ECHO：频谱感知的分层编码方法用于变长信号处理

Yucong Zhang[1,2], Juan Liu[1,3†], Ming Li[1,2†]

[1]School of Computer Science, Wuhan University, Wuhan, China
[2]Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems,
Digital Innovation Research Center, Duke Kunshan University, Kunshan, China
[3]School of Artificial Intelligence, Wuhan University, Wuhan, China

ABSTRACT

预训练的基础模型在视觉和语言方面取得了显著的成功，但它们对于通用机器信号建模——涵盖声学、振动和其他工业传感器数据——的潜力仍处于未充分探索的状态。现有方法使用基于子带的编码器已经达到了具有竞争力的结果，但却受限于固定的输入长度，并且缺乏显式的频率位置编码。在本研究中，我们提出了一种新型基础模型，该模型结合了先进的频段分割架构与相对频率位置嵌入，能够在任意采样配置下实现精确的频谱定位。该模型支持任意长度的输入而不需填充或分段，生成一个简洁的嵌入表示，同时保留时域和频域的保真度。我们在 SIREN 上评估了我们的方法，这是一个新引入的大规模基准测试平台，用于机器信号编码，统一了多个数据集，包括所有 DCASE 任务 2 挑战（2020-2025）以及广泛使用的工业信号语料库。实验结果表明，在异常检测和故障识别方面，该模型具有一致的最先进的性能，证实了所提出模型的有效性和泛化能力。我们已将 ECHO 开源至 https://github.com/yucongzh/ECHO。

Index Terms— 异常声音检测，预训练模型，基础模型，频带分割

1. 介绍

可靠地监测机器健康对于确保安全、减少停机时间以及优化工业领域的运营效率至关重要。近年来，机器信号分析——包括声发射、振动测量和其他传感器模式——已成为检测异常状况和在灾难性故障发生前诊断故障的核心工具。传统方法如手工特征提取结合常规分类器，在狭窄的特定领域场景中已被证明是有效的 [1]。然而，这些方法通常缺乏跨不同类型机器、操作条件和传感模态的一般化能力。

大规模音频基础模型已成为跨不同声学领域表示学习的统一范式，包括语音、音乐和环境声音。利用在 AudioSet 规模语料库上的监督或自监督预训练，基于 Vision Transformer (ViT) [2] 架构的相关模型已经证明了它们对广泛下游任务的强大迁移能力，从标签分类和事件检测到字幕生成 [3–7]。研究人员进一步扩展了训练数据，并展示了在各种语音和音频事件任务中的鲁棒性能 [8]。这种成功自然延伸到了工业监控场景，在这些场景中，这些模型作为机器异常声音检测 (ASD) 和其他条件监测任务的有效前端编码器 [9–11]。在这种标注异常稀缺且操作条件多变的领域中，音频基础模型所提供的丰富和可泛化的表示为特定领域的适应提供了稳健的基础。

直接基于这一趋势，FISHER [12] 最近被提出作为一种专门针对多模态工业信号的基础模型，包括声学和振动数据。通过采用子带建模策略——将频谱图划分为固定带宽的组件并按子带提取表示——该模型能够无缝适应不同的采样率，而无需重新采样或重新训练。FISHER 在其基准测试中通过一致地捕捉多尺度频谱特征，在机器信号理解方面取得了显著改进。

尽管其具有优势，但将通用音频编码器和子带架

†Corresponding authors: Juan Liu and Ming Li.

