ECHO: FREQUENCY-AWARE HIERARCHICAL ENCODING FOR VARIABLE-LENGTH SIGNALS

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

[1]School of Computer Science, Wuhan University, Wuhan, China
[2]School of Artificial Intelligence, Wuhan University, Wuhan, China
[3]Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems,
Digital Innovation Research Center, Duke Kunshan University, Suzhou, China
{yucong.zhang,liujuan}@whu.edu.cn, ming.li369@dukekunshan.edu.cn

ABSTRACT

Pre-trained foundation models have demonstrated remarkable success in audio, vision and language, yet their potential for general machine signal modeling with arbitrary sampling rates—covering acoustic, vibration, and other industrial sensor data—remains under-explored. In this work, we propose a novel foundation model ECHO that integrates an advanced band-split architecture with frequency positional embeddings, enabling spectral localization across arbitrary sampling configurations. Moreover, the model incorporates sliding patches to support inputs of variable length without padding or cropping, producing a concise embedding that retains both temporal and spectral fidelity and naturally extends to streaming scenarios. We evaluate our method on various kinds of machine signal datasets, including previous DCASE task 2 challenges (2020–2025), and widely-used industrial signal corpora. Experimental results demonstrate consistent state-of-the-art performance in machine signal anomaly detection and fault classification, confirming the effectiveness and generalization capability of the proposed model. We open-sourced ECHO on `https://github.com/yucongzh/ECHO`.

Index Terms— Anomalous sound detection, pre-trained model, foundation model, frequency-aware, band-splitting

1. INTRODUCTION

The reliable monitoring of machine health is critical for ensuring safety, reducing downtime, and optimizing operational efficiency across industrial domains. In recent years, machine signal analysis—encompassing acoustic emissions, vibration measurements, and other sensor modalities—has emerged as a central tool for detecting anomalous conditions and diagnosing faults before catastrophic failure. Traditional approaches, such as handcrafted feature extraction combined with conventional classifiers, have proven effective in narrow, domain-specific scenarios [1]. However, these methods often lack generalization capability across heterogeneous machine types, operating conditions, and sensing modalities.

Large-scale audio foundation models have emerged as a unifying paradigm for representation learning across diverse acoustic domains. Leveraging supervised [2] or self-supervised [3–7] Vision Transformer (ViT)-based [8] pre-training on large corpora, these models demonstrate strong transferability to downstream tasks ranging from tagging to captioning. Scaling training data further improves robustness across speech and audio tasks [9]. This success extends to industrial monitoring, where such models serve as front-end encoders for anomalous sound detection (ASD) [10–12] and play key roles in recent DCASE Task 2 challenges. In domains with scarce labeled anomalies and varying conditions, the generalizable representations of audio foundation models provide a solid basis for domain adaptation.

Despite their strengths, existing foundation models face two major challenges in real-world machine signal monitoring tasks. First, current ViT-based pre-trained models rely on fixed-size spectrogram input for patching, and adopt conventional 2D positional embeddings from image processing to learn 2D spatial relation among those patches. Modeling variable-length spectrogram thus requires truncation or interpolation, breaking the spatial relation between patches. We argue that this kind of spatial modeling is not ideal for audio which is temporally sequential by nature. Second, these models are trained on samples with a fixed sampling rate and can only infer at that rate. Inputs with higher or lower rates must be resampled, which inevitably introduces information loss.

In this work, we address these limitations by proposing **ECHO**, an audio foundation model that uses fr**E**quen**C**y-aware **H**ierarchical enc**O**ding for variable-length signals. ECHO is a general-purpose foundation model for machine signals trained on large-scale audio corpus, achieving state-of-the-art performance across various benchmarks, includ-

†Corresponding authors: Juan Liu and Ming Li.

This research is funded in part by the National Natural Science Foundation of China (62571223), Science and Technology Program of Suzhou City (SYC2022051) and Guangdong Science and Technology Plan (2023A1111120012). Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

