FIRST-SHOT ANOMALOUS SOUND DETECTION WITH FROZEN GENERAL AUDIO ENCODERS AND DISTANCE-BASED BACK-ENDS

Technical Report

Rene Glitza, Luca Becker and Rainer Martin

Ruhr-Universität Bochum
Institute of Communication Acoustics
Unisersitätsstr. 150, 44780 Bochum, Germany
{*firstname.lastname*}@rub.de

ABSTRACT

We present a training-free approach to first-shot unsupervised anomalous sound detection in the noise-aware, two-channel setting of DCASE 2026 Task 2. Instead of training or fine-tuning on the challenge data, we pass each recording through a frozen, general-purpose audio encoder, fit a simple reference model on the few normal clips available for each machine, and score a test clip by how far it sits from that reference, so adapting to a new machine or domain only means replacing the reference set. We vary two choices: the embedding source, where we put the industrial-signal foundation model FISHER against an Audio Spectrogram Transformer (AST) that we self-pre-train on AudioSet, and the scoring back-end, a non-parametric k-nearest-neighbor (kNN) memory bank against a per-machine Gaussian mixture model (GMM), both with domain z-score calibration. On the development set the frozen FISHER encoder with a kNN back-end reaches the best official score ($\Omega = 58.23\%$) and is the only configuration to beat both the autoencoder and selective-Mahalanobis baselines, which shows that frozen embeddings with a simple estimator can compete with task-specific reconstruction models in the first-shot regime.

***Index Terms*—** anomalous sound detection, first-shot unsupervised learning, domain generalization, frozen audio encoders, foundation models, kNN, GMM

1. INTRODUCTION

Anomalous sound detection (ASD) for machine condition monitoring decides whether the sound of a target machine is normal or anomalous, a key building block for predictive maintenance in automated production. DCASE 2026 Task 2 [1] continues the *first-shot* line of the 2023–2025 tasks [2, 3], now in a *noise-aware* setting: systems must work on two-channel recordings, generalize across domain shifts such as changes in operating condition or background noise, and run on unseen machine types without any machine-specific tuning. Only normal sounds and a small amount of reference data are available for training, which makes the problem unsupervised, domain-generalized, and data-scarce.

Most strong systems obtain their discriminative power by training or fine-tuning an encoder on the task data, often with outlier exposure or attribute classification [1, 4]. This ties detection quality to a per-machine training stage that needs enough normal data and has to be repeated whenever a new machine or domain appears, which is at odds with the first-shot goal of plug-and-play deployment. We therefore avoid any fine-tuning on the challenge data. Recent work on domain-generalized ASD [5], on training-free detection [6], and on foundation models for anomaly detection [7] all suggests that generic, *frozen* embeddings from large pre-trained encoders already separate normal from anomalous machine sounds. Our pipeline extracts such embeddings, fits a simple reference model on a few normal clips per machine, and scores a test clip by its distance to that reference; moving to a new machine or domain only means swapping the reference set.

We study three design choices within this paradigm. The first is the encoder: an industrial-signal foundation model trained on heterogeneous machine signals (FISHER [8]) against a general-purpose transformer that we self-pre-train on AudioSet [9] with a multi-label AdaCos [10] objective and masked-patch reconstruction, built on the AST architecture [11]; neither encoder sees DCASE Task 2 data. The second is the back-end: a non-parametric k-nearest-neighbor memory bank against a parametric Gaussian mixture model (GMM). The third is how to use the two channels, given that the near microphone is dominated by the machine and the far one by factory noise; we treat the channel layout as a front-end choice and feed the encoder either a stacked dual-channel input or a single time-concatenated channel.

Crossing the two encoders with the two back-ends gives the four systems we submit, none of them fine-tuned on the challenge data. The rest of the report covers the method (Section 3), the setup (Section 4), and results on the development set (Section 5).

