FOUR SYSTEMS FOR NOISE-AWARE UNSUPERVISED ANOMALOUS SOUND DETECTION IN DCASE 2026 TASK 2

Technical Report

Tsz Fhl

HFU
HEFEI, China
616901109@qq.com

ABSTRACT

This technical report describes four submissions to DCASE 2026 Task 2 [1]: Tsz_DCASE_task2_baseline (official autoencoder baseline with Mahalanobis scoring), Tsz_DCASE_task2_1 (stereo discriminative GeoSNet with BEATs and prototype scoring), Tsz_DCASE_task2_fusion (fine-tuned BEATs with ArcFace and rank-min fusion), and Tsz_DCASE_task2_sera (SERA-style ASP with frozen BEATs and multi-task ArcFace). All systems operate on 16 kHz stereo recordings under domain shift [2, 3, 4], use per-machine or global training on development and eval-additional data, and apply gamma-fit decision thresholds (q=0.9) on train-normal scores. Development-set results are reported for architecture validation; eval submissions provide anomaly scores and binary decisions for five official machine types.

***Index Terms*—** Anomalous sound detection, domain shift, BEATs, autoencoder, prototype scoring, DCASE

1. INTRODUCTION

DCASE 2026 Task 2 targets noise-aware unsupervised anomalous sound detection (ASD) for machine condition monitoring under domain shift with stereo microphone recordings [1]. Training data are drawn from the ToyADMOS2 [2] and MIMII DG [3] corpora; the first-shot evaluation setting follows [4]. We submit four systems: Tsz_DCASE_task2_baseline (AE_baseline, official AE with MAHALA scoring), Tsz_DCASE_task2_1 (P_proto, stereo CNN+BEATs with prototype scoring), Tsz_DCASE_task2_fusion (T_fusion, BEATs ArcFace with rank-min fusion), and Tsz_DCASE_task2_sera (sera_ft1, ASP with frozen BEATs).

All systems process 16 kHz waveforms cropped to 10 s clips. Binary decisions follow the official protocol [1, 5]: a gamma distribution is fit to train-normal anomaly scores and the threshold is set at the 90th percentile (q=0.9). Eval submissions contain anomaly scores and decisions for five machine types (Sander, ToyDrone, BlowerDustCollector, SewingMachine, ToothBrush). Each eval machine uses its own additional training split.

2. SYSTEM DESCRIPTION

2.1. Tsz_DCASE_task2_baseline (AE_baseline)

We implement the official DCASE2026 autoencoder (AE) baseline [6] without modifying the core architecture. The frontend stacks five log-mel frames (128 mel bins, FFT 1024, hop 512) into a 640-dimensional vector per clip. Each machine type has an independent AE trained for 100 epochs on normal clips only (batch 256, learning rate 10^{-3}, 10% validation split). At test time, Mahalanobis distance (MAHALA) is computed in the AE latent space using source- and target-domain covariance estimates; the per-clip score is the minimum of the two domain distances. The system has approximately 1 M parameters per machine.

2.2. Tsz_DCASE_task2_1 (P_proto)

This system combines a trainable CNN with multi-scale feature extraction (MSFE) on stereo log-mel energies and spatial cues (ILD, ITD, IPD, mid–side summaries) with a frozen BEATs encoder [7] on the near-microphone waveform. Three similarity-guided feature fusion (SGFF) blocks plus learnable scale fusion merge the branches; FiLM conditioning and a Conformer stack produce a 256-dimensional clip embedding (∼70 M parameters). Training uses mixup and target-domain oversampling for 50 epochs with distributed data parallel. Visible-attribute machines (Sander, ToyDrone) use ArcFace and center loss with domain×speed prototypes; hidden-attribute machines use contrastive embedding and domain cross-entropy with domain-only prototypes. Anomaly scores are minimum L2 distance to normal prototypes (proto_min) fit on eval-additional training data.

