DISTILLATION GUIDED TWO-CHANNEL BEATS SYSTEM FOR DCASE2026 TASK 2

Technical Report

Juncai Yang

Nanjing University
jc_yang@smail.nju.edu.cn

ABSTRACT

This technical report describes our submitted system for DCASE2026 Challenge Task2, Noise-aware Unsupervised Anomalous Sound Detection for Machine Condition Monitoring. The system is based on a pre-trained BEATs representation and a pseudo-clean distillation strategy for two-channel machine-sound recordings. In the first stage, ILRMA-based blind source separation is applied to the two-channel recordings to obtain pseudo-clean waveforms, which are used to adapt a BEATs teacher model. In the second stage, the fixed teacher supervises a student model trained on the original two-channel recordings. The student model consists of a neural enhancement front-end, a shared BEATs encoder, an enhanced-mixture sequence fusion module, and attentive statistics pooling. Utterance-level embeddings extracted by the student model are finally scored by a KNN backend with cosine distance. The submitted system achieves an official score of 62.27% on the development set.

***Index Terms*—** Anomalous sound detection, machine condition monitoring, BEATs, two-channel audio, pseudo-clean distillation, KNN

1. INTRODUCTION

DCASE2026 Challenge Task2 [1, 2, 3, 4] focuses on noise-aware unsupervised anomalous sound detection for machine condition monitoring. The task follows the first-shot unsupervised ASD setting of previous DCASE Task2 challenges, while introducing synchronized two-channel recordings captured at different distances from the target machine. The main task characteristics are summarized as follows:

- **Unsupervised ASD:** only normal machine sounds are available for training, and unknown anomalous sounds are detected at test time.
- **Domain generalization:** the system should be robust to distribution shifts caused by operating states, machine attributes, recording environments, and background noise.
- **Unseen machine types:** the machine types in the development and evaluation sets are different.
- **Attribute availability:** systems should work for machines with and without attribute information.
- **Two-channel recording:** both training and inference use synchronized near- and far-microphone recordings.

The two-channel setting provides useful acoustic information for noise-aware representation learning. The near-channel signal usually contains stronger target-machine components, while the far-channel signal reflects different interference and recording characteristics. In this work, we exploit this setting through a pseudo-clean teacher-student framework.

The proposed system first constructs a pseudo-clean teacher using blind-source-separated signals. The fixed teacher is then used to supervise a student model trained on the original two-channel recordings. The student combines a neural enhancement front-end and an enhanced-mixture sequence fusion module. At inference time, only the student model is used, and anomaly scores are computed by a KNN backend in the embedding space.

2. PROPOSED ASD SYSTEM

2.1. System architecture

The submitted system consists of a pseudo-clean teacher, a two-channel student model, and a KNN backend. Figure 1 shows the overall framework.

In the first stage, a BEATs [5] teacher is adapted using pseudo-clean waveforms derived from the two-channel recordings. In the second stage, the fixed teacher provides response-level supervision for a student model trained on the original two-channel recordings. The student model is used to extract embeddings for the submitted anomaly scores.

2.2. Backbone

We employ BEATs as the acoustic representation backbone. Given a waveform x, the BEATs encoder extracts a sequence of acoustic tokens,

$$\mathbf{H} = f_{\mathrm{B}}(x), \tag{1}$$

where $f_{\mathrm{B}}(\cdot)$ denotes the BEATs encoder. The token sequence is summarized by attentive statistics pooling (ASP) [6], followed by a projection layer:

$$\mathbf{z} = f_{\mathrm{p}}\left(\mathrm{ASP}(\mathbf{H})\right). \tag{2}$$

LoRA [7] is used for parameter-efficient adaptation of the pre-trained BEATs model. During training, a sub-center classification head is attached to the embedding layer as an auxiliary objective. The classification head is discarded during anomaly scoring.

