FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION FOR MACHINE CONDITION MONITORING BASED ON ENSEMBLE LEARNING AND DOMAIN GENERALIZATION

Technical Report

Ting Wu[1,2], *Lu Han*[1,2], *Zhaoli Yan*[3*], *Xiaobin Cheng*[1,2], *Jian Wen*[1,2], *Jun Yang*[1,2]

[1]Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Beijing University of Chemical Technology, Beijing, China
{wuting, hanlu2023, xb_cheng, wenjian, jyang}@mail.ioa.ac.cn
yanzl@mail.buct.edu.cn

ABSTRACT

Unsupervised pretrained models have achieved remarkable success across a wide range of applications. In this report, an approach is presented for DCASE 2025 Task 2: First-shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring. To address this challenge, an anomaly detection algorithm is proposed, which combines density estimation with cross-domain interpolation to robustly detect anomalies. Additionally, a two-stage pretraining strategy within a teacher-student framework is adopted to enhance audio data representation. A dual-headed network architecture is developed to leverage both labeled and unlabeled loss functions, mitigating the scarcity of labeled data. Finally, to optimize the ensemble of several large-scale models, an adaptive weighted combination perturbation search algorithm is introduced to determine the optimal fusion weights. Collectively, these methods achieve a score of 69.94% on the official development dataset, significantly surpassing the baseline model.

***Index Terms*—** Anomalous sound detection, Pretrained model, Transformer, Ensemble, KNN

1. INTRODUCTION

In recent years, anomalous sound detection (ASD) has emerged as a critical task within the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge [**?**, 1–7]. As a specialized branch of anomaly detection, ASD builds upon conventional algorithms by incorporating techniques specifically tailored to acoustic anomaly identification. Traditional approaches to anomaly detection include statistical methods such as Gaussian Mixture Models (GMM), distance-based methods, including k-Nearest Neighbors (k-NN), Local Outlier Factor (LOF), and Density-Based Spatial Clustering of Applications with Noise (DBSCAN), as well as clustering techniques such as K-means [9]. These methods typically operate by modeling normal data distributions and assigning anomaly scores to unknown samples based on their distance from or deviation relative to the learned distribution.

However, applying these methods directly to ASD presents two significant challenges. First, the high dimensionality of raw time-domain audio signals leads to the "curse of dimensionality," rendering direct modeling computationally infeasible and less effective. Second, anomalous patterns are often subtly embedded within the temporal structure of the signal, making it difficult to distinguish between normal and abnormal sounds. Even when using acoustic features such as short-time spectra or Mel-frequency cepstral coefficients (MFCCs), the extracted representations often fail to capture the nuanced characteristics of anomalies. As a result, traditional methods generally exhibit suboptimal performance when applied directly to ASD tasks.

Currently, several core algorithmic paradigms dominate the field of anomalous sound detection (ASD). The primary category includes generative methods that rely on reconstruction error, such as Autoencoders (AE), Variational Autoencoders (VAE), Generative Adversarial Networks (GAN), and diffusion models. These approaches typically model normal sound signals and compute anomaly scores by quantifying the discrepancy between the input and its reconstruction. However, such methods often demonstrate limited robustness against noise and domain shifts. Under conditions of low signal-to-noise ratio (SNR) or domain mismatch, even normal signals may deviate significantly from the learned distribution, resulting in increased false alarms and diminished discriminative performance [10].

To mitigate these limitations, auxiliary classification-based methods have garnered growing interest. Although anomalous samples are generally unavailable, metadata such as device identity and operational conditions associated with normal signals is often available. These methods construct auxiliary classification tasks using normal data, enabling the network to simultaneously compress high-dimensional time-domain inputs and extract semantically rich embeddings. Traditional anomaly detection techniques can then be applied to these embeddings. This strategy enhances the ability to capture intrinsic acoustic features and improves robustness to noise and domain shifts during training [11].

Despite these advantages, classification-based approaches face inherent limitations. Specifically, distributional gaps across different device types pose a significant challenge. While these methods perform well within the same device category, their generalization across different devices is limited, which restricts their generalizability across diverse operational conditions. In response, fine-

*Corresponding author: yanzl@mail.buct.edu.cn

tuning large-scale pretrained models has emerged as a promising solution, trained on diverse and extensive audio datasets, learn generalizable representations, thereby enhancing cross-domain performance. Nevertheless, they are not specifically optimized for targeted ASD applications. Fine-tuning can adapt pretrained models to specific domains, although adaptation is often limited to a subset of parameters. As a result, performance gains are modest, and detection accuracy often varies considerably across device categories. Motivated by these challenges, this study introduces a novel ASD algorithm that combines ensemble learning and domain generalization to improve performance and robustness in diverse acoustic environments.

