ML-SCAD: MULTI-LEVEL STEREO CONDITIONAL ANOMALY DETECTION FOR DCASE 2026 TASK 2

Technical Report

Junghyun Moon

Independent Researcher
Seoul, Republic of Korea
racehorsed100@gmail.com

ABSTRACT

This technical report describes ML-SCAD (Multi-Level Stereo Conditional Anomaly Detection), a system submitted to DCASE 2026 Task 2 that treats the synchronized two-channel stereo recording as a core design element rather than an auxiliary input. The near microphone (machine-dominant) provides the primary anomaly representation; the far microphone (noise-dominant) is exploited in five complementary roles across signal, structural, semantic, score, and spectral levels. The system combines a BEATs/CLAP/autoencoder base ensemble with stereo-aware modules: Wiener-filtered AE (L1), inter-channel relation LOF (L2), machine-specific BEATs routing with Far-channel Attribute-Conditioned Anomaly detection (FACA) and Noise-Aware Score Normalization (NASN) (L3/L4), and Mel-Band Cross-Spectrum (MBCS) plus Cross-Channel Predictive Coding (CCPC) auxiliary branches (L5). MBCS approximates a stereo acoustic-transfer signature via ILD temporal variability (ILD_σ) and magnitude-squared coherence (MSC), achieving standalone avg $\Omega{=}0.5950$, outperforming BEATs-only Mahalanobis ($\Omega{=}0.5502$) on all 7 development machines with identical hyperparameters. The submitted system ($s_{\text{final}}{=}0.85\tilde{s}_{\text{base}}{+}0.10\tilde{s}_{\text{CCPC}}{+}0.05\tilde{s}_{\text{MBCS}}$) achieves per-machine avg $\Omega{=}0.6581$ ($\mathrm{AUC}_{\text{src}}{=}0.7793$, $\mathrm{AUC}_{\text{tgt}}{=}0.6501$, pAUC$=0.6051$; official single-HM $\Omega{=}0.6431$) on the 7-machine development set. No anomaly labels are used for training, and evaluation labels are never accessed.

***Index Terms*—** anomalous sound detection, stereo audio, Mel-Band Cross-Spectrum, Wiener filter, BEATs, Mahalanobis distance, cross-channel predictive coding

1. INTRODUCTION

Unsupervised anomalous sound detection (ASD) identifies abnormal machine states from normal-only training data [1]. DCASE 2026 Task 2 introduces a *noise-awareness challenge*: factory-floor recordings are collected with two synchronized microphones—a near microphone placed close to the target machine (machine-dominant) and a far microphone at a greater distance (noise-dominant). A high anomaly score may arise from a genuine mechanical fault *or* from an acoustic environment change; the two channels provide complementary evidence to separate these causes.

ML-SCAD organises the far-channel roles into five levels (Fig. 1):

- **L1 Signal (Wiener-AE):** The far channel serves as a per-bin noise reference for Wiener spectral subtraction. The noise-suppressed near-channel signal feeds a per-machine autoencoder; the Wiener-AE reconstruction error is blended with the standard AE error via a machine-specific weight γ.
- **L2 Structural (Relation LOF):** A 25-dim inter-channel feature vector encodes per-band magnitude-squared coherence (MSC), per-band near/far log-energy ratio (EFR), GCC-PHAT cross-correlation peak strength, and per-band temporal coherence variability. Normality is scored by Local Outlier Factor [2] with iterative self-training on pseudo-normal test clips.
- **L3 Semantic (BEATs Routing + FACA):** BEATs [3] embeddings are routed to machine-specific scorers (Mahalanobis, PatchCore [4], Spectral-Kurtosis LOF, or Attribute-Conditioned Mahalanobis). For bearingEmu, FACA fits k-means on *far*-channel embeddings to derive noise-condition centroids, enabling per-cluster near-channel Mahalanobis scoring without attribute labels.
- **L4 Score (NASN):** Noise-Aware Score Normalization subtracts a scaled far-channel score from the near-channel score on robustly normalised values, $s_{\text{NASN}}{=}\tilde{s}_{\text{near}}{-}\alpha\tilde{s}_{\text{far}}$, isolating machine-state differences from noise-induced score inflation (applied to sliderEmu and ToyCar).

