CROSS-CHANNEL AND RAW-SPECTRAL DETECTOR ENSEMBLES FOR FIRST-SHOT NOISE-AWARE ANOMALOUS SOUND DETECTION

Technical Report

Seunggyu Jeong[1,2]*, Seong-Eun Kim*[1,2]

[1] Medisensing, Seoul, Korea
[2] Seoul National University of Science and Technology, Seoul, Korea
wa3229433@gmail.com, sekim@seoultech.ac.kr

ABSTRACT

We describe our submission to DCASE 2026 Challenge Task 2, first-shot unsupervised anomalous sound detection (ASD) for machine condition monitoring under a new noise-aware, two-channel (near and far) recording setup. Our system combines a self-supervised representation, an EAT audio transformer adapted with LoRA and trained with a composite supervised-contrastive objective, source-target alignment, and pseudo-attribute clustering, with an ensemble of orthogonal anomaly detectors. On top of the density core (Mahalanobis distance and a per-machine normalizing flow) we add two cross-channel detectors that use the second channel through complementary mechanisms: a learned cross-channel representation (LCCR) that classifies inter-channel level and phase maps, and a cross-channel predictive consistency score (C1) whose residual flags anomalies that the discriminative branch misses. A third detector scores raw log-mel energies with a selective Mahalanobis distance and is strong on the machines where the deep embedding is weak. All scores are combined by a domain-agnostic rank fusion, so the pipeline can be applied to the unlabelled evaluation set. On the development set the primary system reaches a harmonic-mean Ω of 62.95, above the official baseline (57.7). We submit four systems that trade off reliance on the deep embedding against the raw-spectral detector.

***Index Terms*—** Anomalous sound detection, first-shot, domain generalization, cross-channel, normalizing flow, representation diversity, ensemble

1. INTRODUCTION

DCASE 2026 Task 2 [1, 2] continues the first-shot formulation of unsupervised anomalous sound detection (ASD): training data contain only normal sounds, no machine-specific hyper-parameter tuning is allowed, and the evaluation machine types are entirely disjoint from the development ones [3]. The development and additional datasets follow the ToyADMOS2 [4] and MIMII DG [5] designs. Two further difficulties make the task hard. First, domain generalization: each machine is recorded under a source domain (990 clips) and a scarce target domain (10 clips), and the score is the harmonic mean Ω over all section and domain AUCs and the partial AUC (pAUC, p=0.1). Because Ω is a harmonic mean, the weakest machine dominates, so methods that keep the normal distribution compact and protect pAUC are preferred. Second, 2026 replaces the supplementary clean and noisy data of 2025 with a two-channel, clip-paired recording (a near and a far microphone). The second channel is the main new source of information, while the official baseline still uses the near channel only.

We build on an observation that held throughout our experiments: with a well-trained supervised-contrastive backbone, the time-pooled embedding is already a good density representation, and changing the core (angular-margin heads, query bottlenecks, concentration gating) did not help in our tests. The gains instead come from orthogonal detectors that measure different properties than the embedding density. Our contributions are two cross-channel detectors that use the far channel through independent mechanisms, one discriminative (LCCR) and one predictive (C1), together with a raw-spectral detector on a different representation that handles the machines the deep embedding separates poorly. Section 2 describes the representation, Section 3 the detectors and their fusion, Section 4 the four submitted systems, and Section 5 development-set results.

2. REPRESENTATION LEARNING

Backbone. We use EAT [6], an efficient audio transformer pre-trained on AudioSet [7], which we found consistently stronger for density/flow scoring and pAUC than BEATs [8]. The backbone is kept frozen and adapted with low-rank adapters (LoRA, rank 32, α=64) [9], so only a small number of parameters are trained. Patch embeddings are aggregated over time by mean-and-standard-deviation pooling, giving a 1536-dimensional clip embedding.

