DUAL-CHANNEL CROSS-ATTENTION EMBEDDINGS FOR NOISE-AWARE UNSUPERVISED ANOMALOUS SOUND DETECTION

Technical Report

Ziran Jiang, Rong Han

Beijing, China
632953645@qq.com

ABSTRACT

This technical report describes our submission to DCASE 2026 Challenge Task 2: first-shot unsupervised anomalous sound detection for machine condition monitoring. The proposed system learns compact representations from normal training recordings using a dual-channel convolutional encoder followed by cross-channel attention. The near- and far-microphone log-mel spectrograms are encoded separately, exchanged through bidirectional cross-attention layers, and averaged into a single normalized embedding. Anomaly scores are computed without anomalous training data by fitting Gaussian models to the training embeddings and using the minimum Mahalanobis distance as the score. For the official evaluation, we submit four systems that differ in model resolution and checkpoint-ensemble strategy: two 128-bin mel ensembles and two 256-bin mel ensembles.

***Index Terms*—** anomalous sound detection, machine condition monitoring, domain generalization, cross attention, Mahalanobis distance

1. INTRODUCTION

Anomalous sound detection (ASD) aims to detect abnormal operating sounds from machines using acoustic observations [1][2][3]. DCASE 2026 Challenge Task 2 [4] focuses on first-shot unsupervised ASD under domain shifts and unseen machine types. In this setting, only normal clips are available for training, the evaluation machine types differ from those in the development set, and the 2026 data provide synchronized two-channel recordings captured at different distances from the target machine.

Our submission is designed around two requirements. First, the model must use both microphone channels while remaining robust when one channel is noisier. Second, the scoring backend must be applicable to unseen machine types without using anomalous validation data. We therefore train machine-specific embedding extractors using only normal additional-training data, and use simple density models in the learned embedding space for anomaly scoring.

2. PROPOSED METHOD

2.1. Input features

All audio is resampled to 16 kHz and converted to log-mel spectrograms. The main configuration, denoted DCA-128, uses 128 mel bins, an FFT size of 1024, and 1024 time frames per clip. The higher-resolution configuration, denoted DCA-256, uses 256 mel bins and an FFT size of 2048 while keeping 1024 time frames. Machine-dependent hop sizes are used for machines with different clip durations so that the temporal dimension remains fixed. During training, SpecAugment with two frequency masks and two time masks is applied independently to both channels.

2.2. Dual-channel encoder

The embedding network contains a five-layer convolutional encoder and two bidirectional cross-attention layers. Each channel is first encoded by the same convolutional stack. For DCA-128, the encoder maps each input spectrogram to 1024 tokens with 256 channels; for DCA-256, the encoder preserves a larger frequency resolution and produces 2048 tokens. Cross-attention is then applied in both directions: channel 0 attends to channel 1, and channel 1 attends to channel 0. Each attention block uses four heads, residual connections, layer normalization, and a feed-forward subnetwork with GELU activations and dropout.

The attended tokens are averaged over time-frequency positions and projected to 256-dimensional embeddings. The two channel embeddings are ℓ_2 normalized and averaged to obtain the final clip-level representation. For machines where the far-channel signal was found unreliable in prior development experiments, the second channel can be replaced by a copy of the near channel; this policy is fixed before evaluation.

2.3. Training objective

Models are trained separately for each evaluation machine type using only the corresponding normal training recordings. For machines with attribute information, each attribute group is treated as a class and the training data are restricted to the source-domain samples. For machines without attributes, all available normal training samples are used with a single machine-level class. When more than one class is present, the embeddings are optimized with a Sub-center ArcFace classification loss. In all cases, center loss is also used to compact normal embeddings. The final loss is the average over the two channel branches. We train with AdamW, mixed precision, gradient clipping, a 5-epoch warm-up, and cosine learning-rate decay.

2.4. Anomaly scoring

After training, the neural network is used only as an embedding extractor. For machines without attributes, the training embeddings are partitioned into three clusters by k-means, and one Gaussian

Table 1: Submitted systems. "Fast LR" denotes training with a cosine schedule whose full length is 30 epochs (fast decay). "Slow LR" denotes training only to the saved epoch while using a slower cosine schedule parameterized as a 100-epoch run (slow decay).

System	Description
1	**DCA-128 fast+slow LR ensemble.** Average of DCA-128 ep30 scores from both the fast-LR and slow-LR configurations, over all seeds.
2	**DCA-128 fast LR ensemble.** Average of DCA-128 ep30 scores from the fast-LR configuration only, over all seeds.
3	**DCA-256 fast+slow LR ensemble.** Average of DCA-256 ep20 scores from both the fast-LR and slow-LR configurations, over all seeds.
4	**DCA-256 fast LR ensemble.** Average of DCA-256 ep20 scores from the fast-LR configuration only, over all seeds.

model is estimated per cluster. For machines with attributes, Gaussian models are estimated per source-domain attribute group. The ten target-domain normal samples are not used for discriminative tuning; they are used only to add target-domain Gaussian centers when the corresponding source attribute group exists. The covariance matrix is regularized by adding $10^{-4}I$. When adding target-domain centers, the source covariance is scaled using a chi-square threshold so that the target normal samples are not over-penalized.

For a test embedding z, the anomaly score is

$$A(z) = \min_k (z - \mu_k)^T \Sigma_k^{-1} (z - \mu_k), \tag{1}$$

where μ_k and Σ_k are the mean and covariance of the k-th normal cluster. Larger scores indicate more anomalous clips. Binary decisions are obtained by applying a fixed threshold to the submitted anomaly scores; the threshold is not optimized on the evaluation test labels. This scoring protocol follows the first-shot anomaly detection paradigm [5][6].

