ENHANCING MACHINE SOUND ANOMALY DETECTION VIA SOURCE SEPARATION AND HYBRID SSL FUSION

Technical Report

Yucong Zhang[1,2]*, Zhang Chen*[1]*, Ming Li*[1,2]

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

ABSTRACT

This technical report presents our solution for DCASE 2025 Task 2: Anomalous Sound Detection for Machine Condition Monitoring. Our approach integrates BEATs and AudioMAE models through two fusion strategies: 1) score-level ensemble of independently fine-tuned models, and 2) feature-level fusion with unified attentive statistical pooling. Both models employ LoRA-based adaptation on combined historical and current DCASE datasets, enhanced by source separation for clean-referenced machines and universal noise augmentation. The anomaly detection mechanism leverages prototype embeddings generated from KMeans clustering and target samples. Achieving a 66.34% average AUC/pAUC score on the development set, our system demonstrates 10.47% improvement over the baseline, highlighting the effectiveness of hybrid fusion strategies in capturing diverse normal sound patterns.

Index Terms— Anomalous sound detection, fine-tuning, machine sound separation, pre-trained

1. INTRODUCTION

Anomalous Sound Detection (ASD) aims to identify abnormal acoustic events in industrial machinery without prior exposure to fault patterns during training. This unsupervised paradigm presents a fundamental challenge: models must learn comprehensive representations of normal operation using exclusively nominal samples, yet detect subtle deviations caused by unseen anomalies during testing. The domain shift between training and deployment environments further complicates this task, as acoustic signatures vary significantly across machine types, operating conditions, and background noise profiles.

Transfer learning via pre-training and fine-tuning has emerged as a dominant framework across audio domains, where models first learn general acoustic representations from large-scale datasets before adapting to downstream tasks. Recent DCASE competitions demonstrate this paradigm's efficacy for ASD, with top-ranked solutions [1, 2] leveraging models pre-trained on AudioSet [3] or other sound corpora. Such approaches outperform traditional methods by capturing normal sound characteristics through self-supervised objectives, effectively addressing the data scarcity inherent to industrial settings.

This work introduces novel methodologies addressing two pivotal changes in DCASE 2025 Task 2 [4]: (1) Permission to utilize historical competition data (2020-2024) significantly expands the normal sound corpus, and (2) Provision of clean machine references or isolated background noise enables targeted audio enhancement. Our solution integrates these advancements through a hybrid framework featuring:

- **Two-stage Pre-training**: We propose a sequential self-supervised approach where models first learn machine sound distributions via spectrogram reconstruction (generative self-supervised learning (SSL)), followed by discriminative attribute classification using cross-entropy loss. This dual-phase strategy bridges representation learning and domain adaptation.
- **Source Separation Pipeline**: For machine types with clean references, we implement TF-GridNet-based [5] separation trained on clean-noisy pairs. This preprocessing stage isolates mechanical signatures while maintaining original inputs during inference, effectively leveraging the newly provided acoustic resources.

2. METHODOLOGY

Our solution employs two distinct self-supervised learning paradigms through AudioMAE [6] and BEATs [7] models, each addressing different aspects of normal machine sound modeling.

2.1. AudioMAE: Generative-Discriminative Hybrid

The AudioMAE branch combines spectrogram reconstruction and discriminative learning. Pre-trained on AudioSet through masked autoencoding with 80% random frequency masking, the model first undergoes domain adaptation via generative SSL on combined DCASE datasets. During this phase, random rectangular masks spanning 64-128 frequency bins are applied, with reconstruction targets set to original log-Mel spectrograms using L1 loss:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|\mathcal{M}|} \sum_{(i,j)\in\mathcal{M}} |\hat{X}_{i,j} - X_{i,j}| \tag{1}$$

where $\mathcal{M}$ denotes masked positions. Subsequently, the model transitions to discriminative learning using machine-type labels from historical datasets, followed by final adaptation on DCASE 2025 data. Temporal averaging compresses features to $\mathbf{h}_a \in \mathbb{R}^D$.