2. RELATION TO PRIOR WORK

Unsupervised ASD for machine condition monitoring has been shaped by the DCASE Task 2 series, which since 2023 has used a first-shot, domain-generalized formulation: a system must handle a new machine type from a single section of normal data, without machine-specific tuning or score ensembling [2, 12, 3]. The official baseline scores anomalies either from autoencoder reconstruction error or from a selective Mahalanobis distance to per-machine statistics [13], and the 2026 edition adds the noise-aware, two-channel scenario [1]. The strongest entries learn discriminative embeddings with auxiliary machine-identification or attribute objectives, often with outlier exposure [14]; as later edi-

This work has been supported by the German Federal Ministry of Research, Technology and Space (BMFTR) (grant 02L19C200, "HUMAINE") and by the German Research Foundation (DFG) (project numbers 429873205 and 549576906).

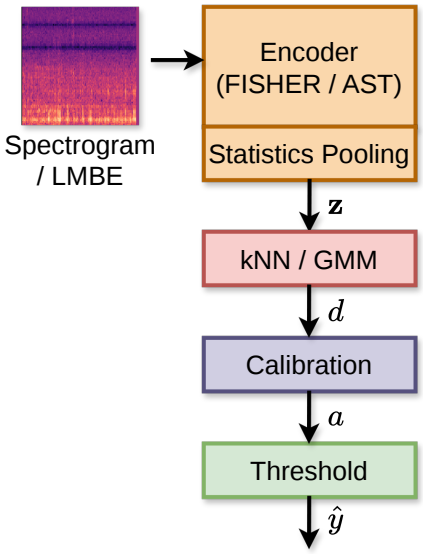

Figure 1: Proposed system: encoder with pooling, distance model (kNN or GMM), calibration, and thresholding.

tions hide attribute metadata, recent work recovers supervision through pseudo-labeling, vector quantization, or domain-adapted self-supervision [15, 16], and noise-aware systems prepend a target-enhancement front-end [17]. All of these tie detection quality to a per-task training stage and must be retrained for each new machine or domain.

A complementary line, which we follow, scores anomalies from *frozen* embeddings and skips task-specific fine-tuning. GenRep shows that frozen features from pre-trained audio encoders, with a nearest-neighbor back-end and score normalization, match or beat fine-tuned systems on first-shot ASD [5], and tailored pooling on pre-trained backbones helps further [4]. Dedicated industrial-signal foundation models such as FISHER [8] and general audio transformers such as AST [11] provide strong reusable encoders, and the wider use of foundation models for anomaly detection is surveyed in [7]. Training-free detectors take this furthest by comparing a query against a small support set in a frozen space and adapting by swapping that set rather than retraining [6, 18]. Our systems combine two frozen encoders, FISHER and a self-pre-trained AST, with two reference-based back-ends, kNN and GMM, in the noise-aware, two-channel DCASE 2026 setting without any fine-tuning.

3. PROPOSED METHOD

3.1. Pipeline Overview

We never optimize any parameter on the DCASE 2026 Task 2 data. A frozen, general-purpose encoder maps each recording to a fixed embedding, and scoring is done by a distance- or density-based back-end fitted only on the per-machine normal training clips (Figure 1):

$$x \xrightarrow{\text{view}} v \xrightarrow{\text{feature}} S \xrightarrow{\text{encoder}} T \xrightarrow{\text{pool}} z \xrightarrow{\text{back-end}} a(x), \quad (1)$$

where x is the two-channel waveform, v a derived view, S a feature (spectrogram or log-mel band energies), $T \in \mathbb{R}^{N\times M}$ the frozen token sequence, z the pooled embedding, and $a(x)$ the anomaly score. Every system shares this graph and runs with the encoder frozen (`max_epochs` $= 0$, a single forward pass). Rather than reduce the stereo signal to one channel, we keep two views: `Dual-Channel` stacks the channels into a $[B\times 2\times T]$ tensor, and `Concat` joins them along time into a $[B\times 1\times 2T]$ tensor. Each view becomes either a spectrogram (n_fft $= 1024$, window 400, hop 160) or log-mel band energy (LMBE, 128 mel bands, n_fft $= 1024$, hop 512), both at 16 kHz. The token sequence is pooled into a clip embedding by concatenating its mean and standard deviation over the N tokens (statistics pooling) [19, 20]. Mean-only pooling is kept as an ablation, and the embedding is left un-normalized.