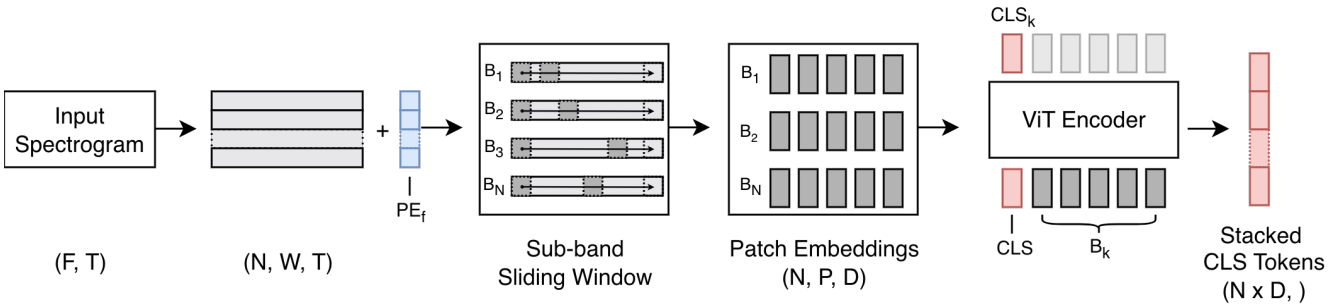

Fig. 1. Feature extraction pipeline of the ECHO framework. F: number of frequency bins after STFT; T: number of time frames after STFT; N: number of sub-bands after band splitting; W: band width of sub-bands; P: number of sliding patches; D: feature dimension for each patch.

ing few-shot anomaly sound detection tasks at DCASE, vibration-based fault detection datasets, as well as multi-modal machine condition monitoring datasets. The main contributions of this work are summarized as follows:

i. frequency-aware band-splitting strategy: splitting spectrogram into frequency sub-bands with a relative frequency positional embedding mechanism tailored for arbitrary sampling rates and frequency resolutions, enabling the model to encode explicit positional context of sub-bands within the full spectrum;

ii. a sliding patch design within each sub-band that suits variable-length signal inputs;

iii. a scalable training framework capable of handling diverse machine signal modalities within a unified representation space; and

iv. achieving state-of-the-art performance on an open-sourced benchmark SIREN for general machine signal embeddings evaluation.

2. RELATED WORK

Concurrent with our study, the FISHER model [13] introduced band-splitting to handle multi-modal signals with varying sampling rates, where each sub-band is modeled independently using a ViT backbone. Although both works share a similar high-level motivation, they were developed independently and adopt substantially different designs, as reflected in the released code bases[1][2]. In contrast to FISHER, our approach incorporates frequency positional encoding within each sub-band before feeding them into the ViT, enabling explicit frequency-aware modeling. Moreover, instead of patch-based image-style tokenization—which is less suitable for variable-length or streaming inputs—our method employs a sliding-window strategy within each sub-band. Conceptually, our framework is designed to simultaneously support variable-length and variable-sampling-rate signals, making it naturally extendable to streaming scenarios.

[1]Our implementation: `https://github.com/yucongzh/ECHO`
[2]FISHER code: `https://github.com/jianganbai/FISHER`

3. MODEL ARCHITECTURE

In this section, we introduce our proposed framework **ECHO** which is designed for robust representation learning of variable-length signals under arbitrary sampling rates. The overall architecture is illustrated in Fig. 1. ECHO consists of four key components: (1) spectrogram extraction, (2) frequency-aware sub-band splitting, (3) temporal sliding patches extraction, and (4) hierarchical encoding.

3.1. Spectrogram Extraction

Given an input waveform sampled at frequency f_s, we compute its Short-Time Fourier Transform (STFT) using a predefined window length t_{win} and hop length t_{hop} specified in seconds. We use the magnitude spectrogram. Because these time durations are converted per signal to integer samples, the spectrogram frame rate is fixed by the chosen hop and thus independent of f_s. Consequently, for inputs of equal duration, the resulting spectrograms contain the same number of time frames across sampling rates.

3.2. Frequency-Aware Sub-band Splitting with Positional Encoding

The spectrogram S is uniformly split along the frequency axis into a set of sub-bands with no overlaps, with the number of sub-bands proportional to the sampling rate f_s. For the k-th sub-band spanning b_{start} to $b_{\text{end}} - 1$, the center frequency f_c, its normalized position p, and the corresponding positional encoding $\text{PE}(p, j)$ are computed as

$$f_c = \frac{(b_{\text{start}} + b_{\text{end}} - 1)}{2} \cdot \frac{f_s}{N_{\text{FFT}}}, \quad p = \frac{f_c}{f_s/2},$$
$$\text{PE}(p, j) = \begin{cases} \sin\left(\frac{\gamma \cdot p}{10000^{2i/d}}\right), & j = 2i, \\ \cos\left(\frac{\gamma \cdot p}{10000^{2i/d}}\right), & j = 2i + 1, \end{cases} \tag{1}$$

where N_{FFT} is calculated as $f_s \times t_w$, d is the embedding dimension and γ is the scaling factor. This design ensures that sub-bands from different sampling rates, but at equivalent relative frequency positions, share consistent positional encodings.