2.3. Pseudo-clean teacher adaptation

For each two-channel recording, ILRMA-based blind source separation [8] is applied to obtain a pseudo-clean waveform x_{pc}. The separated waveform is used as an auxiliary acoustic view for adapting the teacher model.

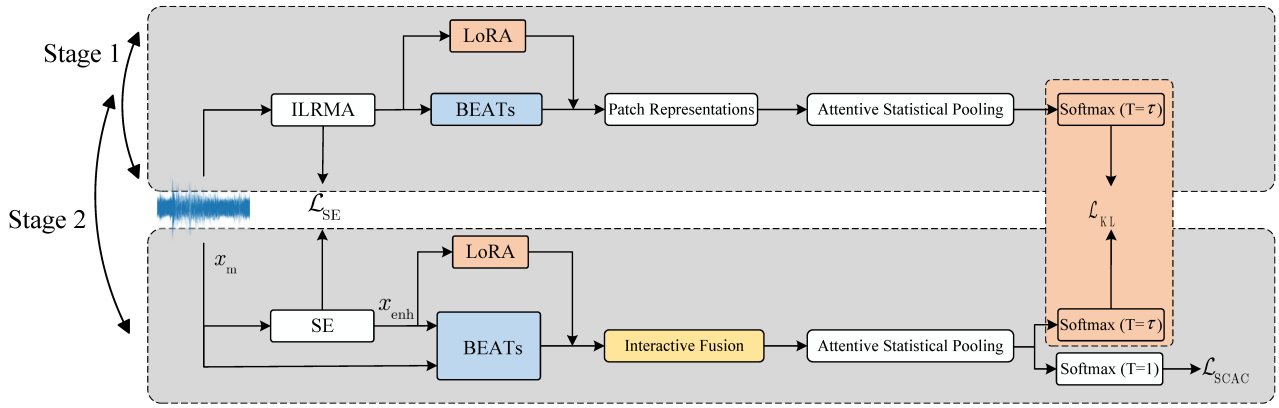

Figure 1: Overview of the proposed pseudo-clean distillation guided two-channel BEATs system.

The teacher embedding is computed as

$$\mathbf{z}_{\mathrm{tea}} = f_{\mathrm{p}}^{\mathrm{tea}}\left(\mathrm{ASP}\left(f_{\mathrm{B}}^{\mathrm{tea}}(x_{\mathrm{pc}})\right)\right), \tag{3}$$

where the superscript tea denotes the teacher branch. The teacher is trained with the auxiliary sub-center classification objective. After the first stage, the teacher parameters are fixed.

2.4. Two-channel student model

The student model is trained using the original two-channel recording. A neural enhancement front-end estimates an enhanced waveform x_{enh} from the two-channel input. In parallel, the near-channel mixture is retained as a mixture-reference signal x_{m}. The enhanced signal and the mixture-reference signal are encoded by a shared BEATs student encoder:

$$\mathbf{H}_{\mathrm{enh}} = f_{\mathrm{B}}^{\mathrm{stu}}(x_{\mathrm{enh}}), \quad \mathbf{H}_{\mathrm{m}} = f_{\mathrm{B}}^{\mathrm{stu}}(x_{\mathrm{m}}), \tag{4}$$

where the superscript stu denotes the student branch.

The two token sequences are combined by a sequence-level fusion module:

$$\mathbf{H}_{\mathrm{fus}} = f_{\mathrm{fus}}\left(\mathbf{H}_{\mathrm{enh}}, \mathbf{H}_{\mathrm{m}}\right). \tag{5}$$

The student embedding is obtained as

$$\mathbf{z}_{\mathrm{stu}} = f_{\mathrm{p}}^{\mathrm{stu}}\left(\mathrm{ASP}(\mathbf{H}_{\mathrm{fus}})\right). \tag{6}$$

The enhanced branch is used to suppress acoustic interference, while the mixture-reference branch preserves the original machine-sound characteristics. Their sequence-level fusion provides the representation used for anomaly scoring.

