DUAL-CHANNEL ATTRIBUTE FINE-TUNING OF PRE-TRAINED MODELS FOR NOISE-AWARE ANOMALOUS SOUND DETECTION

Technical Report

Jie Yang

HuaiBei, China
852549193@qq.com

ABSTRACT

This report details our technical approach to the DCASE 2026 Challenge Task 2, which focuses on Noise-aware Unsupervised Anomalous Sound Detection (UASD) for machine condition monitoring. In real-world scenarios, obtaining clean recordings of target machines or isolated background noise is often difficult because machines cannot be easily stopped. To provide a practical alternative for developing noise-robust systems under such constraints, the 2026 dataset introduces synchronous two-channel audio samples captured at locations near and far from the target machine, where the distant microphone serves as a noise reference.

Our implemented system utilizes the Efficient Audio Transformer (EAT) base model, pre-trained on AudioSet-2M, as the robust feature extraction backbone. To exploit the dual-channel nature of the data, the two channels are explicitly separated to double the training data volume and zero-padded to a uniform length. The EAT model extracts deep features, followed by Attentive Statistics Pooling (ASP) for dimensionality reduction. We construct a unified composite label for each sample encompassing its machine type, intrinsic attribute, domain, and the designated "near" or "far" spatial tag. The backbone is then fine-tuned using the ArcFace loss function to maximize intra-class compactness. For anomaly scoring, a K-Nearest Neighbors (KNN) model is employed in the latent space. We submitted four systems combining different fine-tuning strategies and inference approaches. Evaluated on the development set, the proposed method yields a source domain AUC of 66.70%, a target domain AUC of 68.48%, and a pAUC of 59.05%.

***Index Terms*—** Anomalous sound detection, noise-aware, Efficient Audio Transformer, Attentive Statistics Pooling, LoRA, KNN

1. INTRODUCTION

Unsupervised Anomalous Sound Detection (UASD) aims to identify whether a machine is operating normally or anomalously by learning only from normal operating sounds. The DCASE 2026 Challenge Task 2 introduces the critical challenge of noise-aware anomaly detection [1]. Real-world industrial environments frequently contain diverse and non-stationary background noise [2, 3]. While previous challenges provided clean target sounds or pure noise recordings to improve noise robustness, acquiring such data in advance requires stopping the target machines or noise sources, which is often infeasible in practical industrial settings. To address this limitation, the DCASE 2026 dataset introduces a more applicable setting: synchronous two-channel recordings captured by microphones placed near to and far from the target machine. Since the distant microphone captures less direct machine sound and relatively stronger environmental noise, it provides essential physical cues to identify noise components without requiring machine downtime.

Our baseline architecture builds upon the top-performing solution from the DCASE 2025 Challenge Task 2 [5], which demonstrated the powerful generalization capabilities of pre-trained audio models. To tackle the novel challenges of this year's task, we adapt and expand this foundation by introducing a dual-channel attribute augmentation strategy specifically designed for the 2026 data characteristics. By separating the channels and utilizing a unified composite label, we compel the backbone to explicitly learn noise-robust representations. We then detail the formulation of four submitted systems, which explore different weight adaptation techniques and varying KNN backend scoring strategies.

2. SYSTEM METHODOLOGY

Our implemented system pipeline consists of feature extraction using the EAT backbone coupled with Attentive Statistics Pooling (ASP), dual-channel data augmentation via a unified composite label, advanced fine-tuning using ArcFace loss, and KNN-based anomaly scoring.

2.1. Feature Extraction Backbone

To extract comprehensive acoustic features, we utilize the Efficient Audio Transformer (EAT) [4]. EAT is crafted for self-supervised audio learning, dedicated to efficient representation learning from unlabeled audio data through a customized bootstrap training strategy. The input audio waveforms are first converted into Mel-spectrogram features. These two-dimensional features are then split into localized patches. Each patch is linearly projected into an embedding and processed by the transformer encoder layers to capture deep, context-aware temporal and spectral dependencies.

In our implementation, we utilize the EAT base model pre-trained on AudioSet-2M, which comprises approximately 88 million parameters. To construct a fixed-length, utterance-level representation suitable for distance metric calculations, we employ Attentive Statistics Pooling (ASP) [5]. ASP effectively reduces the dimensionality and aggregates the frame-level transformer patch embeddings into a singular, low-dimensional robust feature vector for each audio clip, highlighting the most informative temporal segments.

2.2. Composite Label Formulation and Objective

To maximize the utility of the dataset's spatial information without introducing complex multi-channel network architectures, we employ a dual-channel attribute augmentation strategy. Each synchronous dual-channel recording is separated into two independent single-channel samples, explicitly tagged with a "near" or "far" pseudo-attribute label, thereby doubling the training data volume.

We formulate a single composite label for each sample to encapsulate its comprehensive state. This unified label is constructed by concatenating the sample's machine type, intrinsic attributes, domain (source/target), and the near/far spatial tag. Consequently, each unique combination of these parameters forms a distinct class label.