3.3. Temporal Sliding Patch Extraction

To model signals of variable duration, each sub-band undergoes temporal segmentation. Specifically, we apply a sliding

Table 1. SIREN Benchmark datasets and their characteristics. SR: Sample rate; MAFAULDA: Machinery Fault Database; CWRU: CWRU Bearing dataset; IIEE: IDMT Electric Engine dataset; IICA: IDMT Compressed Air dataset.

Dataset	Modality	SR	#Classes	Split	Scoring
DCASE2020 [14]	Sound	16k	2	official	KNN
DCASE2021 [15]	Sound	16k	2	official	KNN
DCASE2022 [16]	Sound	16k	2	official	KNN
DCASE2023 [17]	Sound	16k	2	official	KNN
DCASE2024 [18]	Sound	16k	2	official	KNN
DCASE2025 [19]	Sound	16k	2	official	KNN
MAFAULDA [20]	Sound/Vibration	50k	10	LOOCV	KNN
CWRU [21]	Vibration	12k	10	LOOCV	KNN
IIEE [22]	Sound	44.1k	3	official	KNN
IICA [23]	Sound	48k	3	5-Fold-CV	KNN

window of length L (equal to the sub-band width) along the time axis, with a stride of $L/2$ to achieve 50% overlap ping.

This operation is efficiently implemented via a two-dimensional convolution with kernel size $(\text{sub-band height}, L)$ and stride $(\text{sub-band height}, L/2)$. The convolution collapses the frequency dimension, resulting in a patch sequence with shape (N, D), where N is the number of temporal patches and D is the channel dimension (i.e., the patch embedding size). Each patch thus represents a localized temporal feature of the sub-band.

3.4. Hierarchical Encoding

Each frequency-aware patch sequence, prepended with a learnable classification (CLS) token, is fed to the ViT backbone. The CLS token summarizes sub-band information, and the final embedding concatenates all sub-band CLS tokens. This hierarchical design enables ECHO to capture local temporal dependencies within sub-bands while distinguishing frequency ranges via frequency-aware splitting.

3.5. Training and Inference

We adopt a teacher–student framework from EAT [7]. During training, each frequency-aware sub-band is treated independently. The student receives masked inputs, while the teacher is updated via Exponential Moving Average (EMA) of the student: $\theta_{\text{teacher},t} = \alpha\theta_{\text{teacher},t-1} + (1-\alpha)\theta_{\text{student},t}$, with momentum α. We employ two self-supervised objectives: (1) global alignment between the temporal mean of the teacher's layer outputs and student CLS token, and (2) frame-level alignment on masked positions. This dual-level supervision enforces consistency at both coarse and fine scales.

During inference, the full spectrogram is processed by ECHO. CLS tokens from all K sub-bands are concatenated into a hierarchical embedding $\mathbf{z} = [\text{CLS}_1, \ldots, \text{CLS}_K]$ for downstream tasks.

4. BENCHMARK FOR EVALUATION

For fair comparison, we open-source an evaluation benchmark called SIREN (**SI**gnal **R**epresentation **E**valuatio**N** toolkit). SIREN is tailored for general signal diagnosis, including tasks like few-shot anomalous detection (DCASE Task 2 series), and machine fault diagnosis/classification. The detailed information of the datasets and the corresponding tasks are shown in Table 1 and the GitHub repository[3]. The evaluation protocol and metrics are as follows.

DCASE Task 2 series: We follow the official development/evaluation splits [14–19], computing file-level anomaly scores, and aggregate them with machine/section/domain grouping. Per year, we report ROC-AUC and partial AUC (pAUC), and summarize performance by harmonic means of AUC and pAUC across development and evaluation sets.

Fault classification: For datasets (MAFAULDA [20], CWRU [21], IICA [23]) without official train-test split, we use cross validation (CV) for evaluation. On MAFAULDA and CWRU, we use leave-one-out CV (LOOCV) due to limited data for each fault class; on IICA, we use 5-fold CV. For IIEE [22], we use the given train-test split for evaluation. These settings (see Table 1) balance reliability and computational cost. Accuracy is reported for each dataset.

