDGAN）为骨干，以实现高保真频谱图重建，并结合双分支结构与自适应时间混合模块，融合多模态声学特征，增强对细微时序模式的敏感性。在 DCASE 2020 任务 2 数据集上的实验结果表明，本方法性能优越，并能有效定位异常声音事件。

5. 参考文献

[1] Y. Koizumi 等，“DCASE2020 挑战任务 2 的描述与讨论：面向机器状态监测的无监督异常声音检测”，载于*2020*年声学场景与事件检测和分类研讨会论文集（*DCASE2020*），2020年11月，第81–85页。[2] Ritwik Giri、Srikanth V. Tenneti、Fangzhou Cheng、Karim Helwani、Umut Isik 和 Arvindh 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论文集（*DCASE*），2020年，第46–50页。[3] Haihui Chen、Likai Ran、Xixia Sun 和 Chao Cai，“Sw-wavenet：利用 WaveNet 从频谱图和波形图学习表征以进行异常声音检测”，载于*ICASSP 2023 - 2023*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），2023年，第1–5页。[4] J. Guan、F. Xiao、Y. Liu、Q. Zhu 和 W. Wang，“利用基于机器 ID 的对比学习预训练音频表征进行异常声音检测”，载于*IEEE*声学、语音与信号处理国际会议论文集（*ICASSP*），希腊罗得岛，2023年，IEEE，第1–5页。[5] Kaori Suefusa、Tomoya Nishida、Harsh Purohit、Ryo Tanabe、Takashi Endo 和 Yohei Kawaguchi，“基于插值深度神经网络的异常声音检测”，载于*ICASSP 2020 - 2020*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），2020年5月。[6] Sławomir Kapka，“用于无监督异常检测的 ID 条件自编码器”，*arXiv* 预印本 *arXiv:2007 05314*，2020年。。[7] Gordon Wichern、Ankush Chakrabarty、Zhong-Qiu Wang 和 Jonathan Le Roux，“基于注意力神经过程的异常声音检测”，载于*2021*年*IEEE*音频与声学信号处理应用研讨会（*WASPAA*）。IEEE，2021年，第186–190页。[8] Thomas Schlegl、Philipp Seeb¨ock、Sebastian M. Waldstein、Georg Langs 和 Ursula Schmidt-Erfurth，“f-AnoGAN：基于生成对抗网络的快速无监督异常检测”，《医学图像分析》，第54卷，第30–44页，2019年。[9] Xiaowei Huang、Fabin Guo 和 Long Chen，“一种用于机器声音异常检测的 Res-GANomaly 方法”，《*IEEE Access*》，第12卷，第80099–80114页，2024年。[10] Anbai Jiang、Wei-Qiang Zhang、Yufeng Deng、Pingyi Fan 和 Jia Liu，“机器音频的无监督异常检测与定位：一种基于 GAN 的方法”，载于*ICASSP 2023 - 2023*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），2023年，第1–5页。[11] Jonathan Ho、Ajay Jain 和 Pieter Abbeel，“去噪扩散概率模型”，神经信息处理系统进展，第33卷，第6840–6851页，2020年。[12] Youde Liu、Jian Guan、Qiaoxi Zhu 和 Wenwu Wang，“基于谱时信息融合的异常声音检测”，载于*ICASSP 2022 - 2022*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），2022年，第816–820页。[13] Ishaan Gulrajani、Faruk Ahmed、Martin Arjovsky、Vincent Dumoulin 和 Aaron C Courville，“Wasserstein GAN 的改进训练”，神经信息处理系统进展，第30卷，2017年。

[14] Soonhyeon Choi 和 Jung-Woo Choi，《Noisy-arcmix：结合 Mixup 的加性噪声角度间隔损失，用于异常声音检测》，载于*ICASSP 2024 - 2024*年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*），2024年，第 516–520 页。[15] Yuzhuo Liu、Hangting Chen、Yun Wang 和 Pengyuan Zhang，《Power pooling：一种用于弱标注声音事件检测的自适应池化函数》，载于*2021*年国际神经网络联合会议（*IJCNN*）。IEEE，2021年，第 1–7 页。[16] Yuan Gong、Yu-An Chung 和 James Glass，《AST：音频频谱图变换器》，*arXiv* 预印本 *arXiv:2104 01778*，2021年。。[17] Xian Li、Nian Shao 和 Xiaofei Li，《用于片段级和帧级任务的自监督音频教师-学生变换器》，*IEEE/ACM* 音频、语音与语言处理汇刊，2024年。[18] Sanyuan Chen、Yu Wu、Chengyi Wang、Shujie Liu、Daniel Tompkins、Zhuo Chen 和 Furu Wei，《BEATs：使用声学分词器进行音频预训练》，*arXiv* 预印本 *arXiv:2212.09058*，2022年。[19] Wenxi Chen、Yuzhe Liang、Ziyang Ma、Zhisheng Zheng 和 Xie Chen，《EAT：使用高效音频变换器进行自监督预训练》，*arXiv* 预印本 *arXiv:2401.03497*，2024年。[20] Sridhar Ramaswamy、Rajeev Rastogi 和 Kyuseok Shim，《从大型数据集中挖掘离群点的高效算法》，载于*2000*年 *ACM SIGMOD* 数据管理国际会议论文集，2000年，第 427–438 页。[21] Markus M Breunig、Hans-Peter Kriegel、Raymond T Ng 和 J¨org Sander，《LOF：识别基于密度的局部离群点》，载于*2000*年 *ACM SIGMOD* 数据管理国际会议论文集，2000年，第 93–104 页。[22] Bo Zong、Qi Song、Martin Renqiang Min、Wei Cheng、Cristian Lumezanu、Daeki Cho 和 Haifeng Chen，《用于无监督异常检测的深度自编码高斯混合模型》，载于国际学习表征会议，2018年。[23] Jeroen Janssens、Ferenc Husz´ar、Eric Postma 和 HJ van den Herik，《随机离群点选择》，蒂尔堡创造性计算中心技术报告，第 1 卷，第 2012 页，2012年。[24] Harsh Purohit、Ryo Tanabe、Takeshi Ichige、Takashi Endo、Yuki Nikaido、Kaori Suefusa 和 Yohei Kawaguchi，《MIMII数据集：用于故障工业机器调查与检测的声音数据集》，载于声学场景与事件检测和分类 *2019*年研讨会论文集（*DCASE2019*），2019年11月，第 209–213 页。[25] Yuma Koizumi、Shoichiro Saito、Hisashi Uematsu、Noboru Harada 和 Keisuke Imoto，《ToyADMOS：用于异常声音检测的微型机器运行声音数据集》，载于*IEEE*音频与声学信号处理应用研讨会（*WASPAA*）论文集，2019年11月，第 308–312 页。[26] Fengrun Zhang、Xiang Xie 和 Kai Guo，《ASD-Diffusion：使用扩散模型进行异常声音检测》，载于国际模式识别会议。Springer，2025年，第 343–355 页。