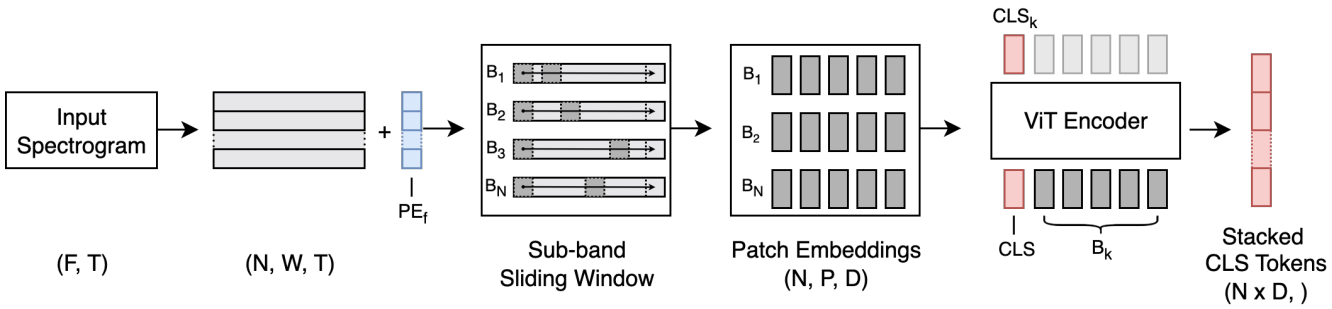

Fig. 1. ECHO 框架的特征提取流程。F：STFT 后的频率 bin 数量；T：STFT 后的时间帧数量；N：频带分割后的子带数量；W：子带的带宽；P：滑动块的数量；D：每个块的特征维度。

构（如 FISHER）应用于实际机器信号监控仍面临两个关键挑战。

首先，大多数基于 Transformer 的编码器使用固定大小的频谱图块和传统的位置嵌入，在处理变长输入时需要进行截断或插值——这些操作在不同的推理条件下会带来性能下降，特别是在长时间段或流式传输场景中尤为明显。

其次，虽然子带建模缓解了采样率不匹配的问题，但它通常忽略了显式编码每个子带在全球频率谱中的相对位置，这可能会限制跨频带交互和多模态下的尺度意识。

在这项工作中，我们通过提出**回声**（即 frEquenCy-awareHierarchicalencOding for variable-length signals）来解决这些限制。ECHO 是一款面向机器信号的通用基础模型，在包括 DCASE 挑战中的声学异常检测、基于振动的故障数据集以及多模态机器状况数据集等多样化基准测试中，实现了最先进的性能。这项工作的主要贡献总结如下：

(1) 一种适用于任意采样率和频率分辨率的相对频率位置嵌入机制，使模型能够编码全频谱内子带的显式位置上下文；

(2) 一种任意长度的输入设计，消除了信号填充或分段的需要，产生简洁的一对一信号嵌入；

(3) 一个可扩展的训练框架，能够在统一表示空间内处理多样的机器信号模态；

(4) 第一个开源的通用机器信号嵌入基准。

广泛的实验表明，我们的模型不仅在准确性和鲁棒性方面超过了先前的预训练方法，而且还能有效地跨声学和振动领域进行泛化，无需特定任务的微调。这项工作有助于推进机器信号理解的单一通用表示模型的愿景，弥合了通用音频建模研究与实际工业故障检测系统之间的差距。

2. 方法

在本节中，我们介绍了所提出的框架**回声**（即 E 频率感知的分层编码 frCquenHierarchical encOding），该框架旨在在任意采样率下进行可变长度信号的鲁棒表示学习。整体架构如图 1 所示。ECHO 包含四个关键组件：(1) 频谱提取，(2) 具有位置编码的频率感知子带分割，(3) 时间补丁提取，以及 (4) 分层 Transformer 编码。此外，我们描述了在教师-学生框架下的不同训练和推理策略。

2.1. 谱图提取

给定一个以频率 f_s 采样的输入波形 $x(t)$，我们应用一组预定义参数的短时傅里叶变换（STFT），包括快速傅里叶变换大小 N_{FFT}、窗口长度和跳跃大小（以秒为单位）。STFT 产生一个绝对值频谱矩阵 $S \in \mathbb{C}^{F\times T}$，其中 F 和 T 分别表示频率 bins 的数量和时间帧的数量。