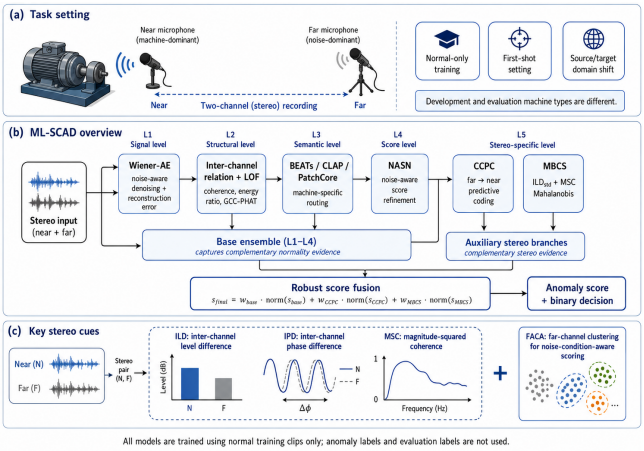

Figure 1: Overview of the ML-SCAD framework. Near (machine-dominant) and far (noise-dominant) channel signals feed five complementary levels (L1–L5). Component scores are aligned by robust score normalization (RSN) and fused into the final anomaly score. Trainable models are fitted using normal training clips only.

- **L5 Blend (MBCS + CCPC):** Two stereo auxiliary branches are fused at fixed low weights. MBCS (Mel-Band Cross-Spectrum Mahalanobis) approximates a stereo acoustic-transfer signature via ILD_σ and MSC. CCPC (Cross-Channel Predictive Coding) trains an MLP to predict near-channel BEATs embeddings from far-channel embeddings; deviations from the learned normal prediction pattern indicate anomalies.

2. TASK SETTING AND COMPLIANCE

DCASE 2026 Task 2 is positioned within a line of machine-condition monitoring ASD benchmarks, including ToyADMOS2 [5] and MIMII DG [6].

Each machine type provides ∼990 source and 10 target normal training clips; the test set contains ∼200 clips (100 per domain) with withheld labels. The official metric Ω is the harmonic mean of AUC_{src}, AUC_{tgt}, and pAUC@FPR=0.1 across all machines and domains [1]. The first-shot constraint (only 10 target training clips) makes per-machine target adaptation highly fragile, motivating unsupervised, domain-agnostic design choices.

ML-SCAD follows the DCASE 2026 Task 2 normal-only training setting. No anomaly labels are used for training the model, and no evaluation labels are accessed. Development labels are used only for offline validation and hyperparameter selection. No DCASE 2020–2025 Task 2 audio datasets are used; only the public baseline architecture/code is referenced. BEATs [3] and LAION-CLAP [7] are both on the official allowed external resource list. The binary decision threshold τ=0 is fixed before evaluation and is not derived from anomaly labels or test-clip label distributions.

3. PROPOSED SYSTEM: ML-SCAD

3.1. Audio Representations

Five complementary feature types are extracted per stereo clip (Table 1). BEATs [3] and LAION-CLAP [7] serve as fixed feature extractors with no fine-tuning. Log-mel features use 16 kHz; CLAP inputs are resampled to 48 kHz. MBCS uses a 256-sample hop to resolve finer temporal structure for cross-spectrum estimation.

3.2. L1: Wiener-Filtered Autoencoder

The far channel acts as a real-time noise reference for Wiener spectral subtraction applied to the near channel:

$$H(f,t) = \max\left(1 - \alpha \frac{|X_{\text{far}}(f,t)|^2}{|X_{\text{near}}(f,t)|^2+\varepsilon},\ \beta\right), \tag{1}$$

with over-subtraction factor α=0.7 and spectral floor β=0.05 (fixed, all machines). The filtered near-channel signal $\hat{X}_{\text{near}}=H\cdot X_{\text{near}}$ is converted to a 128-bin log-mel spectrogram and fed to a per-machine AE following the first-shot ASD baseline architecture [8, 9] (640-dim input, 3-layer bottleneck). The Wiener-AE reconstruction score is blended with the standard near-channel AE score:

$$s_{\text{L1}} = (1-\gamma)\, s_{\text{AE}} + \gamma\, s_{\text{WienerAE}}, \tag{2}$$

where the machine-specific blend weight γ is: fan (0.80), gearboxEmu (0.20), ToyCarEmu (0.10), bearingEmu (0.15), and 0 for all other machines. A high γ prioritises denoised input; a low γ retains rich near-channel content. The spectral floor β=0.05 prevents signal distortion and excessive attenuation in noise-dominant bins.