2.3. Tsz_DCASE_task2_fusion (T_fusion)

The near-microphone channel is encoded by BEATs iter3+AS2M [7] with the top four transformer blocks fine-tuned on all seven development machine types. Patch features are pooled with generalized mean (p=3) to a 768-dimensional embedding. ArcFace training uses composite labels (machine, section, domain, and visible attributes where available; scale s=64, margin m=0.2) with AdamW and early stopping on development Ω. For each eval machine, KNN and diagonal GMM negative log-likelihood scorers produce raw anomaly scores; each scorer is rank-normalized and combined by element-wise minimum. One global model is trained on development data and applied to each eval machine using its own additional training bank (∼95 M parameters).

2.4. Tsz_DCASE_task2_sera (sera_ft1)

Following the top-ranked DCASE 2025 Task 2 recipe, we use a frozen BEATs iter3+AS2M checkpoint (AudioSet finetuned) on the

near-microphone channel. Attentive Statistics Pooling (ASP) aggregates patch sequences into a fixed-dimensional embedding with batch normalization. A shared fully connected layer feeds an ArcFace main head (12 machine classes: 7 dev + 5 eval) and auxiliary heads for speed, product class, car ID, and grit where attributes are visible. Training uses AdamW (lr=10^{-4}), early stopping (patience 5), and an 80/20 train/validation split over cached training clips. Scoring applies per-clip embedding z-score normalization; minimum Euclidean distances to source (990) and target (10) normal banks are computed separately and fused by element-wise minimum at test time after z-score normalization.

3. EXPERIMENTS

3.1. Development set

Tables 1 and 2 report development-set AUC (source/target) and pAUC for AE_baseline and P_proto. T_fusion and sera_ft1 achieve macro-average official scores of 61.46% and 56.47%, respectively, on the seven-machine development set. These results are for architecture validation only.

Table 1: Development results for Tsz_DCASE_task2_baseline (%). Macro Ω: 58.17.

Machine	AUC src	AUC tgt	pAUC
fan	60.22	45.80	52.89
ToyCarEmu	69.48	71.44	53.16
gearboxEmu	75.68	52.36	53.16
bearingEmu	68.10	61.92	60.47
sliderEmu	66.54	49.64	51.16
valveEmu	58.54	57.08	50.68
ToyCar	76.64	53.96	57.74

Table 2: Development results for Tsz_DCASE_task2_1 (%, proto_min).

Machine	AUC src	AUC tgt	pAUC
fan	54.52	61.24	59.89
ToyCarEmu	57.58	52.02	49.74
gearboxEmu	70.60	53.18	53.63
bearingEmu	62.82	58.14	53.05
sliderEmu	55.00	47.62	48.63
valveEmu	49.94	44.88	48.95
ToyCar	51.22	61.46	51.37

3.2. Eval submission

All four systems submit anomaly scores and binary decisions for the five official eval machine types. AE_baseline and P_proto train one dedicated model per machine on eval-additional normals only. T_fusion and sera_ft1 use globally trained backbones and machine-specific normal banks for scoring.

4. CONCLUSION

We presented four complementary approaches to DCASE 2026 Task 2: a reconstruction baseline, a stereo discriminative fusion system, and two BEATs-based embedding methods with different training and scoring strategies. Future work includes stronger domain-generalization objectives and ensemble scoring across checkpoints.

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)*, 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)*, 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] T. Endo, R. Tanabe, Y. Nikaido, and Y. Kawaguchi, "Efficient anomalous sound detection using deep autoencoder and thresholding decision," in *Proc. DCASE Workshop*, 2021.

[6] Y. Kawaguchi, K. Imoto, D. Niizumi, N. Harada, T. Endo, Y. Nikaido, R. Tanabe, T. Nishida, K. Dohi, H. Purohit, and M. Yamamoto, "Description and discussion on DCASE 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," in *Proc. DCASE Workshop*, 2022.

[7] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, W. Wu, M. Zeng, X. Yu, and F. Yan, "BEATs: Audio pre-training with acoustic tokenizers," in *Proc. ICLR*, 2023.