3.2. Frozen Audio Encoders

We compare two frozen encoders under an otherwise identical pipeline. The Audio Spectrogram Transformer (AST) [11] is a 12-layer, 12-head ViT with embedding dimension $M = 768$ and a 16×16 patch grid. Since the challenge forbids arbitrary external weights, we do not use a public checkpoint but self-pre-train the AST on AudioSet [9] with a multi-label AdaCos [10] objective and a masked-patch reconstruction branch, using no DCASE Task 2 labels; the encoder is then loaded frozen and its decoder discarded. FISHER [8] is used off the shelf as a foundation model for industrial signals. Instead of a full spectrogram it splits the log-amplitude STFT into fixed-bandwidth sub-bands, encodes each with a ViT, and concatenates the per-sub-band `[CLS]` tokens, which lets one model cover arbitrary sampling rates. Its released checkpoint is pre-trained by teacher–student self-distillation on a 17k-hour mixture of general-audio corpora, again with no industrial-anomaly or DCASE labels. We use the FISHER-small variant (12 layers, 6 heads, $K = 384$, $\approx$22M parameters) on the spectrogram view with its native 25 ms window and 10 ms hop. Both encoders are loaded frozen with their training-only stochastic layers disabled, so embeddings are deterministic.

3.3. Scoring Back-Ends and Calibration

Scoring is grouped per machine type (`group_keys` $=$ [`machine`]), keeping domain and section labels out of the scoring path. We use two complementary back-ends. The k-nearest-neighbor (kNN) memory bank stores the normal training embeddings of a machine and scores a test clip by its Euclidean distance to the $k = 1$ nearest neighbor, a purely non-parametric detector. The Gaussian mixture model (GMM) instead fits a per-machine full-covariance mixture (4 components, n_init $= 3$, reg_covar $= 10^{-6}$, `float64`) and scores by negative log-likelihood.

Raw scores are not comparable across machines or domains, since their scale depends on how densely the embedding space is populated. We therefore apply a per-group z-score calibration fitted on training scores only: for a group g (e.g. machine and domain) with training mean μ_g and standard deviation σ_g, a raw score s is mapped to

$$\tilde{s} = \frac{s-\mu_g}{\sigma_g+\varepsilon}, \qquad \varepsilon = 10^{-12}, \quad (2)$$

with unseen groups falling back to global statistics. As an alternative we also consider a local-density calibration that compares s against its $k = 5$ nearest training scores in the same group,

$$\tilde{s} = \frac{s-m_k(s)}{d_k(s)+\varepsilon}, \qquad \varepsilon = 10^{-12}, \quad (3)$$

Table 1: The four submitted systems. All use frozen encoders and per-machine grouping; they differ in the embedding source, feature extractor, and the distance/density back-end.

System	View	Feature	Encoder	Back-end	Pooling
`Glitza_IKA_task2_1`	Dual-Channel	Spec	FISHER [8]	kNN (k=1, Euclidean)	mean–std
`Glitza_IKA_task2_2`	Dual-Channel	Spec	FISHER [8]	GMM (full, 4)	mean
`Glitza_IKA_task2_3`	Concat	LMBE	AST (AudioSet AdaCos + masked patch)	kNN (k=1, Euclidean)	mean–std
`Glitza_IKA_task2_4`	Concat	LMBE	AST (AudioSet AdaCos + masked patch)	GMM (full, 4)	mean

so that the same distance counts as less anomalous in a dense region and more anomalous in a sparse one. For the binary submission decision we threshold the calibrated training scores at their p-th percentile,

$$\theta \;=\; \mathrm{percentile}_p\big(\{s_1,\ldots,s_N\}\big), \qquad p=95, \tag{4}$$

and flag a clip as anomalous when $a(x) \geq \theta$. All calibration and threshold statistics come from training data only, so no test-split information leaks into scoring, and the threshold affects only the hard decisions: the official score uses the continuous $a(x)$. We optimize and report that score Ω, the harmonic mean h over the per-(machine, section) source/target AUCs and partial AUCs,

$$\Omega = h(\mathrm{AUC}_{m,n,d}, \mathrm{pAUC}_{m,n}), \tag{5}$$

with $p = 0.1$, $m \in \mathcal{M}$, $n \in \mathcal{N}$ for the partial AUC and $d \in \{\mathrm{source}, \mathrm{target}\}$ in addition for the partial AUC. Normal clips are split by domain and anomalous clips pooled by section.