2.2. 频率感知子带分割与位置编码

频谱图 S 沿频率轴均匀分割成一组子带。子带的数量与采样率 f_s 成正比，即较高的采样率会导致更多的子带。

对于第 k 个子带，其频率 bin 索引范围从 b_{start} 到 $b_{\text{end}}-1$，我们计算其中心频率为

$$f_c = \frac{(b_{\text{start}} + b_{\text{end}} - 1)}{2} \cdot \frac{f_s}{N_{\text{FFT}}}. \tag{1}$$

奈奎斯特频率是 $f_{\text{Nq}} = f_s/2$。然后我们通过奈奎斯特极限对中心频率进行归一化：

$$p = \frac{f_c}{f_{\text{Nq}}}. \tag{2}$$

这个归一化的位置 $p \in [0,1]$ 被注入到一个正弦位置编码方案中：

$$\begin{aligned} \text{PE}(p, 2i) &= \sin\left(\frac{\gamma \cdot p}{10000^{2i/d}}\right), \\ \text{PE}(p, 2i+1) &= \cos\left(\frac{\gamma \cdot p}{10000^{2i/d}}\right), \end{aligned} \tag{3}$$

其中 d 表示嵌入维度，γ 表示归一化位置的缩放因子。这种设计确保了不同采样率但占据等效相对频率位置的子带具有一致的位置编码。

2.3. 变长输入的时间补丁提取

为了建模不同持续时间的信号，每个子频带都进行时域分割。具体而言，我们在时间轴上应用一个长度为 L（等于子频带宽度）的滑动窗口，并以步长 $L/2$ 实现 50% 的重叠。

该操作通过一个核大小为 $(\text{sub-band height}, L)$、步幅为 $(\text{sub-band height}, L/2)$ 的二维卷积高效实现。卷积将频率维度压缩，生成一个形状为 (N, D) 的补丁序列，其中 N 是时间补丁的数量，D 是通道维度（即补丁嵌入大小）。因此每个补丁代表了子带的一个局部时间特征。

2.4. 分层编码

每个补丁序列，连同额外的可学习 CLS 标记一起，被送入 ViT 编码器。在每个子带内，CLS 标记聚合子带级别的上下文信息。对于每个子带，我们收集输出中的 CLS 标记。结果，最终输出以分层方式拼接来自每个子带的所有 CLS 标记，并形成最终嵌入。这样，Transformer 编码器能够捕捉到每个子带内的局部时间依赖性，并且借助所提出的相对频率位置编码能够在不同频带之间进行区分。

2.5. 训练阶段

在训练过程中，每个子带都被视为一个独立的输入样本。我们采用了一个受 EAT [7] 启发的教师-学生框架，但使用了一种简化的随机遮罩策略而不是专门设计的遮罩。

学生模型在掩码输入上进行训练，而教师模型的参数通过学生模型参数的指数滑动平均（EMA）进行更新：

$$\theta_{\text{teacher},t} = \alpha \cdot \theta_{\text{teacher},t-1} + (1-\alpha) \cdot \theta_{\text{student},t}, \tag{4}$$

其中 $\alpha \in [0.99, 0.999]$ 是动量系数。

我们使用与 EAT 和 FISHER 相同的训练方案，我们利用两个自监督训练目标：(1) 教师模型平均层输出的时间均值池化与学生 CLS 标记之间的全局对齐，以及 (2) 出现在掩码位置的标记上的帧级特征对齐。这种双层级监督确保了在粗粒度和细粒度上的一致性。

2.6. 推理阶段

我们仅在推理过程中使用学生模型，但使用完整的频谱图。在进行频带分割和使用 ECHO 建模后，我们简单地收集编码后的每个子频带的 CLS 标记。这些 CLS 标记被连接在一起以形成最终的全局表示：

$$\boldsymbol{z} = [\text{CLS}_1, \text{CLS}_2, \ldots, \text{CLS}_K], \tag{5}$$

其中，K 表示子频带的数量。生成的嵌入 $\boldsymbol{z}$ 作为下游任务（如分类或异常检测）的输入。

3. SIREN 标准基准测试

为了公平比较，我们开源了一个名为 SIREN 的评估基准，SIREN 是 SIgnalRepresentationEvaluatioN 用于机器 [1]。SIREN 专门针对通用信号诊断，包括少