Table 1: Feature representations in ML-SCAD.

Feature	Notes	Dim
Log-mel AE	near; 128 mel, 5-frame ctx	640
BEATs	near; iter3+AS2M, mean pool	768
LAION-CLAP	near; clap-htsat, 48 kHz	512
MBCS	near+far; 40 mel, hop=256	200
Relation	near+far; MSC/EFR/GCC	25

Table 2: Machine-specific Level 3 routing. γ: Wiener-AE blend weight; α: NASN suppression coefficient.

Machine	Primary Scorer	Stereo Module
fan	PatchCore-256 + WienerMahal	Wiener (γ=0.80)
gearboxEmu	ACM-BEATs (9 attr. groups)	Attr. cond. (γ=0.20)
ToyCar	DSDE-PatchCore-256	NASN (α=0.4)
sliderEmu	BEATs Mahal-128	NASN (α=0.9)
bearingEmu	BEATs Mahal-128 + FACA	FACA (k=3)
ToyCarEmu	BEATs Mahal-256 + ACM-AE	Wiener (γ=0.10)
valveEmu	Spectral Kurtosis LOF	Periodic bypass

3.3. L2: Inter-Channel Relation Features

The 25-dim per-clip relation vector is:

$$R = [\,\text{MSC}_{1:8},\ \text{EFR}_{1:8},\ \text{GCC},\ \sigma_{\text{MSC},1:8}\,]^\top \in \mathbb{R}^{25} \tag{3}$$

where $\text{MSC}_{1:8}$ are 8-band coherence values; $\text{EFR}(b)=\log(E_{\text{near}}(b)+\varepsilon)-\log(E_{\text{far}}(b)+\varepsilon)$ is the per-band near/far log-energy ratio capturing normal SNR structure; GCC is the GCC-PHAT [10] cross-correlation peak strength (max normalized cross-correlation within ±50 samples); and $\sigma_{\text{MSC},1:8}$ is the temporal standard deviation of per-band MSC within the clip, capturing within-clip coherence instability. The relation vector is scored by LOF [2] (k=5). A 2-iteration unlabeled refinement adds the lowest-scoring 30% clips as auxiliary LOF reference samples without using anomaly or domain labels. Machine-specific blend weight $w_{\text{rel}}\in\{0.15,\ldots,0.30\}$ is tuned on the development set.

3.4. L3: Machine-Specific BEATs Routing and FACA

BEATs [3] embeddings (768-dim, pretrained on AudioSet-2M [11]) are reduced by PCA to $d\in\{128,256\}$ and scored by machine-specific routing (Table 2).

Attribute-Conditioned Mahalanobis (ACM) fits per-attribute cluster statistics on source training clips and scores each test clip against its nearest attribute cluster: $s_{\text{ACM}}=\min_c d_M(\mathbf{z};\hat{\mu}_c,\hat{\mathbf{\Sigma}}_c)$, where c indexes the 9 attribute groups for gearboxEmu and the AE-based attribute clusters for ToyCarEmu. **valveEmu** shows highly periodic vibration; Spectral Kurtosis (SK) routing detects clips with SK>0.60 and scores them against high-kurtosis training clips only, bypassing the general embedding scorer and yielding Ω=0.8713.

Far-Channel Attribute-Conditioned Anomaly (FACA). FACA removes noise-condition confounding without attribute labels. For bearingEmu, k-means (k=3) is fitted on training-normal far-channel BEATs-PCA32 embeddings, deriving three noise-condition centroids. At inference, each test clip is assigned to its nearest

centroid: $c^*(x)=\arg\min_c \|\mathbf{z}_{\text{far}}(x)-\mu_c\|$, and per-cluster Mahalanobis scoring on near-channel BEATs-PCA256 is applied. This conditioning reduces the effect of environment-induced variation on the near-channel anomaly score. FACA contributes $\Delta\Omega{=}{+}0.0065$ over global Mahalanobis for bearingEmu ($\Delta\text{AUC}_{\text{src}}{=}{-}0.017$, $\Delta\text{AUC}_{\text{tgt}}{=}{+}0.027$), pointing to enhanced noise-condition decoupling in the target domain where environment variation is larger.