Scoring. We use k-nearest neighbors (k-NN): (1) compute training embeddings $\{e_i\}_{i=1}^{N}$ to build a memory bank $\mathcal{M}$; (2) for a test embedding e^*, retrieve its k nearest neighbors in $\mathcal{M}$. For DCASE anomaly detection, the score is the nearest-neighbor distance (k=1). For fault classification, the label is the majority vote of the k neighbors.

5. EXPERIMENTS

5.1. Implementation Details

We adopt the same backbone architecture as the ViT. Currently, we have released two editions of ECHO: small and tiny. The spectrogram is extracted using a window size of 25 ms and window shift of 10 ms on normalized raw signals. The sub-band width in our model is fixed to 32. The model is trained for a total of 400,000 steps using four NVIDIA GeForce RTX 3090 GPUs with a global batch size of 256, with 4,000 warm-up steps. The learning rate is scaled relative to the effective batch size [3], with a base learning rate of 10^{-4}. Training employs a cosine learning rate scheduler with a linear warm-up phase spanning 40,000 steps. The minimum learning rate is set to 10^{-5}, and weight decay is applied with a coefficient of 0.05.

5.2. Baseline Models

Currently, we include 5 pre-trained foundation models for comparison: BEATs [5], CED [6], EAT [7], Dasheng [9], and FISHER [13]. All models use ViT-style structure as the backbone model, and are trained using open-sourced audio datasets across various domains, including full AudioSet (AS2M) [25], MTG-Jamendo (MTG) [26], VGGSound (VGG) [27], Music4all (M4A) [28], ACAV [29], and Freesound (FS)[4].

5.3. Experimental Results

Table 2 reports the performance on the SIREN benchmark. Several observations can be made:

[3]Codes available at `https://github.com/yucongzh/SIREN`

[4]`https://freesound.org/`

Table 2. Performance (%) comparison of pre-trained foundation models across DCASE challenges and machine fault diagnosis datasets in our SIREN benchmark (k=5). DCASE tasks are reported by the (harmonic) means of AUCs and partial AUCs among machines, while fault classification tasks are reported by accuracy. DCASE tasks are evaluated according to the officials. "Mean" stands for arithmetic mean. Sample rates are listed under dataset names. FS*: Freesound derived from WavCaps [24].

Model	Datasets	Scale	#Param.	DCASE Tasks							Fault Classification Tasks					Mean
				2020	2021	2022	2023	2024	2025	Mean	IIEE	IICA	CWRU	MAFAULDA	Mean	
				16k	16k	16k	16k	16k	16k	-	44.1k	48k	12k	50k	-	
BEATs [5]	AS2M	Base	90M	**74.26**	**61.31**	58.97	62.89	55.89	57.84	61.86	65.81	91.55	88.57	99.69/63.66	81.86	71.86
CED [6]	AS2M	Base	86M	67.75	56.67	57.26	60.84	57.83	57.72	59.68	80.21	86.08	81.90	99.74/66.48	82.88	71.28
		Small	22M	67.69	56.66	56.79	60.04	57.24	57.89	59.39	74.42	85.66	85.71	99.59/64.43	81.96	70.67
		Mini	10M	67.59	56.35	57.03	59.85	56.39	57.43	59.11	74.17	84.91	82.86	99.64/63.66	81.05	70.08
		Tiny	5.5M	67.21	56.17	56.77	59.61	56.19	57.82	58.96	72.74	84.02	82.86	99.64/63.51	80.55	69.76
EAT [7]	AS2M	Large	0.3B	73.94	57.47	58.54	61.66	**57.89**	**60.01**	61.58	68.33	89.96	**91.43**	99.33/85.03	86.82	74.20
		Base	86M	72.13	57.79	58.57	59.69	57.12	59.75	60.84	78.97	89.01	85.71	99.90/84.52	87.62	74.23
Dasheng [9]	AS2M+MTG	1.2B	1.2B	69.48	57.06	57.29	61.23	57.13	57.12	59.88	96.09	91.91	90.48	99.69/77.24	91.08	75.48
	+VGG	0.6B	0.6B	68.18	56.76	56.61	59.94	56.75	56.73	59.16	99.11	92.11	89.52	99.74/78.22	91.74	75.45
	+ACAV	Base	86M	69.15	57.27	57.87	60.70	57.71	57.01	59.95	99.36	90.88	88.57	99.85/81.96	92.12	76.04
FISHER [13]	AS2M+MTG +M4A+FS	Small	22M	70.54	59.51	59.79	61.83	55.66	58.68	61.00	97.48	94.20	86.67	100.0/85.29	92.73	76.86
		Mini	10M	69.98	58.39	57.91	60.35	55.91	56.90	59.91	99.90	94.50	73.33	100.0/87.75	91.10	75.50
		Tiny	5.5M	70.64	58.51	57.11	58.46	55.34	57.69	59.62	99.80	**95.43**	75.24	**100.0/88.52**	91.80	75.71
ECHO	AS2M+MTG +FS*	Small	22M	72.23	60.20	**59.96**	63.71	57.86	58.70	**62.11**	99.85	93.67	90.48	99.54/82.42	**93.19**	**77.65**
		Tiny	5.5M	70.14	59.01	59.76	**63.75**	56.91	58.40	61.33	**100.0**	93.58	90.48	99.85/78.83	92.55	76.94