Following the generative pre-training phase, the AudioMAE branch undergoes two distinct discriminative fine-tuning stages to adapt to the anomaly detection task.

Phase I: Full Data Adaptation
In the first discriminative phase, we fine-tune the entire model (excluding frozen pretrained weights) on the combined dataset containing both historical DCASE data (2020-2024) and the 2025 training set. This stage employs machine-type classification as the learning objective, with compound labels encoding device attributes and operating conditions (e.g., "fan_rpm_1800_load_medium"). The cross-entropy loss is computed as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}) \tag{2}$$

where $y_{i,c}$ denotes the ground-truth label and $p_{i,c}$ the predicted probability for class c. All model parameters except the frozen pretrained backbone are updated during this phase.

Phase II: 2025-Specific LoRA Tuning
The second discriminative stage focuses on task-specific adaptation using only the DCASE 2025 dataset. Building upon the Phase I model, we implement Low-Rank Adaptation (LoRA) [8] with rank $r = 8$ for efficient parameter updates. The adaptation process modifies the self-attention layers in AudioMAE's transformer blocks through:

$$\mathbf{W}'_q = \mathbf{W}_q + \mathbf{B}_q\mathbf{A}_q, \quad \mathbf{W}'_k = \mathbf{W}_k + \mathbf{B}_k\mathbf{A}_k \tag{3}$$

where $\mathbf{W}_q, \mathbf{W}_k$ are original query/key projection matrices, and $\mathbf{B}_q, \mathbf{A}_q$ their low-rank counterparts. This phase maintains the same classification objective but reduces the learning rate to $1e^{-5}$ to prevent overfitting to the limited 2025 data.

Both phases utilize identical data preprocessing pipelines but differ in their augmentation strategies: Phase I and II introduce targeted noise injection using 2025-provided background samples.

2.2. BEATs: Discriminative Attribute Learning

The BEATs branch focuses on learning discriminative features through machine attribute classification. Initialized with weights from the publicly released BEATs model pre-trained on AudioSet, we implement a two-phase adaptation process similar to the process of training AudioMAE. In Phase I, the model undergoes fine-tuning using historical DCASE datasets (2020-2024 editions) with machine-type labels formatted as compound attributes. This process employs standard cross-entropy loss over 23 machine categories. Phase II continues the adaptation using DCASE 2025 training data with identical objectives but updated label spaces to only the machine types in DCASE 2025. To keep the best of the original model, we apply LoRA when fine-tuning the data in phase II. Temporal features from the final transformer layer $\mathbf{H}_b \in \mathbb{R}^{B\times T_b\times D}$ are compressed through attentive statistical pooling (ASP):

$$\mathbf{h}_b = \sum_{t=1}^{T_b}\alpha_t\mathbf{h}_t, \quad \alpha_t = \text{softmax}(\mathbf{w}^\top\tanh(\mathbf{W}\mathbf{h}_t)) \tag{4}$$

where $\mathbf{w} \in \mathbb{R}^{D_a}$ and $\mathbf{W} \in \mathbb{R}^{D_a\times D}$ are learnable parameters.

2.3. Ensemble Strategies

2.3.1. Score-Level Fusion

Our primary ensemble combines predictions from independently adapted BEATs and AudioMAE models. Let $s_b^{(i)}$ and $s_a^{(i)}$ denote anomaly scores for sample i from each branch. The final score is computed as the mean of the output scores of those two models.