3. SUBMISSION SYSTEMS

The challenge allows up to four submitted systems. All four systems use the same architecture family, preprocessing, training loss, and Mahalanobis scoring backend. They differ only in the mel resolution and in how checkpoint scores are averaged. For every machine, scores are first generated for ten random seeds: 42, 123, 456, 789, 1024, 2048, 3141, 4096, 5555, and 6789. Score-level averaging is performed independently for each evaluation file.

The two DCA-128 configurations share the same 128-bin architecture but use different cosine learning-rate schedules. The fast-LR configuration trains for 30 epochs with a 30-epoch cosine schedule and saves ep10, ep20, and ep30 checkpoints. The slow-LR configuration trains to ep30 with a 100-epoch cosine schedule and saves the same checkpoints. The DCA-256 systems use the 256-bin architecture. System 3 averages ep20 checkpoints from both the fast-LR and slow-LR runs, while System 4 uses only the ep20 checkpoints from the fast-LR DCA-256 run.

4. EXPERIMENTAL SETUP

The final models are trained on the additional training dataset released for DCASE 2026 Task 2. The evaluation set contains unlabeled test clips for the same additional-training machine types, so no evaluation labels are used for model selection, threshold tuning, or score calibration. The machine types handled by the final scoring scripts are BlowerDustCollector, Sander, SewingMachine, ToothBrush, and ToyDrone. BlowerDustCollector, Sander, and ToyDrone are treated as attributed machines, while SewingMachine and ToothBrush are treated as no-attribute machines.

We do not use prohibited previous DCASE Task 2, ToyADMOS, MIMII, MIMII DUE, or MIMII DG [7] data. The submitted systems are trained from the provided DCASE 2026 normal training data only. No anomalous clips from development or evaluation test sets are used for training.

5. RESULTS

Table 2 reports development-set results. The official score Ω is the harmonic mean of all source-domain AUCs, target-domain AUCs, and pAUCs across the seven development machine types. We compare with the two official baselines: selective Mahalanobis (MAHALA) and simple autoencoder reconstruction error (AE). Our DCA-128/DCA-256 spreadsheets did not include ToyCar and ToyCarEmu, so these two machines were completed by recomputing the official metrics from the stored old-model scores in the DCA-128 30-epoch checkpoints using the official recalculation script.

The submitted systems improve substantially on Fan and GearboxEmu, while the official baselines remain stronger on ToyCarEmu, ToyCar, BearingEmu, and SliderEmu. This development behavior motivated submitting multiple checkpoint and resolution ensembles rather than relying on a single model selection criterion.

Because the official evaluation labels are not available to participants at submission time, official evaluation-set AUC, pAUC, and ranking scores cannot be reported here. The submitted anomaly-score files follow the official format, with one anomaly score for each evaluation test clip. The official score will be computed by the organizers from the source-domain AUC, target-domain AUC, and pAUC values for each machine section.

6. CONCLUSION

We presented a first-shot unsupervised ASD system based on dual-channel cross-attention embeddings and Mahalanobis-distance scoring. The method explicitly uses the synchronized near- and far-microphone recordings provided in DCASE 2026 Task 2 while keeping the scoring backend simple and label-free. Our four submissions explore checkpoint and resolution diversity through score-level averaging across seeds, training schedules, and mel resolutions.

7. REFERENCES

[1] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[2] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmen-

Table 2: Development-set comparison with the official baselines. Values are percentages.

Machine	Metric	MAHALA	AE	System 1	System 2	System 3	System 4
ToyCarEmu	AUC_source	69.49	69.62	47.64	47.64	48.52	48.52
	AUC_target	66.62	61.20	31.36	31.36	34.80	34.80
	pAUC	53.47	55.89	47.37	47.37	47.68	47.68
ToyCar	AUC_source	77.28	75.62	47.96	47.96	45.20	45.20
	AUC_target	53.17	37.87	43.00	43.00	46.56	46.56
	pAUC	58.25	54.03	49.42	49.42	50.42	50.42
BearingEmu	AUC_source	65.92	62.34	63.56	68.64	58.28	56.76
	AUC_target	62.28	59.56	58.48	56.52	54.76	55.80
	pAUC	60.42	59.85	52.05	53.00	49.32	51.53
Fan	AUC_source	60.00	61.45	86.48	89.32	95.84	95.60
	AUC_target	45.09	46.94	54.56	57.40	45.28	45.48
	pAUC	52.29	53.33	62.63	62.53	54.47	53.79
GearboxEmu	AUC_source	74.48	68.23	73.12	71.40	72.84	70.48
	AUC_target	52.74	49.78	58.28	60.16	73.40	74.80
	pAUC	53.97	52.94	52.47	52.63	60.79	60.42
SliderEmu	AUC_source	66.36	67.25	45.56	42.72	49.96	49.68
	AUC_target	49.18	45.05	53.76	41.84	50.44	53.08
	pAUC	50.36	50.38	52.42	52.16	49.84	51.42
ValveEmu	AUC_source	56.60	67.74	48.36	52.28	50.08	53.04
	AUC_target	56.50	68.78	53.24	56.80	50.84	50.12
	pAUC	50.20	55.08	48.68	49.11	51.11	48.47
Overall	Ω	57.66	56.66	51.76	51.56	52.08	52.28

tation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[3] P. C. Mahalanobis, "On the generalized distance in statistics," *Proceedings of the National Institute of Sciences of India*, vol. 2, no. 1, pp. 49–55, 1936.

[4] 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," *arXiv e-prints: 2606.01578*, 2026.

[5] 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)*, 2021, pp. 1–5.

[6] 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.

[7] 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 Detection and Classification of Acoustic Scenes and Events Workshop (DCASE2022)*, Nancy, France, 2022.