We apply the ArcFace loss function [6] to the low-dimensional embeddings generated by ASP:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))}+\sum_{j=1,j\neq y_i}^{c}e^{s\cos\theta_j}} \quad (1)$$

where y_i is the composite label of sample i, s is the scaling factor, and m is the angular margin. θ_j represents the angle between the embedding of sample i and the registered embedding of the j-th class. By optimizing this loss against the unified composite labels, the network learns to actively pull samples with identical specific conditions (including spatial distance) closer together, intrinsically achieving spatial feature disentanglement and enhancing inter-class discrepancy.

2.3. Model Fine-Tuning Strategies

To adapt the pre-trained generalized acoustic representations to the specific industrial domain of DCASE Task 2, we implement two distinct parameter optimization strategies:

2.3.1. Full Fine-Tuning

Full Fine-Tuning is an optimization method that adapts a pre-trained model to specific tasks or domain requirements by adjusting all of its parameters. In our implementation, we fine-tuned all parameters of the EAT backbone and ASP layers simultaneously. The model is trained for 20,000 steps using the AdamW optimizer with a maximum learning rate of 5e-5 and a batch size of 16. While highly effective at capturing domain-specific nuances, Full Fine-Tuning alters the entire pre-trained weight space to fit the target dataset.

2.3.2. LoRA Fine-Tuning

In addition to Full Fine-Tuning, we adopt the Low-Rank Adaptation (LoRA) method [7] for model parameter tuning, which we refer to as EAT-LORA. LoRA freezes the pre-trained model weights and injects trainable rank decomposition matrices into specific layers. This allows the model to adapt to new tasks while maintaining most of the pre-trained knowledge.

LoRA's core idea is to approximate parameter updates using low-rank matrices, significantly reducing the number of trainable parameters. For the weight matrix W of a pre-trained model, LoRA decomposes its update into the product of two low-rank matrices A and B:

$$W = W_0 + \Delta W = W_0 + B \cdot A \cdot \alpha \quad (2)$$

where W_0 represents the pre-trained weight matrix, B and A are matrices of dimensions $d \times r$ and $r \times k$ respectively ($r \ll \min(d, k)$), and α is a scaling factor used to adjust the magnitude of the update. This approach drastically improves training efficiency while preserving the robust representational capacity of the pre-trained model.

2.4. KNN Anomaly Scoring Backend

For the anomaly detection phase, we employ K-Nearest Neighbors (KNN) [8] as the backend scorer. We construct a reference library $\mathcal{E}_{norm}$ comprising the low-dimensional embeddings of normal training samples.

For a given test sample x_{test}, we extract its embedding $E(x_{test})$ via the fine-tuned EAT-ASP backbone. The anomaly score $\mathcal{A}(x_{test})$ is calculated based on the cosine distance to its nearest neighbor ($K = 1$) in the normal library:

$$\mathcal{A}(x_{test}) = 1 - \max_{E_n \in \mathcal{E}_{norm}} \frac{E(x_{test}) \cdot E_n}{\|E(x_{test})\|\|E_n\|} \quad (3)$$

Anomalous sounds will lie further away from the normal feature cluster, resulting in a lower maximum cosine similarity and consequently a higher anomaly score.

2.5. Submitted Systems

Based on the methodologies described above, we constructed and submitted four distinct systems to the challenge, evaluating different combinations of fine-tuning techniques and KNN scoring mechanisms.

System 1 (EAT-Fusion) utilizes the Full Fine-Tuning strategy but adopts a multi-channel evaluation approach. In the inference stage, KNN anomaly scores are calculated independently for both the "near" and "far" channels, and the final anomaly score is obtained via score-level averaging [9].

System 2 (EAT-LORA-Fusion) employs the LoRA strategy for parameter updates. Similar to System 1, the final anomaly score is generated through the score-level fusion [9] of both the "near" and "far" channel KNN outputs, aiming to capture complementary spatial anomaly cues.

System 3 (EAT-Near) utilizes the Full Fine-Tuning strategy. During the KNN scoring phase, only the embeddings from the "near" channel are used to construct the normal library and evaluate test samples. This approach strictly isolates the evaluation space from distant environmental noise.

System 4 (EAT-LORA-Near) employs the Low-Rank Adaptation (LoRA) strategy to update the EAT backbone efficiently. It relies exclusively on the "near" channel for KNN anomaly scoring to minimize background interference while preserving the pre-trained knowledge base.

3. EXPERIMENTAL SETUP

To address the DCASE 2026 Task 2 challenge, our systems were developed and evaluated using the official development dataset [1]. The dataset comprises seven machine types: fan, gearbox (Emu), bearing (Emu), slide rail (Emu), valve (Emu), ToyCar, and ToyCar (Emu). For each machine section, the provided training data consists of 990 normal clips from a source domain and 10 normal clips from a target domain.