2.3.2. Feature-Level Fusion

Both models utilize frozen backbone weights trained from phase I, and add LoRA adaptors for further fine-tuning. We implement two distinct feature fusion approaches using frozen encoders with LoRA adapters:

Attentive Fusion: Temporal features are concatenated before pooling:

$$\mathbf{H}^{(i)}_{\text{fused}} = [\mathbf{H}^{(i)}_b; \mathbf{H}^{(i)}_a] \in \mathbb{R}^{(T_b+T_a)\times D} \tag{5}$$

A trainable attention mechanism generates fused embeddings:

$$\mathbf{h}^{(i)}_{\text{fused}} = \sum_{t=1}^{T_b+T_a}\alpha_t\mathbf{h}_t, \quad \alpha_t = \text{softmax}(\mathbf{v}^\top\tanh(\mathbf{V}\mathbf{h}_t)) \tag{6}$$

where $\mathbf{v} \in \mathbb{R}^{D_a}$ and $\mathbf{V} \in \mathbb{R}^{D_a\times D}$ are newly initialized parameters.

Mean-Pooling Fusion: Features are compressed via temporal averaging before concatenation:

$$\mathbf{h}^{(i)}_{\text{fused}} = \left[\frac{1}{T_b}\sum_{t=1}^{T_b}\mathbf{h}^{(i,t)}_b; \frac{1}{T_a}\sum_{t=1}^{T_a}\mathbf{h}^{(i,t)}_a\right] \in \mathbb{R}^{2D} \tag{7}$$

Both approaches are trained end-to-end on the DCASE 2025 dataset with the following constraints:

- Encoder weights remain frozen (LoRA adapters active)
- Pooling/attention parameters are trainable

2.4. Prototype-Based Anomaly Scoring

For each test sample $\mathbf{x}_{\text{test}}$, we compute its embedding $\mathbf{h}_{\text{test}}$ and compare against a prototype set $\mathcal{P}$ constructed as:

$$\mathcal{P} = \{\mathbf{c}_1, ..., \mathbf{c}_{16}\} \cup \{\mathbf{t}_1, ..., \mathbf{t}_{10}\} \tag{8}$$

where $\mathbf{c}_k$ are KMeans centroids from 990 source samples and $\mathbf{t}_j$ are target sample embeddings. The anomaly score is computed as:

$$s(\mathbf{h}_{\text{test}}) = \min_{\mathbf{p}\in\mathcal{P}}\left(1 - \frac{\mathbf{h}_{\text{test}}\cdot\mathbf{p}}{\|\mathbf{h}_{\text{test}}\|\|\mathbf{p}\|}\right) \tag{9}$$

This minimum cosine distance approach effectively identifies deviations from established normal patterns.

2.5. Data Preprocessing and Augmentation

2.5.1. Source Separation for Clean Reference Machines

For machine types providing clean audio samples (e.g., "ToyCar", "bearing"), we implement a TF-GridNet-based [5] separation pipeline. Each machine type trains a dedicated model using clean-noisy pairs from DCASE 2025 with a continuous loss function [10].

2.5.2. Universal Noise Augmentation

To homogenize acoustic conditions across all machine types, we inject DCASE 2025-provided background noise into every training sample. For machines without clean references (e.g., "Fan", "Slide rail"), random noise segments are mixed at SNRs between -5 to 10 dB. Clean-reference machines first undergo separation before noise addition, ensuring consistent noise floors across the dataset.

Table 1: The ASD performance (shown in %) of submitted systems on the test dataset of the DCASE 2025 Task 2 development dataset. H. Mean in the first title row refers to the harmonic mean over all machine types. For metrics, sAUC: "source AUC"; tAUC: "target AUC"; pAUC: "partial AUC", H. Mean: "harmonic mean over all metrics". System