[1] 基准工具包和源代码可在 https://github.com/yucongzh/SIREN 获得。

Table 1. SIREN 基准数据集及其特征。MAFAULDA：机械故障数据库；CWRU：轴承数据集；IIEE：IDMT 电动机数据集；IICA：IDMT 压缩空气数据集。LOOCV：留一法交叉验证；CV：交叉验证。

Dataset	Modality	Sampling Rate	#Classes	Split	Scoring
DCASE2020	Sound	16k	2	-	KNN
DCASE2021	Sound	16k	2	-	KNN
DCASE2022	Sound	16k	2	-	KNN
DCASE2023	Sound	16k	2	-	KNN
DCASE2024	Sound	16k	2	-	KNN
DCASE2025	Sound	16k	2	-	KNN
MAFAULDA	Sound/Vibration	50k	10	LOOCV	KNN
CWRU	Vibration	12k	10	LOOCV	KNN
IIEE	Sound	44.1k	3	LOOCV	KNN
IICA	Sound	48k	3	5Fold-CV	KNN

量样本异常检测任务以及机器故障诊断/分类等任务。数据集的详细信息及其对应的任务见表 1 和 GitHub 仓库。

3.1. 基准组成

DCASE **任务** 2 (2020 – 2025)。我们采用了官方的开发/评估机构，为期六年。年份 2021 – 2025 进一步包括领域（源目标）和部分因素，在聚合过程中我们保留这些以匹配挑战实践。

故障诊断/分类。我们包含了四个具有代表性的数据集，涵盖了不同的传感器和操作条件：MAFAULDA（声学/振动）、CWRU（振动）、IIEE（声学）和 IICA（声学）。这些数据集共同强调了跨模式、采样率、类别数量和操作制度的表示质量。

3.2. 评估协议

DCASE **异常检测**。我们遵循每年的官方开发/评估划分，并计算文件级别的异常分数，随后这些分数会根据机器-/部分-/领域感知分组进行汇总，与官方一致 [13–18]。我们报告每一年的开发、评估和整体（合并）结果。

故障分类。为了在有限数据下获得稳健且可比较的估计，我们在 MAFAULDA、CWRU 和 IIEE 上使用了留一法交叉验证（LOOCV），并在 IICA 上使用了五折交叉验证（如表 1 所示）。这些选择平衡了统计可靠性和计算成本，并且是文献中的常见做法。

3.3. 指标和理由

DCASE **异常检测**。我们按照比赛规则报告了 ROC-AUC 和部分 AUC（pAUC）。这些无阈值指标非常适合异常检测中常见的类别不平衡和操作点不确定性。聚合尊重特定年份的分组（例如，章节和源/目标），使忠实且可跨年度比较的摘要成为可能。

故障分类。我们报告了准确率、宏平均精确率、宏平均召回率、宏平均 F1 和各类别分数。宏平均指标缓解了类别不平衡，并且比微平均度量更好地反映了多类行为，而按类别报告揭示了特定条件下的优缺点。

总结来说，SIREN 标准化了异常检测和故障分类中的数据集、协议和指标，使得信号表示的对比简洁明了且具有可比性，不会将结果与特定任务的模型工程混淆。

4. 实验结果

4.1. 模型细节

我们采用与 ViT 相同的骨干架构。目前，我们发布了 ECHO-S 和 ECHO-B 两个版本，分别使用小模型和基础模型的 ViT 超参数。频谱图是通过在归一化的原始信号上使用 25 毫秒的窗口大小和 10 毫秒的窗口偏移量提取的。我们的模型中的子带宽度固定为 32，从而导致有 32 个频率区间。用于 ViT 的块具有与子带宽度相同的宽度和高度值，并且跳跃长度固定为块宽度的一半。