4. EXPERIMENTS

4.1. Dataset

We develop and evaluate on the DCASE 2026 Task 2 dataset for first-shot unsupervised ASD [21, 22]. Each machine type provides a small set of normal training clips and an evaluation set of normal and anomalous clips split into a *source* and a *target* domain. Audio is processed at 16 kHz, and the two-channel signal is kept as a stereo tensor rather than averaged to mono before the front-end.

4.2. Self-Pre-Trained AST

Because the challenge restricts external pre-trained weights, the AST systems use an encoder we pre-train ourselves on AudioSet [9] rather than a public checkpoint. Training combines a masked-patch reconstruction loss (weight 0.1), where a MAE-style decoder rebuilds masked log-mel patches, with a multi-label AdaCos classification loss (weight 1.0), where an AdaCos head [10] predicts the 527 AudioSet classes from the CLS latent using normalized embeddings and class centers with a binary cross-entropy objective. The encoder follows Section 3.2 (128-band log-mel input, 16×16 patches, $D = 768$, 12 layers, 12 heads, mask ratio 0.5); after training, it is exported and loaded frozen by the ASD systems.

4.3. Submitted Systems

We submit four systems forming a 2×2 grid over the embedding source and the back-end (Table 1). All share the same per-machine grouping and domain z-score calibration and differ only in encoder, feature, and back-end.

Table 2: Overall results on the DCASE 2026 Task 2 development set (percentages). AUC values are averaged over machines and domains; Ω is the official harmonic-mean score. Best value per column in bold.

	System	$\mathrm{AUC}_{\mathrm{src}}$	$\mathrm{AUC}_{\mathrm{tgt}}$	pAUC	Ω
Baseline	MSE (AE)	67.46	52.74	54.50	56.66
	MAHALA	67.16	55.08	54.14	57.66
Ours	S1: FISHER + kNN	69.73	**56.22**	54.60	**58.23**
	S2: FISHER + GMM	**70.29**	53.06	**55.96**	57.06
	S3: AST + kNN	63.23	49.49	51.83	53.57
	S4: AST + GMM	58.77	43.96	50.65	48.39

The AST systems use the self-pre-trained encoder of Section 4.2, loaded frozen with its classification head and decoder discarded and `max_epochs=0` so that no adaptation to the challenge data takes place; Systems 1 and 2 swap in the FISHER encoder under the same frozen regime, leaving every other stage unchanged so the two embedding sources are directly comparable. The kNN back-end uses $k = 1$ with Euclidean distance, the GMM uses four full-covariance components, and hard decisions follow the percentile rule (Eq. 4). Alongside Ω we also report the mean source AUC, target AUC, and pAUC to show where each system gains or loses.

5. EXPERIMENTAL RESULTS

5.1. Overall Results

The FISHER systems clearly lead (Table 2). System 1 (FISHER + kNN) takes the top official score at $\Omega = 58.23\%$, ahead of the Mahalanobis (57.66%) and autoencoder (56.66%) baselines and without any training on the challenge data. System 2 (FISHER + GMM) follows at $\Omega = 57.06\%$, level with MAHALA, and posts the best source AUC (70.29%) and pAUC (55.96%) of any system. A frozen encoder with a plain distance- or density-based back-end is therefore already enough to match, and in places beat, the task-specific reconstruction baselines. The AST systems fall short, with System 3 (AST + kNN) at $\Omega = 53.57\%$ and System 4 (AST + GMM) at 48.39%: under the same pipeline our self-pre-trained AST embeddings carry less discriminative information than FISHER's, even though both encoders are frozen and of similar size (≈ 21.4M parameters). Two patterns hold for both encoders. The kNN back-end stays ahead of the GMM ($+1.2$ points Ω for FISHER, $+5.2$ for AST), so on these small reference sets a neighbor distance travels better than a per-machine mixture; and the target domain is a bottleneck everywhere, with every system losing 13 to 22 AUC points from source to target.