3.5. L4: Noise-Aware Score Normalization (NASN)

NASN performs score-level noise suppression on robustly normalised scores:

$$s_{\text{NASN}} = \tilde{s}_{\text{near}} - \alpha \cdot \tilde{s}_{\text{far}}, \tag{4}$$

where $\tilde{s}_{\text{near}}$ and $\tilde{s}_{\text{far}}$ are RSN-normalised Mahalanobis scores of near and far BEATs embeddings respectively, and α controls the suppression strength. Normalising before subtraction ensures the two channels are on a comparable scale. The intuition is: a simultaneously elevated far-channel score suggests an acoustic environment change rather than a machine fault. **sliderEmu** ($\alpha{=}0.9$): $\Delta\Omega{=}{+}0.0127$, primarily target AUC improvement, consistent with strong far-channel noise variation under domain shift. **ToyCar** ($\alpha{=}0.4$, blend $w{=}0.20$): $\Delta\Omega{=}{+}0.0018$.

3.6. L5: Mel-Band Cross-Spectrum (MBCS)

MBCS approximates an inter-channel stereo-transfer signature between the two microphones via level variability and coherence. The full 200-dim feature covers ILD, IPD, and MSC over 40 mel bands; the submitted best subset uses the 80-dim pair (ILD_σ, MSC) projected to PCA-64. Let $Z_c(f,t)$ denote the STFT at bin f, frame t, channel $c \in \{\text{near}, \text{far}\}$, and let $\mathcal{F}_b$ be the set of STFT bins in mel band b. Define the band-aggregated complex cross-spectrum $C_b(t){=}\sum_{f\in\mathcal{F}_b} Z_{\text{near}}(f,t)Z^*_{\text{far}}(f,t)$ and per-channel band power $P_{c,b}(t){=}\sum_{f\in\mathcal{F}_b}|Z_c(f,t)|^2$. The two features per band are:

$$\text{ILD}_\sigma(b) = \text{std}_t\left\{\log\frac{P_{\text{near},b}(t){+}\varepsilon}{P_{\text{far},b}(t){+}\varepsilon}\right\}, \tag{5}$$

$$\text{MSC}(b) = \frac{1}{T}\sum_{t=1}^{T}\frac{|C_b(t)|^2}{P_{\text{near},b}(t)\,P_{\text{far},b}(t){+}\varepsilon}. \tag{6}$$

Physical motivation. Under normal machine operation the near/far spatial relationship is stable, so large ILD_σ indicates a change in vibration propagation or posture. Band-level MSC aggregates the complex cross-spectrum within each mel band before squaring, providing values in $[0,\ 0.98]$; unlike per-bin MSC (which collapses to 1 for any STFT bin), it provides genuine inter-channel coherence discrimination. **Domain robustness.** Both features are within-clip *relative* statistics invariant to absolute recording distance, making them robust to source/target domain shift in far-channel recording position. Absolute ILD mean and IPD, by contrast, degrade under domain shift and underperform (see Table 4). A Mahalanobis scorer with ridge-regularized covariance is fitted on PCA-64 projections.

3.7. L5: Cross-Channel Predictive Coding (CCPC)

CCPC is designed to exploit cross-channel stereo relationships at the embedding level: an MLP trained on normal clips predicts near-channel BEATs-PCA128 (φ_{near}) from far-channel BEATs-PCA128 (φ_{far}). Architecture: 2 hidden layers, 512 units, LayerNorm + GELU activations, Dropout($p{=}0.1$), AdamW optimiser, 300 epochs. The anomaly score is the cross-channel prediction residual:

$$s_{\text{CCPC}}(x) = \left\|\text{MLP}(\varphi_{\text{far}}(x)) - \varphi_{\text{near}}(x)\right\|^2. \tag{7}$$