1) Effect of dataset scaling. By comparing BEATs (71.86%), CED (71.28%) and EAT (74.23%) with Dasheng (76.04%) with base scale, we observe that incorporating additional large-scale datasets might help mitigate the domain mismatch between general audio pre-training and machine sound analysis. A similar trend can also be found in FISHER (76.86%) and our proposed ECHO (77.65%), both trained on additional audio datasets, showing that scale-up of training data consistently enhances cross-domain representation learning.

2) Sliding-patch vs. conventional patch modeling. Traditional foundation models (BEATs, CED, and EAT) rely on conventional patch tokenization, yielding total average scores in the range of 70–74% in our SIREN benchmark. In contrast, Dasheng introduces a sliding-patch strategy, leading to a higher overall mean of 76.04%, outperforming EAT (74.23%) by $+1.81\%$. A similar observation is confirmed within the band-splitting family: our proposed ECHO reaches a total average of 77.65% on SIREN, surpassing FISHER (76.86%) by $+0.79\%$. These results highlight that sliding-patch modeling is more effective than fixed patch partitioning for machine sound analysis.

3) Band-splitting architecture. Compared with conventional pre-trained foundation models, band-splitting-based methods achieve superior performance in fault classification across multiple modalities and sampling conditions, while maintaining competitive results on DCASE tasks. As shown in Table 2, ECHO (93.19%) and FISHER (92.73%) rank first and second on fault classification tasks, with model with small scale. This confirms that decomposing signals into frequency sub-bands facilitates robust cross-sampling-rate anomaly detection. Furthermore, ECHO consistently outperforms FISHER, which we attribute to the integration of frequency positional encoding, enabling better frequency-aware modeling across sub-bands. It is noteworthy that Dasheng (92.12%), trained only on samples with fixed sampling rate, also exhibits strong adaptability to diverse fault classification tasks, a capability that may stem from its use of large-scale training data.

4) Effect of model scaling. Comparing ECHO-Small (77.65%) and ECHO-Tiny (76.94%) shows that enlarging the model scale yields consistent improvements across both DCASE (62.11% vs. 61.33%) and fault classification tasks (93.19% vs. 92.55%). This indicates that our architecture retains scalability potential, and larger variants may further boost generalization.

Overall, ECHO achieves the highest overall performance (77.65%) in our SIREN benchmark, outperforming the strong baseline (FISHER, 76.86%) by $+0.79\%$. These results show that combining frequency-aware band-splitting with sliding-patch modeling is effective in our principled framework for cross-domain machine signal representation learning.

6. CONCLUSION

In this article, we proposed ECHO, a novel pre-trained foundation model that incorporates a frequency-aware band-splitting strategy to handle diverse sampling rates and uses sliding patches to accommodate inputs of varying lengths. The model was evaluated on the newly introduced open-source benchmark SIREN, which includes tasks such as anomaly detection and fine-grained anomaly classification across multiple machine modalities, including acoustics and vibrations. Experimental results demonstrate that ECHO achieves state-of-the-art performance in both anomaly detection and fault classification, highlighting its potential for a wide range of industrial anomaly detection applications.

7. REFERENCES

[1] P. Yan, A. Abdulkadir, P.-P. Luley, M. Rosenthal, G. A. Schatte, B. F. Grewe, and T. Stadelmann, "A comprehensive survey of deep transfer learning for anomaly detection in industrial time series: Methods, applications, and directions," *IEEE Access*, vol. 12, pp. 3768–3789, 2024.

[2] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. ASLP*, vol. 28, pp. 2880–2894, 2020.