Training objective. The adapters are trained jointly on the normal training recordings of all machine types in the development and additional training sets (a single universal model). We use a composite supervised-contrastive loss [10] whose classes are the Cartesian product of machine identity and recording attribute, which sharpens per-class concentration. Two ingredients address domain generalization: (i) a source-target alignment term (weight 0.5) that pulls the two domain means together to protect machines whose target domain is sharp, and (ii) pseudo-attributes for machines without attribute labels, obtained by agglomerative clustering of embeddings, which provide contrastive structure where none is annotated. Together these raise the development kNN probe and the final density score.

3. ANOMALY DETECTORS AND FUSION

All detectors are fitted per machine on its normal training embeddings and produce a per-clip anomaly score; the scores are then rank-normalized and linearly fused.

3.1. Density core

The core comprises two complementary density estimators on the near-channel embedding. Mahalanobis distance under an empirical covariance is a low-variance scorer. A per-machine normalizing flow [11] (PCA to 64 dimensions followed by a 6-layer RealNVP with hidden width 128, 200 epochs) learns the normal density boundary and is strong on pAUC. The flow is seeded deterministically and re-seeded per machine for reproducibility, and we average three seeds. The two are complementary in the bias-variance sense, and their rank fusion forms the density core used by every submitted system.

3.2. Cross-channel detectors (contributions)

The far channel is informative only relative to the near channel, so both detectors operate on the channel pair.

LCCR (learned cross-channel representation). We compute inter-channel level differences, inter-channel phase differences (as $\cos$ / $\sin$ maps) and log-power mel maps from the stereo clip, and feed them to a small convolutional network (≈ 0.13M parameters) trained with the same composite supervised-contrastive objective. The resulting 128-dimensional cross-channel embedding is scored with its own Mahalanobis and flow density. LCCR is a discriminative use of the two channels, and on the development set it contributes more than the hand-crafted spatial feature it replaces.

C1 (cross-channel predictive consistency). In PCA (128-d) space we fit a ridge regressor that predicts the far-channel embedding from the near-channel embedding on normal training clips. The residual norm at test time is the anomaly score: a healthy machine keeps a stable near-to-far acoustic map, while a fault perturbs it. C1 is a predictive mechanism, complementary to LCCR, and recovers machines such as valve where the discriminative branch is weak.

3.3. Auxiliary detectors

Two further detectors broaden the ensemble. A hand-crafted relative-transfer-function spatial feature ($\log|H|$, $\cos\angle H$, $\sin\angle H$ per band) is scored with a min-centred Mahalanobis under a Ledoit-Wolf covariance, taking the minimum distance to the source and target centroids. A cosine-kNN detector uses one minus the maximum cosine similarity to the training normals, adding a discriminative, noise-robust view.

Raw-spectral detector (representation diversity). Every detector above reads the same EAT embedding and is therefore strongly correlated, so reweighting them cannot fix a machine the embedding separates poorly. We add one detector on a different representation: per-clip raw log-mel energies (mean and standard deviation over time), scored by a selective Mahalanobis distance, the minimum of the source- and target-centroid distances under a Ledoit-Wolf covariance. This low-level spectral view captures subtle stationary faults such as bearing wear that the deep embedding smooths over. On development it is the strongest single detector on the machines where the embedding underperforms the official baseline, and its score is only weakly correlated with the embedding density. It therefore adds representation diversity rather than reweighting the same features, and improves the harmonic mean by raising its weakest terms.

Table 1: Submitted systems and development scores. Ω and pAUC in %. Detectors: M=Mahalanobis, F=flow, L=LCCR, C=C1, S=RTF spatial, K=cosine-kNN, R=raw log-mel selective Mahalanobis.