Model	Metric	Bearing	Fan	Gearbox	Slider	ToyCar	ToyTrain	Valve	H. Mean
System-1	sAUC	69.66	76.80	60.36	45.12	81.24	73.50	83.50	67.30
	tAUC	65.94	65.62	70.12	64.72	80.16	59.12	87.96	69.39
	pAUC	52.68	52.10	58.10	53.05	69.36	53.47	85.63	58.80
	H. Mean	62.45	53.13	76.53	60.95	61.85	63.22	85.66	64.83
System-2	sAUC	76.40	79.56	74.28	62.82	87.54	78.02	82.08	76.53
	tAUC	62.14	59.70	68.92	56.02	75.74	54.10	87.46	64.62
	pAUC	49.78	49.31	61.31	55.26	70.00	52.73	85.89	58.49
	H. Mean	67.75	57.84	77.10	59.69	60.89	60.49	85.08	65.74
System-3	sAUC	71.84	78.18	67.22	71.98	86.44	80.06	86.48	76.83
	tAUC	65.55	63.82	74.94	43.52	82.28	64.68	86.46	65.77
	pAUC	49.68	50.42	56.21	53.21	76.15	55.57	87.73	58.78
	H. Mean	65.20	53.89	81.40	65.30	60.88	62.12	86.89	66.34
System-4	sAUC	68.76	75.82	65.98	63.56	83.36	80.64	84.16	73.76
	tAUC	63.56	60.46	74.22	52.42	79.82	64.92	86.64	67.11
	pAUC	48.47	50.10	56.00	53.36	73.84	55.63	86.94	58.23
	H. Mean	64.53	56.02	78.81	65.53	58.93	60.38	85.90	65.75
Baseline [9] (MSE)	sAUC	66.53	70.96	64.80	70.10	71.05	61.76	63.53	67.69
	tAUC	53.15	38.75	50.49	48.77	53.52	56.46	67.18	51.39
	pAUC	61.12	49.46	52.49	52.32	49.70	50.19	57.35	52.94
	H. Mean	59.75	49.90	55.26	55.68	56.73	53.14	62.42	55.87
Baseline [9] (MAHALA)	sAUC	63.63	77.99	73.26	73.79	73.17	50.87	56.22	65.51
	tAUC	59.03	38.56	51.61	50.27	50.91	46.15	61.00	50.05
	pAUC	61.86	50.82	55.07	53.61	49.05	48.32	52.53	52.72
	H. Mean	61.45	51.34	58.61	57.58	55.87	48.37	56.37	55.34

3. EXPERIMENTS

3.1. Experimental Setup

Our experimental framework integrates three primary data sources: AudioSet (2 million 10-second YouTube audio clips across 527 classes) for large-scale pre-training, historical DCASE Task 2 datasets (2020-2024 editions) containing official development and additional training sets derived from ToyADMOS [13] and MIMII [14] datasets, and the DCASE 2025 dataset following the same structure with normal-only training samples and mixed normal/anomalous test data. The evaluation protocol employs four metrics: source-domain AUC (sAUC), target-domain AUC (tAUC), partial AUC (pAUC) over [0,0.1] false positive rates, and their harmonic mean (H.Mean).

All experiments were conducted on NVIDIA L20 GPUs using AdamW optimization ($\beta_1 = 0.9, \beta_2 = 0.999$) with base learning rate 2×10^{-4}, weight decay 1×10^{-5}, and batch size 512. The detailed training hyper-parameters are shown in Table 2.

3.2. System Description

The four developed systems are structured as follows:
System-1 constitutes a single-model approach using AudioMAE with two-stage self-supervised learning.
System-2 implements score-level fusion by averaging anomaly scores from two independent branches: 1) the AudioMAE pipeline described in System-1, and 2) a BEATs model discriminatively fine-tuned on identical data with LoRA adaptation (rank=8).
System-3 employs feature-level fusion through temporal concatenation of AudioMAE and BEATs embeddings, processed via trainable attentive statistical pooling. The architecture freezes both encoders' original parameters trained from Phase I while enabling LoRA-based adaptation (rank=64) during end-to-end training on this year's dataset.
System-4 explores alternative feature fusion via temporal mean-pooling.

3.3. Results and Analysis

Table 1 details performance comparisons on the DCASE 2025 development test set. The AudioMAE single system (System-1) establishes a baseline H.Mean of 64.83, validating our hybrid SSL

Table 2: Pre-training and Fine-tuning hyper-parameters. DC-T2 refers to task 2 of the DCASE challenge. DC-T2-all refers to all the training data from 2020 to 2025. DC-T2-2025 refers to the training data from this year. For augmentation, NA: "noise augmentation". For loss functions: MSE: "minimum square error"; CE: "cross entropy".