Table 3: Full per-machine results on the DCASE 2026 Task 2 development set (percentages). For each metric we report all seven machine types and the machine-wise average. Baselines are the official autoencoder (MSE) and selective Mahalanobis (MAHALA) modes [13, 3]; S1–S4 are our systems (S1: FISHER+kNN, S2: FISHER+GMM, S3: AST+kNN, S4: AST+GMM). Best value per row in bold.

Machine	MSE	MAHALA	S1	S2	S3	S4
AUC source [%]						
ToyCarEmu	69.62	69.49	**78.06**	77.98	58.35	69.26
ToyCar	75.62	77.28	**83.50**	82.14	68.70	67.70
bearingEmu	62.34	**65.92**	55.84	58.48	60.50	53.26
fan	61.45	60.00	63.46	61.12	**69.13**	53.94
gearboxEmu	68.23	**74.48**	73.42	73.26	73.16	72.06
sliderEmu	**67.25**	66.36	54.76	59.06	64.95	58.58
valveEmu	67.74	56.60	79.04	**80.02**	47.81	36.56
Average	67.46	67.16	69.73	**70.29**	63.23	58.77
AUC target [%]						
ToyCarEmu	61.20	66.62	**67.88**	60.04	53.85	52.72
ToyCar	37.87	**53.17**	48.84	33.24	50.35	30.12
bearingEmu	59.56	**62.28**	57.44	58.76	60.81	55.46
fan	46.94	45.09	50.63	**51.40**	45.72	50.96
gearboxEmu	49.78	**52.74**	50.72	44.20	43.56	43.14
sliderEmu	45.05	49.18	44.56	47.34	**51.44**	48.22
valveEmu	68.78	56.50	73.46	**76.44**	40.67	27.10
Average	52.74	55.08	**56.22**	53.06	49.49	43.96
pAUC ($p = 0.1$) [%]						
ToyCarEmu	55.89	53.47	51.32	**59.11**	53.47	54.79
ToyCar	54.03	**58.25**	52.42	55.11	51.00	50.79
bearingEmu	59.85	**60.42**	53.89	51.79	59.16	53.21
fan	53.33	52.29	52.26	**54.05**	51.68	50.79
gearboxEmu	52.94	53.97	**55.68**	53.53	49.37	48.11
sliderEmu	**50.38**	50.36	48.42	48.74	48.79	48.37
valveEmu	55.08	50.20	68.21	**69.42**	49.34	48.53
Average	54.50	54.14	54.60	**55.96**	51.83	50.66
Official score Ω [%]						
Total	56.66	57.66	**58.23**	57.06	53.57	48.39

5.2. Per-Machine Analysis

The machine-level breakdown (Table 3) shows that most of the FISHER advantage comes from `valveEmu`, where Systems 1 and 2 reach 79.04% and 80.02% source AUC and 73.46% and 76.44% target AUC, well above the baselines, because valves produce impulsive, transient sounds that FISHER captures well. The same machine is the worst case for AST, whose source AUC drops to 47.81% (System 3) and 36.56% (System 4) and target AUC to 27.10%, which by itself pulls down their overall Ω. No system wins everywhere: the Mahalanobis baseline stays strong on `ToyCar`, `bearingEmu`, `gearboxEmu`, and `sliderEmu`, and the widest source-to-target gaps appear on `ToyCar`, where the GMM target AUC falls to 33.24% (System 2) and 30.12% (System 4). Closing that `ToyCar` gap is the clearest next step.