2.5. Distillation-based fine-tuning

During the second stage, the fixed teacher receives x_{pc}, while the student receives the original two-channel recording. Let $\mathbf{o}_{\mathrm{tea}}$ and $\mathbf{o}_{\mathrm{stu}}$ denote the teacher and student logits, respectively. Their softened posterior distributions are defined as

$$p_{\mathrm{tea}}^{(\tau)} = \mathrm{softmax}\left(\frac{\mathbf{o}_{\mathrm{tea}}}{\tau}\right), \quad p_{\mathrm{stu}}^{(\tau)} = \mathrm{softmax}\left(\frac{\mathbf{o}_{\mathrm{stu}}}{\tau}\right), \tag{7}$$

where τ is the distillation temperature. The student training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{SCAC}} + \lambda_{\mathrm{KD}}\tau^2 \mathrm{KL}\left(p_{\mathrm{tea}}^{(\tau)} \| p_{\mathrm{stu}}^{(\tau)}\right) + \lambda_{\mathrm{SE}}\mathcal{L}_{\mathrm{SE}}. \tag{8}$$

Here, $\mathcal{L}_{\mathrm{SCAC}}$ is the student classification loss [9], $\mathcal{L}_{\mathrm{SE}}$ is the waveform-level enhancement loss with respect to the pseudo-clean waveform, and the KL-divergence term transfers the teacher response to the student [10]. The teacher is not updated in this stage.

2.6. KNN backend

For anomaly scoring, embeddings of normal recordings are stored as machine-wise reference libraries. Given a test recording x, the cosine distance between its student embedding $\mathbf{z}_{\mathrm{stu}}(x)$ and a reference embedding $\mathbf{z}_i$ is

$$d(\mathbf{z}_{\mathrm{stu}}(x), \mathbf{z}_i) = 1 - \frac{\mathbf{z}_{\mathrm{stu}}(x)^\top \mathbf{z}_i}{\|\mathbf{z}_{\mathrm{stu}}(x)\|_2 \|\mathbf{z}_i\|_2}. \tag{9}$$

The anomaly score is calculated by a KNN backend [11]:

$$\mathcal{A}(x) = \frac{1}{k} \sum_{\mathbf{z}_i \in \mathcal{N}_k(x)} d(\mathbf{z}_{\mathrm{stu}}(x), \mathbf{z}_i), \tag{10}$$

where $\mathcal{N}_k(x)$ denotes the set of k nearest reference embeddings. We use $k = 1$ in the submitted system.

2.7. Implementation details

The training segment length is set to 12 s. For embedding extraction, valid-length masks are used to prevent padded samples from contributing to the utterance-level representation. For the BEATs frontend, 128-dimensional filter-bank features are extracted using a 25-ms frame length, a 10-ms frame shift, and a Hamming window. The projection dimension of the BEATs backend is set to 128, and the hidden dimension of attentive statistics pooling is also set to 128. LoRA adaptation is applied to the query and value projection modules of BEATs, with rank 64. The sub-center classification head uses 16 sub-centers per class during training. For anomaly scoring, we use a KNN backend with $k = 1$ and cosine distance. Score normalization is applied to the final anomaly scores of the development-selected submitted system.

2.8. Submitted systems

We submitted two variants based on the same pseudo-clean distillation guided BEATs framework. The two variants share the same

Table 1: Development-set performance for each machine type.