Under normal operation the near/far embedding relationship is consistent; a large residual indicates that the far channel fails to predict the near channel in the expected way. To isolate the source of the gain, an ablation compares CCPC against a same-channel temporal-shift baseline (SC-TS): a near→near MLP with a 1-frame temporal shift, which carries no cross-channel spatial information. SC-TS achieves $+0.0081$ avg $\Delta\Omega$ vs. CCPC at $+0.0082$—nearly identical. This suggests that the CCPC gain mainly comes from *predictive representation structure* (smoothness of the BEATs manifold under normal conditions) rather than cross-channel stereo geometry, which is largely lost at the embedding level. This finding motivates earlier-stage cross-channel fusion rather than late-stage embedding prediction in future noise-aware ASD systems.

4. SCORE FUSION AND DECISION

Raw scores from heterogeneous components span incompatible scales. Robust score normalization (RSN) aligns them per component:

$$\tilde{s} = \frac{s - \text{median}(s_{\text{test}})}{\text{MAD}(s_{\text{test}})}, \tag{8}$$

where median and MAD are estimated from unlabeled score *magnitudes* within each evaluation section, using no anomaly labels, domain labels, machine-condition labels, file-name metadata, or threshold tuning. Scale alignment via RSN maps heterogeneous component scores to a common reference before fixed-weight fusion in Eq. (9). The base ensemble integrates Wiener-AE, relation LOF, CLAP-LOF, and BEATs Mahalanobis/PatchCore with internal machine-specific weighting. The final submitted score is:

$$s_{\text{final}} = 0.85\,\tilde{s}_{\text{base}} + 0.10\,\tilde{s}_{\text{CCPC}} + 0.05\,\tilde{s}_{\text{MBCS}}. \tag{9}$$

Blend weights $(0.85, 0.10, 0.05)$ are fixed uniformly for all five evaluation machines to prevent overfitting to development machine characteristics, as evaluation machine types are entirely unseen. The binary decision $\hat{y}{=}\mathbf{1}[s_{\text{final}}{>}\tau]$ uses $\tau{=}0$, the natural threshold after RSN centering.

5. EXPERIMENTS

5.1. Experimental Setup

The DCASE 2026 Task 2 development set [1] contains 7 machine types: ToyCar, ToyCarEmu, fan, gearboxEmu, bearingEmu, sliderEmu, and valveEmu. Each machine provides $\sim$990 source and 10 target normal training clips, and $\sim$100 source and 100 target test clips (50 normal + 50 anomaly per domain). The evaluation set contains 5 unseen machine types (ToyDrone, ToothBrush, SewingMachine, BlowerDustCollector, Sander) whose anomaly labels are never accessed.

STFT parameters: 1024-pt FFT, Hann window; AE and BEATs features use 512-sample hop; MBCS uses 256-sample hop over 40 mel bands. BEATs embeddings are extracted with masked mean pooling (padding mask applied) to handle variable-length clips. No data augmentation is used at any stage. All hyperparameters are selected on development set overall Ω and kept fixed for all evaluation machines.

Table 3: Submitted ML-SCAD system results on the development set. Per-machine avg Ω=0.6581; official single-HM Ω=0.6431.

Machine	AUC_{src}	AUC_{tgt}	pAUC	Ω
ToyCar	0.7004	0.7024	0.5458	0.6405
ToyCarEmu	0.6920	0.8932	0.5737	0.6964
fan	0.9360	0.3572	0.5058	0.5133
gearboxEmu	0.7784	0.6360	0.6400	0.6788
bearingEmu	0.6776	0.5848	0.6042	0.6197
sliderEmu	0.6772	0.5668	0.5342	0.5868
valveEmu	0.9932	0.8100	0.8321	0.8713
Avg	**0.7793**	**0.6501**	**0.6051**	**0.6581**

Table 4: MBCS Mahalanobis ablation (avg over 7 dev machines, identical hyperparameters).