The audio samples provided in the dataset vary in duration from 6 to 16 seconds at a sampling rate of 16 kHz. To ensure uniform

input dimensions and preserve temporal continuity without distorting acoustic features, all shorter audio clips were zero-padded to align with the maximum sample length of 16 seconds prior to feature extraction. The length-normalized raw waveforms were then converted into log-Mel spectrograms before being fed into the EAT model. All experiments were conducted using PyTorch on a single NVIDIA RTX 4090 GPU.

4. EXPERIMENTAL RESULTS

The performance of our developed systems was evaluated on the official development set using the Area Under the Receiver Operating Characteristic Curve (AUC) for source and target domains, and partial AUC (pAUC, $p = 0.1$). We compared our best-performing submitted configuration against the two official baselines: Autoencoder with Mean Squared Error (AE-MSE) and Autoencoder with Mahalanobis distance [10].

Table 1: Anomaly detection results on the development set.

Method		Baseline MSE	Baseline MAHALA	Best Submitted System
bearingEmu	AUC(source)	62.34%	65.92%	59.86%
	AUC(target)	59.56%	62.28%	57.28%
	pAUC	59.85%	60.42%	51.74%
	score	60.56%	62.79%	56.08%
fan	AUC(source)	61.45%	60.00%	55.38%
	AUC(target)	46.94%	45.09%	66.74%
	pAUC	53.33%	52.29%	60.58%
	score	53.26%	51.75%	60.55%
gearboxEmu	AUC(source)	68.23%	74.48%	80.44%
	AUC(target)	49.78%	52.74%	77.92%
	pAUC	52.94%	53.97%	64.21%
	score	55.94%	58.92%	73.46%
sliderEmu	AUC(source)	67.25%	66.36%	55.14%
	AUC(target)	45.05%	49.18%	54.16%
	pAUC	50.38%	50.36%	49.21%
	score	52.71%	54.29%	52.70%
ToyCar	AUC(source)	75.62%	77.28%	75.30%
	AUC(target)	37.87%	53.17%	71.36%
	pAUC	54.03%	58.25%	63.11%
	score	51.60%	61.33%	69.54%
ToyCarEmu	AUC(source)	69.62%	69.49%	75.82%
	AUC(target)	61.20%	66.62%	82.52%
	pAUC	55.89%	53.47%	61.53%
	score	61.73%	62.37%	72.19%
valveEmu	AUC(source)	67.74%	56.60%	76.10%
	AUC(target)	68.78%	56.50%	80.94%
	pAUC	55.08%	50.20%	67.95%
	score	56.66%	54.26%	74.60%
All (hmean)	AUC(source)	67.19%	66.46%	66.70%
	AUC(target)	50.85%	54.24%	68.48%
	pAUC	54.36%	53.91%	59.05%
	score	56.66%	57.66%	64.49%

As detailed in Table 1, our approach significantly outperforms the official baselines, particularly in domain generalization capabilities. While maintaining a comparable Source Domain AUC (66.70%) to the baselines, our best system achieves a Target Domain AUC of 68.48%, marking a substantial improvement over the AE-MSE (48.73%) and Mahalanobis (54.24%) baselines. Furthermore, the overall pAUC improved to 59.05%, bringing the final overall harmonic mean score to 64.49%.

Notably, for machine types like *fan*, *gearboxEmu*, and *valveEmu*, the Target AUC sees dramatic increases. This confirms that formulating a combined composite label for ArcFace fine-tuning, alongside strategically managing inference channels via KNN, successfully mitigates the negative impact of distant environmental noise in the evaluation space.

5. CONCLUSION

This report detailed our technical implementation for DCASE 2026 Task 2. By leveraging the Efficient Audio Transformer (EAT) coupled with Attentive Statistics Pooling (ASP), we ensured strong baseline feature extraction. We heavily exploited the synchronous multi-channel nature of the dataset by splitting dual-channel recordings and zero-padding them to a uniform length. By concatenating the sample attributes, domains, and spatial distances into a single composite label, we constrained the network to build noise-aware representations via ArcFace loss optimization. We submitted four systems to evaluate the interactions between parameter adaptation strategies (Full Fine-Tuning and LoRA) and inference scoring mechanisms (dual-channel fusion vs. near-channel exclusive KNN). The experimental results demonstrate that our methodology provides robust anomaly detection performance under severe industrial noise shifts.

6. REFERENCES

[1] T. Nishida, N. Harada, D. Niizumi, et al., "Description and discussion on DCASE 2026 challenge task 2: first-shot unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2606.10097*, 2026.

[2] 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," *arXiv preprint arXiv:2106.02369*, 2021.

[3] 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," *arXiv preprint arXiv:2205.13879*, 2022.

[4] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "EAT: Self-supervised pre-training with efficient audio transformer," in *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, 2024.

[5] L. Wang, "Pre-trained model enhanced anomalous sound detection system for dcase2025 task2," *Detection and Classification of Acoustic Scenes and Events*, 2025.

[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[7] 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.," *Iclr*, vol. 1, no. 2, p. 3, 2022.

[8] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 427–438.

[9] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 8, no. 4, p. e1249, 2018.

[10] 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 *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 2023, pp. 191–195.