[3] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba *et al.*, "Masked autoencoders that listen," in *Proc. NeurIPS*, vol. 35, 2022, pp. 28 708–28 720.

[4] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Proc. Interspeech*, 2022, pp. 2753–2757.

[5] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen *et al.*, "BEATs: Audio pre-training with acoustic tokenizers," in *Proc. ICML*, 2023, pp. 5178–5193.

[6] H. Dinkel, Y. Wang, Z. Yan, J. Zhang, and Y. Wang, "CED: Consistent ensemble distillation for audio tagging," in *Proc. ICASSP*, 2024, pp. 291–295.

[7] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "EAT: Self-supervised pre-training with efficient audio transformer," in *Proc. IJCAI*, 2024, pp. 3807–3815.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.

[9] H. Dinkel, Z. Yan, Y. Wang, J. Zhang, Y. Wang, and B. Wang, "Scaling up masked audio encoder learning for general audio classification," in *Proc. Interspeech*, 2024, pp. 547–551.

[10] A. Jiang, B. Han, Z. Lv, Y. Deng, W.-Q. Zhang, X. Chen *et al.*, "Anopatch: Towards better consistency in machine anomalous sound detection," in *Proc. Interspeech*, 2024, pp. 107–111.

[11] X. Zheng, A. Jiang, B. Han, Y. Qian, P. Fan, J. Liu, and W.-Q. Zhang, "Improving anomalous sound detection via low-rank adaptation fine-tuning of pre-trained audio models," in *Proc. SLT*, 2024, pp. 969–974.

[12] B. Han, A. Jiang, X. Zheng, W.-Q. Zhang, J. Liu, P. Fan, and Y. Qian, "Exploring self-supervised audio models for generalized anomalous sound detection," *IEEE Trans. ASLP*, 2025.

[13] P. Fan, A. Jiang, S. Zhang, Z. Lv, B. Han, X. Zheng *et al.*, "FISHER: A foundation model for multi-modal industrial signal comprehensive representation," *arXiv preprint arXiv:2507.16696*, 2025.

[14] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe *et al.*, "Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE*, 2020, pp. 81–85.

[15] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi *et al.*, "Description and discussion on dcase 2021 challenge task 2: Unsupervised anomalous detection for machine condition monitoring under domain shifted conditions," in *Proc. DCASE*, 2021, pp. 186–190.

[16] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida *et al.*, "Description and discussion on dcase 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," in *Proc. DCASE*, 2022.

[17] ——, "Description and discussion on DCASE 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE*, 2023, pp. 31–35.

[18] T. Nishida, N. Harada, D. Niizumi, D. Albertini, R. Sannino, S. Pradolini *et al.*, "Description and discussion on dcase 2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE*, 2024, pp. 111–115.

[19] ——, "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.

[20] Signals, Multimedia and Telecommunications Laboratory (SMT), "MAFAULDA: Machinery fault database," SMT Lab. [Online]. Available: http://www02.smt.ufrj.br/~offshore/mfs/page_01.html.

[21] Case Western Reserve University, "Bearing data center: Seeded fault test data," Case School of Engineering. [Online]. Available: https://engineering.case.edu/bearingdatacenter.

[22] S. Grollmisch, J. Abeßer, J. Liebetrau, and H. Lukashevich, "Sounding industry: Challenges and datasets for industrial sound analysis," in *Proc. EUSIPCO*, 2019, pp. 1–5.

[23] D. Johnson, J. Kirner, S. Grollmisch, and J. Liebetrau, "Compressed air leakage detection using acoustic emissions with neural networks," in *Proc. Inter-Noise*, Seoul, South Korea, 2020, pp. 5662–5673.

[24] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao *et al.*, "Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *IEEE/ACM Trans. ASLP*, vol. 32, pp. 3339–3354, 2024.

[25] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*, 2017, pp. 776–780.

[26] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The mtg-jamendo dataset for automatic music tagging," in *Proc. ICML*, 2019.

[27] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *Proc. ICASSP*, 2020, pp. 721–725.

[28] I. A. P. Santana, F. Pinhelli, J. Donini, L. Catharin, R. B. Mangolin, V. D. Feltrim *et al.*, "Music4all: A new music database and its applications," in *Proc. IWSSIP*, 2020, pp. 399–404.

[29] S. Lee, J. Chung, Y. Yu, G. Kim, T. Breuel, G. Chechik, and Y. Song, "Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning," in *Proc. ICCV*, 2021, pp. 10 274–10 284.