6. CONCLUSIONS

We presented a training-free approach to first-shot ASD that pairs a frozen, general-purpose audio encoder with a light kNN or GMM back-end. On the DCASE 2026 Task 2 development set the frozen FISHER encoder with a kNN memory bank scored best ($\Omega = 58.23\%$) and was the only system to beat both baselines, with FISHER + GMM close behind, so frozen foundation embeddings with a simple estimator can stand in for task-specific reconstruction models here.

The two back-ends trade off source and target in opposite ways. The per-machine GMM is fitted mostly on the plentiful source clips, so it earns the best source AUC (70.29%) but treats the sparse target clips as anomalous; the kNN scores by distance to the nearest references, so even a few target anchors keep it steady under the domain shift, which is what wins the harmonic mean. The AST systems land below the baselines, most likely because our AST saw only the *balanced* AudioSet subset ($\approx$ 20k clips) rather than the full corpus ($\approx$ 2M), too little to learn embeddings for subtle machine-specific anomalies; this is clearest on `valveEmu`, where AST collapses to 35–46% while FISHER clears 73%.

Future work will pre-train the AST on the full AudioSet, attack the source-to-target gap with light few-shot adaptation, and combine the sharp source density of the GMM with the cross-domain robustness of the kNN back-end.

7. REFERENCES

[1] T. Nishida, N. Harada, D. Takeuchi, D. Niizumi, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2026 challenge task 2: Noise-aware unsupervised anomalous sound detection for machine condition monitoring," 2026.

[2] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Description and discussion on DCASE 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," 2023.

[3] T. Nishida, N. Harada, D. Niizumi, D. Albertini, R. Sannino, S. Pradolini, F. Augusti, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2025 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," 2025.

[4] G. Zhong, Q. Wang, J. Du, L. Wang, M. Cai, and X. Fang, "An enhanced audio feature tailored for anomalous sound detection based on pre-trained models," 2025.

[5] P. Saengthong and T. Shinozaki, "GenRep for first-shot unsupervised anomalous sound detection of DCASE 2025 challenge," Technical Report, DCASE2025 Challenge, 2025.

[6] H.-H. Wu, W.-C. Lin, A. D. Kumar, and L. Bondi, "Towards few-shot training-free anomaly sound detection," in *Proc. Interspeech*, 2025.

[7] J. Ren, T. Tang, H. Jia, H. Fayek, X. Li, S. Ma, X. Xu, and F. Xia, "Foundation models for anomaly detection: Vision and challenges," 2025.

[8] A. Jiang, P. Fan, B. Han, W.-Q. Zhang, J. Liu, and Y. Qian, "FISHER: A foundation model for multi-modal industrial signal comprehensive representation," 2025.

[9] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[10] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li, "AdaCos: Adaptively scaling cosine logits for effectively learning deep face representations," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 823–10 832.

[11] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proc. Interspeech*, 2021, pp. 571–575.

[12] T. Nishida, N. Harada, D. Niizumi, D. Albertini, R. Sannino, S. Pradolini, F. Augusti, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," 2024.

[13] N. Harada, D. Niizumi, Y. Ohishi, D. Takeuchi, and M. Yasuda, "First-shot anomaly sound detection for machine condition monitoring: A domain generalization baseline," in *Proc. 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 191–195.

[14] I. Kuroyanagi, T. Hayashi, K. Takeda, and T. Toda, "Serial-OE: Anomalous sound detection based on serial method with outlier exposure capable of using small amounts of anomalous data for training," 2025.

[15] 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," 2025.

[16] X. Fang, G. Zhong, Q. Wang, F. Chu, and J. Du, "Improving anomalous sound detection with attribute-aware representation from domain-adaptive pre-training," 2025.

[17] T. Fujimura, I. Kuroyanagi, and T. Toda, "The NU systems for DCASE 2025 challenge task 2," Technical Report, DCASE2025 Challenge, 2025.

[18] B. Deng, J. Chen, Z. Hong, X. Qu, G. Li, J. Wan, C. Xie, and J. Wang, "Enhancing anomalous sound detection with multi-level memory bank," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2024.

[19] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. Interspeech*, 2017, pp. 999–1003.

[20] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[21] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

[22] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Barcelona, Spain, November 2021, pp. 1–5.