Anomalous sound detection (ASD) continues to face several critical challenges. First, a significant domain shift persists: although the source domain contains abundant labeled data, the target domain often suffers from data scarcity, leading to suboptimal performance in cross-domain settings. Second, the absence of annotations related to specific operational conditions in the dataset limits the effectiveness of auxiliary classification objectives, which in turn degrades the quality of the learned embeddings. To address these issues, this study introduces and integrates the following key approaches:

1. **Density-Aware Domain Synthesized K-Nearest Neighbors (DADS-KNN):**
   To address distributional disparities between the source and target domains, we propose a novel DADS-KNN approach that integrates density estimation with cross-domain interpolation to facilitate robust cross-domain anomaly detection.
2. **Task-Adaptive Teacher-Student Pretraining (TATS):**
   By leveraging an auxiliary dataset of machine operation signals and large-scale pretrained models, we construct a teacher-student framework incorporating a task-adaptive pretraining strategy to enhance applicability to ASD-specific tasks.
3. **Adaptive Combination Perturbation (ACP):**
   To mitigate performance variability across large-scale pretrained models, we introduce the ACP algorithm, which dynamically optimizes ensemble weights via perturbation-based search, aimed at maximizing overall ensemble effectiveness.
4. **Two-Stage Dual-Head Semi-Supervised Learning (TS-DHSSL):**
   To overcome the absence of attribute-level annotations, we propose a TS-DHSSL framework that jointly learns from labeled and unlabeled data within a unified semi-supervised structure for improved representation learning under weak supervision.

2. METHOD

The proposed approach utilizes five distinct pretraining strategies. The first strategy involves full-parameter fine-tuning, wherein the backbone of a large-scale pretrained model is retained. An attentive statistical pooling layer is employed to compress the frequency dimension into a fixed-dimensional representation, followed by two fully connected layers for feature transformation. The classification head is supervised using the ArcFace loss function to enhance inter-class separability and intra-class compactness. The training schedule incorporates a warm-up stage and applies cosine annealing for learning rate scheduling to promote stable convergence and improve model generalization.

The second fine-tuning strategy involves augmented data training, in which pure noise and clean, background-free audio are incorporated as auxiliary categories associated with each device type. The classification task is subsequently retrained to include these additional categories, enabling the model to better distinguish truly anomalous signals from variations induced by noise or recording artifacts.

The third strategy employs Low-Rank Adaptation (LoRA), which introduces trainable low-rank decomposition modules into selected weight matrices, such as the attention layers of the pretrained model. By freezing the original model parameters and updating only the introduced low-rank matrices, this approach significantly reduces the number of trainable parameters and computational overhead, while preserving the model's representational capacity.

The fourth fine-tuning approach employs a dual-head network architecture with a two-stage training scheme. In the first stage, the network is trained using conventional supervised learning. Subsequently, clustering is performed on unlabeled data to produce pseudo-labels. In the second stage, labeled and unlabeled data are optimized independently using task-specific loss functions, which are then integrated into a unified objective for joint training.

The fifth approach adopts a secondary pretraining scheme based on a teacher–student framework. Historical datasets are consolidated to enable joint training of both teacher and student models, where the student model is updated via backpropagation, while the teacher model is updated through an exponential moving average (EMA) of the student's weights. This mechanism retains the original pretrained model's generalization ability on audio signals and mitigates the performance degradation typically caused by training on historical datasets.

To enhance anomaly detection, this study proposes the Density-Aware Domain Synthesized K-Nearest Neighbors (DADS-KNN) algorithm. This approach integrates target domain interpolation with local density information surrounding each sample during inference, resulting in a robust and adaptive detection framework.

To address the performance variability among large pretrained models, this study introduces the Adaptive Combination Perturbation (ACP) algorithm. Starting from a fixed ensemble of models, ACP iteratively refines the fusion weights via stochastic perturbations guided by performance feedback. A decay factor is incorporated into the perturbations to control the search range and improve stability. At each iteration, small perturbations are applied to the current weight vector, followed by normalization, and the configuration yielding the optimal fusion performance is retained. This process enables efficient and adaptive optimization of weight allocation within the fixed model ensemble.

3. MODEL

BEATs (Bidirectional Encoder representation from Audio Transformers) is an iterative self-supervised learning framework that jointly optimizes an acoustic tokenizer and an audio encoder to generate semantically rich discrete label predictions from audio data. The tokenizer is initially initialized with random projections and trained through masked prediction. Subsequently, semantic knowledge is distilled from a pretrained model to iteratively refine both the tokenizer and the audio encoder [12]. In this study, all five

Table 1: Source domain AUC, target domain AUC, pAUC, and harmonic mean of the BEATs, SSLAM, EAT and ResNet network under different training strategies.