4.2. 训练详情

该模型使用四块 NVIDIA GeForce RTX 3090 GPU，全局批次大小为 256，总共训练 400,000 步。学习率策略定义如下：基础学习率 1×10^{-4} 根据有效批次大小按比例缩放，公式如下：

$$\mathrm{lr} = \mathrm{base\_lr} \times \frac{\mathrm{eff\_batch\_size}}{256}.$$

训练采用余弦学习率调度器，线性预热阶段跨越 40,000 步。最小学习率设置为 1×10^{-5}，并应用权重衰减，系数为 0.05。梯度裁剪使用 1.0 的初始范数阈值，在 60,000 训练步后降低到 0.3。

Table 2. 特征提取器在 DCASE 挑战和机器故障诊断数据集上的性能比较。DCASE 任务通过各台机器的 AUCs 和部分 AUCs 的平均值报告，而故障分类任务则通过准确率报告。对于 MAFAULDA，结果报告为“声音/振动”，其中“声音”表示麦克风通道的准确率，“振动”表示振动通道的准确率；均值从融合得分计算得出。

模型	数据集	尺度	DCASE 任务							故障分类任务					平均值
			2020	2021	2022	2023	2024	2025	Mean	IIEE	IICA	CWRU	MAFAULDA	Mean	
BEATs [5]	AS2M	Base	0.743	0.613	0.590	0.629	0.559	0.578	0.619	0.658	0.916	0.886	0.856/0.997	0.862	0.740
CED [6]	AS2M	Base	0.677	0.567	0.573	0.608	0.578	0.577	0.596	0.802	0.861	0.819	0.865/0.997	0.868	0.720
		Small	0.677	0.567	0.568	0.600	0.572	0.579	0.593	0.744	0.857	0.857	0.865/0.997	0.858	0.726
戴森 [8]	AS2M+MTG+VGG +ACAV100M	Base	0.694	0.573	0.579	0.607	0.571	0.569	0.598	0.994	0.911	0.886	0.924/0.999	0.942	0.755
		0.6B	0.685	0.568	0.566	0.568	0.568	0.566	0.586	0.991	0.921	0.895	0.924/0.997	0.945	0.749
费舍尔 [12]	AS2M+MTG	Small	0.705	0.595	0.598	0.618	0.557	0.587	0.610	0.975	0.945	0.905	0.951/1.000	0.955	0.766
		Tiny	0.706	0.585	0.571	0.585	0.553	0.577	0.596	0.998	0.954	0.752	0.951/1.000	0.931	0.748
		Mini	0.700	0.584	0.579	0.604	0.559	0.569	0.599	0.999	0.945	0.733	0.951/1.000	0.925	0.747
ECHO	AS2M+MTG	Small	0.722	0.602	0.600	0.637	0.579	0.587	0.621	0.999	0.938	0.905	0.929/1.000	0.954	0.772

4.3. 基线模型

目前，我们包含 4 个用于比较的基础模型，包括 BEATs、CED、Dasheng 和 FISHER。所有模型都使用 ViT 作为骨干模型，并使用开源音频数据集进行训练，包括 AudioSet-2M (AS2M) [19]、MTG-Jamendo (MTG) [20]、VGGSound (VGG) [21] 和 ACAV100M (ACAV) [22]。值得一提的是，BEATs、CED 和 Dasheng 通过重建掩码区域进行训练，而 FISHER 则以 EAT 方式通过师生蒸馏进行训练。

4.4. 实验结果

表 2 报告了五届连续的 DCASE 挑战赛（2020 – 2025）和四个故障分类基准测试（IIEE、IICA、CWRU、MAFAULDA）的结果。仅在 AS2M 上进行预训练的模型（BEATs、CED）显示出有限的任务间泛化能力，平均 DCASE 分数为 0.593 – 0.619。大盛，在更广泛的混合数据上进行了训练，缩小了工业数据集上的差距，但在 DCASE 任务中仍低于顶级表现者。相比之下，FISHER 和我们提出的 ECHO，都在 AS2M+MTG 上进行了带分割建模的训练，对机器信号提供了持续强劲的表现。