Machine	Metric	Baseline		Our system
		Mahala	MSE	
bearingEmu	AUC (source)	**64.16%**	61.56%	58.58%
	AUC (target)	58.80%	57.70%	**58.90%**
	pAUC	**60.16%**	59.63%	52.42%
fan	AUC (source)	60.88%	61.32%	**74.10%**
	AUC (target)	45.40%	47.02%	**61.18%**
	pAUC	52.37%	53.26%	**59.42%**
gearboxEmu	AUC (source)	**75.94%**	70.44%	68.02%
	AUC (target)	51.60%	49.02%	**80.70%**
	pAUC	52.79%	51.79%	**57.21%**
sliderEmu	AUC (source)	**70.28%**	64.60%	64.90%
	AUC (target)	47.20%	40.70%	**59.68%**
	pAUC	49.84%	50.53%	**52.11%**
ToyCar	AUC (source)	**79.12%**	76.34%	75.06%
	AUC (target)	53.30%	37.04%	**88.52%**
	pAUC	58.68%	53.63%	**65.84%**
ToyCarEmu	AUC (source)	67.88%	**74.22%**	55.74%
	AUC (target)	71.50%	66.08%	**78.98%**
	pAUC	52.84%	**56.63%**	48.89%
valveEmu	AUC (source)	54.40%	**69.90%**	60.60%
	AUC (target)	56.20%	69.60%	**70.16%**
	pAUC	49.26%	**55.11%**	50.21%
All (hmean)	AUC (source)	66.56%	**67.89%**	65.44%
	AUC (target)	53.78%	49.97%	**69.53%**
	Official score	57.33%	56.41%	**62.27%**

model architecture, feature configuration, and KNN scoring backend, and differ in checkpoint selection and inference-time scoring protocol.

System 1 uses the checkpoint from the last training epoch and computes one file-level anomaly score for each recording. Valid-length masks are used during embedding extraction to prevent padded samples from affecting the utterance-level representation.

System 2 uses the checkpoint selected according to the development-set performance. It follows the same embedding extraction and KNN scoring framework, but applies an additional multi-crop scoring strategy for ToyDrone recordings, whose duration is longer than the training segment length. The crop-level anomaly scores are aggregated into one file-level score.

In the following results, we report the development-set performance of System 2, which is the development-selected submitted system.

3. RESULTS

The systems are evaluated in terms of source-domain AUC, target-domain AUC, pAUC, and the official score. Table 1 compares the development-set performance of the proposed system with two baseline scoring methods.

The proposed system substantially improves the harmonic-mean target-domain AUC and the official score over the two baseline scoring methods. In particular, large target-domain AUC gains are observed for ToyCar, gearboxEmu, fan, and sliderEmu, suggesting that the pseudo-clean distillation guided two-channel representation is effective under domain-shifted noise conditions. In contrast, the harmonic-mean source-domain AUC is not improved over the strongest baseline, indicating that the submitted system mainly benefits target-domain generalization rather than uniformly improving all evaluation metrics.

4. CONCLUSION

This report presents a pseudo-clean distillation guided two-channel BEATs system for DCASE2026 Task2. The system first adapts a teacher model using pseudo-clean waveforms derived from two-channel recordings. A student model is then trained on the original recordings with teacher-student distillation, neural enhancement, and enhanced-mixture sequence fusion. The final anomaly score is obtained by a KNN backend in the student embedding space. The proposed framework exploits the two-channel noise-aware setting while retaining an inference procedure based on the original recordings. The development-set results show that the system mainly improves target-domain robustness and the official score.

5. 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," *In arXiv e-prints: 2606.01578*, 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," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Barcelona, Spain, November 2021, pp. 1–5.

[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," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

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

[5] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.

[6] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. Interspeech 2018*, 2018, pp. 2252–2256.

[7] X. Zheng, A. Jiang, B. Han, Y. Qian, P. Fan, J. Liu, and W.-Q. Zhang, "Improving anomalous sound detection via low-rank adaptation fine-tuning of pre-trained audio models," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 969–974.

[8] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: Two converging routes to ilrma originating from ica and nmf," *APSIPA Transactions on Signal and Information Processing*, vol. 8, p. e12, 2019.

[9] K. Wilkinghoff, "Sub-cluster adacos: Learning representations for anomalous sound detection," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[10] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[11] 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.