Config	Dim	PCA	avg Ω
BEATs Mahal	768	128	0.5502
MBCS-200	200	–	<0.50
ILD_{μ} only	40	–	0.5301
MSC only	40	40	0.5674
ILD_{σ}+MSC	**80**	**64**	**0.5950**
+W.AE	80	64	0.5993

5.2. Development Set Results

Table 3 reports per-machine performance of the submitted ML-SCAD system. The base ensemble (without CCPC/MBCS) achieves avg Ω=0.6614; the full submitted blend (Eq. 9) yields Ω=0.6581 (Δ=−0.0033). The slight decrease reflects deliberately conservative auxiliary weights: CCPC and MBCS are retained at low weights because they may generalise differently to the 5 unseen evaluation machines, and overfitting these weights to 7 development machines would not be appropriate.

Notable observations. fan exhibits the largest AUC_{src}−AUC_{tgt} gap (0.936 vs. 0.357), indicating that the near-channel machine signal is dominated by environmental noise in the target domain—a case where even Wiener filtering at γ=0.80 cannot fully decouple machine and noise. The sub-chance AUC_{tgt} (0.357) indicates systematic score inversion, suggesting a negative-transfer effect where target-domain noise distribution changes disrupt the far-channel Wiener reference. **valveEmu** benefits strongly from SK-LOF periodic bypass routing (Ω=0.8713), suggesting that machine-type-specific signal structure can be highly discriminative when correctly identified.

5.3. Ablation Studies

ILD_{μ} and full-200 features underperform BEATs-only, suggesting that absolute inter-channel statistics degrade under domain shift. The ILD_{σ}+MSC subset outperforms BEATs-only by +0.0448 avg Ω (7/7 machines) with a single fixed hyperparameter set and no per-machine tuning. The near-identical CCPC and SC-TS improvements suggest that cross-channel spatial information is not preserved at the BEATs embedding level. Disabled components include BiCCPC (bidirectional CCPC: −0.0119), CCSD (near−far BEATs Mahalanobis: negative on all machines), and Normalizing Flow (average −0.009 vs. Mahalanobis), all of which reduced performance on the development set and were excluded from the submitted system.

Table 5: Per-component stereo contribution (dev set). Deltas are local ablations, not additive with the full ensemble.

Component	Scope	avg $\Delta\Omega$
MBCS ILD_{σ}+MSC	vs. BEATs-only	+0.0448
NASN (α=0.9)	sliderEmu	+0.0127
FACA (k=3)	bearingEmu	+0.0065
NASN (α=0.4)	ToyCar	+0.0018
CCPC (w=0.10)	6/7 machines	+0.0082
SC-TS (ablation)	6/7 machines	+0.0081
BiCCPC (disabled)	7/7 machines	−0.0119
CCSD (disabled)	7/7 machines	negative

6. CONCLUSIONS

ML-SCAD is a five-level stereo anomaly detection system that exploits the near/far microphone structure of DCASE 2026 Task 2 at the signal, structural, semantic, score, and spectral levels. The MBCS feature approximates a stereo acoustic-transfer signature via ILD temporal variability and band-level coherence—domain-robust statistics that consistently outperform BEATs-only Mahalanobis as a standalone scorer (+0.0448 avg Ω). FACA, NASN, and Wiener filtering each address a distinct facet of the noise-awareness challenge. The submitted system achieves per-machine avg Ω=0.6581 (official Ω=0.6431).

A primary finding is that CCPC and its same-channel ablation SC-TS yield nearly identical improvements (+0.0082 vs. +0.0081), demonstrating that cross-channel spatial information is lost at the BEATs embedding level and motivating earlier-stage cross-channel fusion in future work. The primary limitation is the severe target-domain degradation of fan-type machines under acoustic domain shift, where Wiener filtering alone is insufficient. Future work will focus on earlier-stage stereo fusion and normal-score-based calibration.

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," *arXiv*, 2026, arXiv:2606.01578.

[2] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. ACM SIGMOD International Conference on Management of Data*, 2000, pp. 93–104.

[3] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proc. 40th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 202, 2023, pp. 5178–5193.

[4] K. Roth, L. Pemula, J. Zepeda, B. Schoelkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14 318–14 328.

[5] 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.

[6] 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.

[7] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[8] 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.

[9] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," DCASE Challenge, Tech. Rep., 2023, arXiv:2305.07828.

[10] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[11] 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.