FISHER 是机器故障诊断中最强大的基准：其小变体在故障套件中达到了最佳平均值（0.955），并在 MAFAULDA 振动准确性上接近完美。然而，这在 DCASE 上带来了权衡，其中 FISHER-Small 的平均分为 0.610。我们的 ECHO 在保持故障诊断准确度的同时实现了更优的 DCASE 性能：ECHO-Small 达到了最高的 DCASE 平均值 0.621（比 FISHER-Small 高 +0.011），并维持了接近的故障平均值 0.954 (-0.001)。因此，ECHO-Small 在所有任务中提供了最佳的整体平均分，为 0.772（相比 FISHER-Small 提高了 +0.006）。这些结果表明，我们的策略利用次带上的频率位置嵌入使模型能够学习到更好的通用机器信号表示。

总体而言，ECHO 提供了更好的跨域折衷：它推进了声学场景分析的最新泛化能力，而没有牺牲工业故障检测的能力。这种平衡在评估模型中产生了最高的整体平均值，并表明 ECHO 适合于必须在一个统一系统内同时处理环境声学和机器状态监测的真实世界部署。

5. 结论

本文提出了一种名为 ECHO 的基础模型，旨在处理具有不同采样率和长度的信号。我们在新引入的一个开源基准 SIREN 上评估了我们的模型，该基准包括异常检测和细粒度异常分类任务，并涵盖了声学和振动等多种机器模态。此外，SIREN 设计为可扩展，在未来将纳入更多的故障检测数据集和模态。在 SIREN 上的实验结果表明 ECHO 的优越性，它在异常检测中达到了最先进的性能，并且在异常分类中的表现与现

有的领先方法相当。值得注意的是，ECHO 在 DCASE 样式的异常检测任务中优于其他模型，突显了其在不同类型任务之间平衡和强大的能力。

6. 未来工作

在未来的工作中，我们计划通过训练和发布基于不同 ViT 架构（如 ViT-tiny、ViT-mini、ViT-base 和 ViT-large）的额外变体来进一步扩大 ECHO 模型的规模，以提供更广泛的有效性与准确性权衡范围。我们还将继续扩展 SIREN 基准测试，通过纳入更多跨多种模态的真实世界机器故障诊断数据集，支持对异常检测和分类方法进行更加全面和通用化的评估。

7. REFERENCES

[1] Luca Radicioni, Francesco Morgan Bono, and Simone Cinquemani, "Vibration-based anomaly detection in industrial machines: A comparison of autoencoders and latent spaces," Machines, vol. 13, no. 2, pp. 139, 2025.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in International Conference on Learning Representations, 2021.

[3] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer, "Masked autoencoders that listen," in Proc. of NIPS, 2022, vol. 35, pp. 28708–28720.

[4] Khaled Koutini and Jan Schlüter and Hamid Eghbal-zadeh and Gerhard Widmer, "Efficient Training of Audio Transformers with Patchout," in Proc. of Interspeech, 2022, pp. 2753–2757.

[5] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei, "BEATs: Audio pre-training with acoustic tokenizers," in Proc. of ICML, 2023, vol. 202, pp. 5178–5193.

[6] Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Junbo Zhang, and Yujun Wang, "Ced: Consistent ensemble distillation for audio tagging," in Proc. of ICASSP. IEEE, 2024, pp. 291–295.

[7] Wenxi Chen, Yuzhe Liang, Ziyang Ma, Zhisheng Zheng, and Xie Chen, "Eat: Self-supervised pre-training with efficient audio transformer," in Proc. of IJCAI, 2024, pp. 3807–3815.

[8] Heinrich Dinkel, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Yujun Wang, and Bin Wang, "Scaling up masked audio encoder learning for general audio classification," arXiv preprint arXiv:2406.06992, 2024.