Model	AUC(source)	AUC(target)	pAUC	hmean
BEATs	70.03	67.23	57.22	64.33
BEATs_all	69.98	67.77	55.53	63.76
BEATs_LoRA	71.16	63.31	55.69	62.75
BEATs_TATS	70.69	66.52	56.98	64.20
BEATs_TATS2	69.30	66.31	57.50	63.96
BEATs_TS	**73.11**	**69.82**	**58.12**	**66.36**
SSLAM	66.57	65.07	56.99	62.58
SSLAM_TATS	71.06	65.45	58.24	64.49
SSLAM_TATS2	72.26	64.71	58.26	64.58
SSLAM_TATS3	69.14	68.22	56.52	64.08
SSLAM_TS	67.81	64.44	55.41	62.10
EAT	54.23	62.76	53.86	56.67
EAT_TATS	70.75	67.95	57.65	64.94
EAT_TS	64.91	62.80	53.63	60.03
ResNet	67.71	67.97	57.60	64.05
ResNet_all	66.07	68.69	57.24	63.61

previously described training strategies were applied to the BEATs large model. The fifth strategy was further extended by incorporating datasets spanning multiple years, yielding a total of six distinct training approaches.

SSLAM (Self-Supervised Learning from Audio Mixtures) is a novel audio self-supervised learning framework specifically designed for complex multi-audio environments to enhance the model's capacity to represent and generalize across real polyphonic audio signals. Notably, SSLAM employs mixed audio inputs during training, substantially improving robustness and generalization in multi-source environments while maintaining performance on single-source tasks [13]. In this study, five training strategies are applied to the SSLAM model, specifically leveraging the first, fourth, and fifth methods described previously. The fifth method is further extended by incorporating multi-year datasets.

The Efficient Audio Transformer (EAT) is a computationally efficient self-supervised audio learning framework that employs a bootstrapping training paradigm to substantially reduce pretraining costs while preserving performance. This method introduces the novel Utterance-Frame Objective (UFO) to enhance acoustic event modeling and highlights the importance of masking strategies in audio self-supervision by proposing large-scale inverse block masking to enhance representation quality [14]. In this study, three training strategies—the first, fourth, and fifth methods—are applied to the EAT model.

ResNet, renowned for its strong feature extraction capabilities, has been extensively employed in anomalous sound detection. By utilizing residual connections, ResNet enables the effective training of deeper network architectures, facilitating the extraction of salient audio features from spectrograms and thereby improving both the accuracy and robustness of anomaly detection. In this study, the first and second training strategies were applied to ResNet.

The the above four methods evaluation results are presented in Table 1.

Incorporating the Adaptive Combination Perturbation (ACP) algorithm, the optimized weight parameters for the four systems submitted in this study are detailed in Table 2:

Table 2: Weight settings for different network models.

Model	System 1	System 2	System 3	System 4
BEATs	1	0.001	1	0.001
BEATs_all	1	0.001	1	0.058
BEATs_LoRA	1	0.001	1	0.010
BEATs_TATS	1	0.001	1	0.106
BEATs_TATS2	1	0.057	1	0.138
BEATs_TS	1	0.101	1	0.157
SSLAM	1	0.114	1	0.162
SSLAM_TATS	1	0.063	1	0.045
SSLAM_TATS2	1	0.163	1	0.010
SSLAM_TATS3	1	0.001	1	0.098
SSLAM_TS	1	0.098	1	0.001
EAT	1	0.001	1	0.001
EAT_TATS	1	0.189	1	0.205
EAT_TS	1	0.001	1	0.010
ResNet	1	0.215	0	0
ResNet_all	1	0.001	0	0

4. SUBMIT SYSTEM

The four submitted systems, as previously described, were evaluated using metrics including AUC and pAUC. Specifically, source domain AUC, target domain AUC, pAUC, and their harmonic mean were computed. Among these, the SSLAM and BEATs networks exhibited superior performance. The application of the proposed Adaptive Combination Perturbation (ACP) algorithm further revealed that SSLAM contributed most significantly to the overall score. Due to the limited generalization capability of the ResNet network, its weight contributions were excluded in Systems 3 and 4. Additionally, a simple averaging of all weights—a method known for its robustness—was also employed. Applying the ACP algorithm to optimize the weights across all networks, System 2 achieved the best performance, attaining a harmonic mean score of 69.94%. The corresponding evaluation results are summarized in Table 3.

Table 3: results of four submitted systems on the development set.