System	Detectors (emphasis)	Ω	pAUC
1	M+F+L+C+S+K+R (balanced)	62.95	56.74
2	M+F+L+C+S (cross-channel)	62.46	57.10
3	M+F+S+L+C+R (spatial)	62.60	57.29
4	R+M+F (raw-spectral, robust)	62.13	55.03

3.4. Domain-agnostic normalization, fusion and decision

The evaluation test clips carry no domain label, so every detector is normalized in a domain-agnostic way: we take the minimum over the per-train-domain z-scores, $\tilde{s} = \min_{d\in\{\mathrm{src},\mathrm{tgt}\}}(s-\mu_d)/\sigma_d$, using only training-domain statistics. Normalized scores are rank-transformed and fused as

$$s = a\, r(\mathrm{Mah}) + a\, r(\mathrm{NF}) + \sum_j w_j\, r(D_j), \tag{1}$$

with $a = (1-\sum_j w_j)/2$ and D_j the cross-channel and auxiliary detectors. A binary decision uses the 90th percentile of the training scores as the threshold.

4. SUBMITTED SYSTEMS

The four systems differ in how much they rely on the EAT embedding versus the raw-spectral detector. System 2 uses only the embedding and cross-channel detectors, Systems 1 and 3 add a moderate raw-spectral term, and System 4 is dominated by the raw-spectral detector. With this range, at least one system should match the unknown evaluation machines. Table 1 lists them.

System 1 is the balanced ensemble of all detectors, including a moderate raw-spectral term, and gives the best development Ω. System 2 increases the weight of the learned and predictive cross-channel detectors and is the only purely embedding-based system. System 3 emphasizes the hand-crafted spatial feature with a small raw-spectral term added to raise its floor, and gives the best pAUC. System 4 is dominated by the raw-spectral detector with a light EAT density. It is the least correlated with the others (Spearman 0.87 to System 1, against 0.94 to 0.99 among the embedding-based systems) and the most uniform across machine types, which is useful if the deep embedding does not transfer well to the unseen evaluation machines.

5. RESULTS

Table 2 reports per-machine development results for the primary System 1. All numbers use the deployable domain-agnostic normalization (no test-domain information), so they are directly comparable to what the pipeline will produce on the evaluation set. The harmonic-mean Ω is 62.95 with mean pAUC 56.74, well above the official baseline (Ω=57.7). Compared with the embedding-only density core, the orthogonal detectors raise the weakest machines: the cross-channel detectors recover `fan` and `ToyCarEmu`, and the raw-spectral detector closes most of the gap on `bearing`, the only machine on which the deep embedding alone falls below the baseline.

Table 2: Development results of System 1 (AUC in %).

Machine	$\mathrm{AUC}_{\mathrm{src}}$	$\mathrm{AUC}_{\mathrm{tgt}}$	pAUC
ToyCar	72.12	84.80	63.16
ToyCarEmu	60.64	84.96	54.89
bearingEmu	62.68	55.60	54.89
fan	79.84	42.96	51.95
gearboxEmu	75.16	64.52	58.32
sliderEmu	63.84	55.84	50.89
valveEmu	91.96	79.92	63.11

6. CONCLUSION

Our DCASE 2026 Task 2 system keeps a supervised-contrastive representation as a fixed density core and obtains its improvements from orthogonal detectors. The two cross-channel detectors, LCCR (discriminative) and C1 (predictive), use the near and far recordings through independent mechanisms, while a raw-spectral selective-Mahalanobis detector on a different representation handles the machines the deep embedding separates poorly. We submit four systems that trade off reliance on the deep embedding against the raw-spectral detector; the last is a low-correlation, machine-uniform fallback for the unseen evaluation machines.

7. 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] "DCASE 2026 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," https://dcase.community/challenge2026/, 2026.

[3] 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.

[4] 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.

[5] 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.

[6] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "EAT: Self-supervised pre-training with efficient audio transformer," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[7] J. F. Gemmeke, D. P. W. 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. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[8] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proc. International Conference on Machine Learning (ICML)*, 2023.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. International Conference on Learning Representations (ICLR)*, 2022.

[10] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[11] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using Real NVP," in *Proc. International Conference on Learning Representations (ICLR)*, 2017.