[9] Bing Han, Zhiqiang Lv, Anbai Jiang, Wen Huang, Zhengyang Chen, Yufeng Deng, Jiawei Ding, Cheng Lu, Wei-Qiang Zhang, Pingyi Fan, et al., "Exploring large scale pre-trained models for robust machine anomalous sound detection," in Proc. of ICASSP. IEEE, 2024, pp. 1326–1330.

[10] Anbai Jiang, Bing Han, Zhiqiang Lv, Yufeng Deng, Wei-Qiang Zhang, Xie Chen, Yanmin Qian, Jia Liu, and Pingyi Fan, "Anopatch: Towards better consistency in machine anomalous sound detection," in Proc. of Interspeech, 2024, pp. 107–111.

[11] Xinhu Zheng, Anbai Jiang, Bing Han, Yanmin Qian, Pingyi Fan, Jia Liu, and Wei-Qiang Zhang, "Improving anomalous sound detection via low-rank adaptation fine-tuning of pre-trained audio models," in Proc. of SLT. IEEE, 2024, pp. 969–974.

[12] Pingyi Fan, Anbai Jiang, Shuwei Zhang, Zhiqiang Lv, Bing Han, Xinhu Zheng, Wenrui Liang, Junjie Li, Wei-Qiang Zhang, Yanmin Qian, et al., "Fisher: A foundation model for multi-modal industrial signal comprehensive representation," arXiv preprint arXiv:2507.16696, 2025.

[13] Yuma Koizumi, Yohei Kawaguchi, Keisuke Imoto, Toshiki Nakamura, Yuki Nikaido, Ryo Tanabe, Harsh Purohit, Kaori Suefusa, Takashi Endo, Masahiro Yasuda, et al., "Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," arXiv preprint arXiv:2006.05822, 2020.

[14] Yohei Kawaguchi, Keisuke Imoto, Yuma Koizumi, Noboru Harada, Daisuke Niizumi, Kota Dohi, Ryo Tanabe, Harsh Purohit, and Takashi Endo, "Description and discussion on dcase 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions," arXiv preprint arXiv:2106.04492, 2021.

[15] Kota Dohi, Keisuke Imoto, Noboru Harada, Daisuke Niizumi, Yuma Koizumi, Tomoya Nishida, Harsh Purohit, Takashi Endo, Masaaki Yamamoto, and Yohei Kawaguchi, "Description and discussion on dcase 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," arXiv preprint arXiv:2206.05876, 2022.

[16] Kota Dohi, Keisuke Imoto, Noboru Harada, Daisuke Niizumi, Yuma Koizumi, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, and Yohei Kawaguchi, "Description and discussion on dcase 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," arXiv preprint arXiv:2305.07828, 2023.

[17] Tomoya Nishida, Noboru Harada, Daisuke Niizumi, Davide Albertini, Roberto Sannino, Simone Pradolini, Filippo Augusti, Keisuke Imoto, Kota Dohi, Harsh Purohit, Takashi Endo, and Yohei Kawaguchi, "Description and discussion on dcase 2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," arXiv preprint arXiv:2406.07250, 2024.

[18] Tomoya Nishida, Noboru Harada, Daisuke Niizumi, Davide Albertini, Roberto Sannino, Simone Pradolini, Filippo Augusti, Keisuke Imoto, Kota Dohi, Harsh Purohit, et al., "Description and discussion on dcase 2025 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," arXiv preprint arXiv:2506.10097, 2025.

[19] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in Proc. of ICASSP. IEEE, 2017, pp. 776–780.

[20] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra, "The mtg-jamendo dataset for automatic music tagging," in Proc. of ICML, 2019.

[21] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman, "Vggsound: A large-scale audio-visual dataset," in Proc. of ICASSP. IEEE, 2020, pp. 721–725.

[22] Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale

Song, “Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning,” in Proc. of ICCV, 2021, pp. 10274–10284.