Machine + Metric	System 1	System 2	System 3	System 4
Bearing AUC (source)	63.24	67.52	63.48	68.16
Bearing AUC (target)	74.32	72.44	75.24	74.24
Bearing pAUC	62.26	63.11	62.74	63.00
Bearing hmean	66.18	67.47	66.69	68.16
Fan AUC (source)	62.76	64.60	62.52	64.96
Fan AUC (target)	57.16	58.80	57.40	60.08
Fan pAUC	54.42	55.42	53.63	55.63
Fan hmean	57.91	59.37	57.62	59.98
Gearbox AUC (source)	77.68	76.08	78.08	76.08
Gearbox AUC (target)	89.00	90.32	84.68	84.80
Gearbox pAUC	74.05	74.26	71.11	71.84
Gearbox hmean	79.76	79.61	77.56	77.21
Slider AUC (source)	81.40	80.04	81.44	80.44
Slider AUC (target)	63.08	63.52	62.24	61.60
Slider pAUC	56.26	55.74	56.32	56.26
Slider hmean	65.34	64.97	65.07	64.60
ToyCar AUC (source)	61.20	69.20	61.20	66.68
ToyCar AUC (target)	73.12	72.68	74.56	75.04
ToyCar pAUC	54.95	59.84	54.68	58.33
ToyCar hmean	62.22	66.78	62.45	66.07
ToyTrain AUC (source)	81.76	81.60	82.20	82.60
ToyTrain AUC (target)	74.76	74.32	74.32	74.84
ToyTrain pAUC	58.42	58.84	58.11	59.21
ToyTrain hmean	70.22	70.25	70.04	70.83
Valve AUC (source)	95.16	99.56	93.76	96.16
Valve AUC (target)	85.20	82.92	81.12	77.36
Valve pAUC	76.05	85.53	72.37	76.89
Valve hmean	84.76	88.77	81.50	82.57
Overall AUC (source)	72.93	75.53	72.91	75.15
Overall AUC (target)	72.29	72.24	71.59	71.61
Overall pAUC	61.31	63.24	60.49	62.21
Overall Office score	68.41	69.94	67.85	69.21

5. CONCLUSION

This paper proposes an ensemble and domain-generalized framework for anomalous sound detection. It introduces a KNN-based anomaly detection algorithm that integrates density estimation with cross-domain interpolation. To leverage related audio data, a Task-Adaptive Teacher-Student (TATS) pretraining strategy is developed, which preserves the pretrained model's generalization capability while enhancing detection accuracy. For data lacking attribute labels, a two-stage dual-head semi-supervised network is employed, effectively combining losses from both labeled and unlabeled samples to outperform the baseline. Finally, an Adaptive Combination Perturbation (ACP) algorithm is presented to dynamically optimize the weighting of multiple pretrained models in the ensemble. Integrating these components, the proposed system achieves a final harmonic mean score of 69.94%.

6. REFERENCES

[1] 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. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Barcelona, Spain, Nov. 2021, pp. 1–5.

[2] 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. 7th DCASE Workshop (DCASE2022)*, Nancy, France, Nov. 2022.

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

[4] Y. Koizumi, Y. Kawaguchi, K. Imoto, *et al.*, "Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2006.05822*, 2020.

[5] Y. Kawaguchi, K. Imoto, Y. Koizumi, *et al.*, "Description and discussion on dcase2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions," *arXiv preprint arXiv:2106.04492*, 2021.

[6] K. Dohi, K. Imoto, N. Harada, *et al.*, "Description and discussion on dcase2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2305.07828*, 2023.

[7] 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 dcase2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2406.07250*, 2024.

[8] ——, "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.

[9] K. Wilkinghoff and F. Kurth, "Why do angular margin losses work well for semi-supervised anomalous sound detection?" *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2023.

[10] X. Yang, F. Lv, F. Liu, *et al.*, "Self-training vision language berts with a unified conditional model," *IEEE Trans. Circuits Syst. Video Technol.*, 2023.

[11] K. Wilkinghoff, "Self-supervised learning for anomalous sound detection," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 276–280.

[12] S. Chen, Y. Wu, C. Wang, *et al.*, "Beats: Audio pre-training with acoustic tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.

[13] T. Alex, S. Atito, A. Mustafa, *et al.*, "Sslam: Enhancing self-supervised models with audio mixtures for polyphonic soundscapes," in *Proc. 13th Int. Conf. on Learning Representations (ICLR)*, 2025.

[14] W. Chen, Y. Liang, Z. Ma, *et al.*, "Eat: Self-supervised pre-training with efficient audio transformer," *arXiv preprint arXiv:2401.03497*, 2024.