Configuration	**pre-training**		**fine-tuning**	
	AS-2M	DC-T2-all	DC-T2-all	DC-T2-2025
Optimizer	AdamW [11], $\beta_1 = 0.9, \beta_2 = 0.999$			
Weight decay	0.00001			
Base learning rate	0.0002			
Learning rate schedule	half-cycle cosine decay [12]			
Minimum learning rate	0.000001			
Warm-up epochs	3	10	10	2
Epochs	32	100	100	10
Batch size	512			
Augmentation	-	NA	NA	-
Loss Function	MSE		CE	CE

approach. Score-level fusion (System-2) achieves significant sAUC improvement at slight tAUC/pAUC costs. Feature-level fusion via temporal concatenation (System-3) delivers optimal overall performance with a harmonic score of 66.34, outperforming mean-pooling fusion (System-4). Valve detection demonstrates exceptional robustness, while Slider presents the most challenging scenario.

Comparative analysis reveals our systems surpass conventional auto-encoder baselines by 9.47 on average. The largest improvement occurs in Gearbox ASD detection, surpassing by 26.14%, demonstrating SSL's superiority in capturing complex mechanical patterns. Temporal feature fusion shows particular effectiveness for ToyCar and ToyTrain , suggesting improved representation learning for small mechanical components.

4. SUMMARY

This technical report presents a novel framework for DCASE 2025 Task 2, addressing two critical challenges through systematic innovations. First, by unifying historical DCASE datasets (2020-2024) with current competition data, we develop a hybrid self-supervised learning paradigm combining generative spectrogram reconstruction and discriminative attribute classification. Second, leveraging newly provided clean machine references, we implement TF-GridNet-based source separation to enhance acoustic pattern learning. Our architecture integrates BEATs and AudioMAE models through dual fusion strategies: score-level ensemble averaging and feature-level temporal concatenation with attentive pooling, both enhanced by LoRA-based parameter-efficient adaptation (rank=8). The proposed system achieves 66.34% average AUC/pAUC on the development set, demonstrating 10.47% improvement over reconstruction-based baselines.

5. REFERENCES

[1] Z. Lv, A. Jiang, B. Han, Y. Liang, Y. Qian, X. Chen, J. Liu, and P. Fan, "Aithu system for first-shot unsupervised anomalous sound detection," DCASE2024 Challenge, Tech. Rep., June 2024.

[2] A. Jiang, X. Zheng, Y. Qiu, W. Zhang, B. Chen, P. Fan, W.-Q. Zhang, C. Lu, and J. Liu, "Thuee system for first-shot unsupervised anomalous sound detection," DCASE2024 Challenge, Tech. Rep., June 2024.

[3] J. F. Gemmeke, D. P. 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. of ICASSP*. IEEE, 2017, pp. 776–780.

[4] 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," *In arXiv e-prints: 2506.10097*, 2025.

[5] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "Tf-gridnet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.

[6] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," vol. 35, 2022, pp. 28 708–28 720.

[7] 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. of ICML*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 5178–5193.

[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models." vol. 1, no. 2, 2022, p. 3.

[9] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, and M. Yasuda, "First-shot anomaly detection for machine condition monitoring: A domain generalization baseline," *Proc. of EUSIPCO*, pp. 191–195, 2023.

[10] Z. Pan, M. Ge, and H. Li, "A hybrid continuity loss to reduce over-suppression for time-domain target speaker extraction," in *Proc. of Interspeech*, 2022, pp. 1786–1790.

[11] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. of ICLR*, 2019.

[12] ——, "Sgdr: Stochastic gradient descent with warm restarts," in *Proc. of ICLR*, 2022.

[13] 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 *Proc. of DCASE*, Barcelona, Spain, November 2021, pp. 1–5.

[14] 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 *Proc. of DCASE*, Nancy, France, November 2022.