A SPATIAL-TEMPORAL ATTENTION AND CONFIDENCE-BASED DOMAIN ADAPTATION FRAMEWORK FOR NOISE-AWARE FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION

Technical Report

Nighil Natarajan, Raghav Sridharan, Nithilan M, Chandrakala S

Department of Computer Science and Engineering, Shiv Nadar University Chennai, India

ABSTRACT

This report presents our system for the DCASE 2026 Challenge Task 2 on noise-aware first-shot unsupervised anomalous sound detection. We propose a domain-adversarial contrastive learning framework that learns robust acoustic representations under domain shift and varying machine operating conditions. Raw audio captured from the near microphone is converted into RGB log-Mel spectrogram patches and processed by a pre-trained ResNet-34 encoder. Spatial features are extracted through attention pooling, while a temporal encoder captures sequential acoustic dependencies. To improve robustness against domain shifts and environmental noise, the framework combines confidence-weighted contrastive learning with Domain Adversarial Neural Networks (DANN), CORAL, and Maximum Mean Discrepancy (MMD) losses. Anomaly detection is performed using a hybrid Mahalanobis-Cosine scoring strategy over clustered normal embeddings. Experimental results on the DCASE 2026 Task 2 development dataset demonstrate improved anomaly detection performance under noise-aware and domain-shift conditions. The source code is publicly available at `https://github.com/nighiln05/DCASE_26_Task2`.

***Index Terms*—** Anomalous Sound Detection; Domain Adversarial Learning; Contrastive Learning; Temporal Attention; Domain Shift; Predictive Maintenance.

1. INTRODUCTION

Anomalous Sound Detection (ASD) is a critical component of predictive maintenance in industrial systems, enabling early fault identification from acoustic signals without invasive sensors [1, 2]. This also drives the need for lightweight, efficient models deployable directly on edge hardware [3]. The DCASE 2026 Challenge Task 2 [4, 5] builds upon the MIMII DG and ToyADMOS2 datasets [6, 7] and extends the first-shot unsupervised anomalous sound detection setting to a noise-aware scenario using synchronized recordings captured simultaneously from near and far microphones. Models must detect anomalies from normal-only training data while generalizing across source and target operational domains with significant distributional shift and varying acoustic conditions.

This setting creates two core challenges. First, the scarcity of anomalous data makes supervised training impractical, requiring models to define normality from limited examples [1]. Second, domain shift arises from changes in machine speed, load, and environment between training and test conditions, causing standard detectors to confuse distributional shift with actual faults [2]. Recent reviews highlight this specific domain mismatch as a primary performance bottleneck [8]. Standard contrastive learning approaches address representation learning well but lack explicit mechanisms for cross-domain alignment, and are sensitive to noisy or anomaly-corrupted training samples [9].

We address both challenges through a unified framework. Spectrograms are segmented into patches and processed by a ResNet-34 encoder [1]. A dual-pathway architecture extracts spatial features via attribute-conditioned attention pooling and temporal features via a convolutional temporal encoder. Domain shift is explicitly countered through CORAL, MMD, and DANN losses applied jointly during training. A confidence-based sample weighting mechanism using exponential moving average (EMA) centers suppresses unreliable samples during contrastive optimization. At inference, anomaly scores are produced by a hybrid Mahalanobis-Cosine metric over K-means clustered embeddings, with adaptive model selection per machine type.

2. PROPOSED SYSTEM

2.1. System Overview

Fig. 1 illustrates the proposed framework. Raw audio signals are transformed into RGB log-Mel spectrograms and partitioned into overlapping patches. A ResNet-34 encoder extracts patch-level representations, which are subsequently processed by a dual-path spatial-temporal feature extractor. The resulting embeddings are optimized using confidence-aware contrastive learning together with domain-alignment

objectives. During inference, domain-specific clustering and hybrid distance metrics are used to compute anomaly scores.

2.2. Preprocessing

The audio recordings are processed using only the near-microphone channel (Channel 0). The selected audio is resampled to 16 kHz and converted into log-Mel spectrograms using a 1024-point FFT, hop length of 512, and 128 Mel frequency bins spanning 20 Hz to 8 kHz. The resulting spectrograms are normalized, converted into RGB images using the plasma colormap, and resized to 224×224. Each image is partitioned into overlapping 32×32 patches with stride 16, producing a sequence of patch tokens for subsequent representation learning.

2.3. Data Augmentation and View Generation

To improve representation robustness and enable contrastive learning, two augmented views, denoted as v_1 and v_2, are generated from each RGB log-Mel spectrogram. The augmentation pipeline includes:

- Random Resized Crop
- Horizontal Flip
- Color Jitter
- Grayscale Conversion

The two augmented views are processed independently through the shared patch extraction and feature encoding pipeline. This produces two embedding representations, $\mathbf{z}_1$ and $\mathbf{z}_2$, which form a positive pair for confidence-weighted NT-Xent contrastive learning. By enforcing consistency between different augmented views of the same audio sample, the model learns invariant representations that are more robust to acoustic variability and domain shifts.

2.4. Dual-Path Spatial-Temporal Encoding

2.4.1. Patch Encoding

Each patch is processed by an ImageNet pre-trained ResNet-34 encoder followed by a projection head to obtain patch embeddings

$$\mathbf{H} \in \mathbb{R}^{B \times N \times D}, \tag{1}$$

where B denotes the batch size, N the number of extracted patches, and $D = 128$ is the embedding dimension. The ResNet-34 encoder captures local spectro-temporal patterns from each patch, while the projection head maps the extracted features into a compact representation space optimized for contrastive learning. All patch embeddings are subsequently ℓ_2-normalized before further processing.

2.4.2. Spatial Attention Pooling

To identify informative spectro-temporal regions, an attribute-conditioned attention mechanism assigns importance weights to each patch. Given patch embedding h_i and attribute vector $\mathbf{a}$, the attention score is computed as

$$\tilde{s}_i = \mathbf{W}_2 \tanh\left(\mathbf{W}_1 h_i + \mathbf{b}_1\right) + \mathbf{W}_{\text{attr}}\mathbf{a}. \tag{2}$$

where $\mathbf{W}_1$, $\mathbf{W}_2$, and $\mathbf{W}_{\text{attr}}$ are learnable weight parameters, $\mathbf{b}_1$ is a bias term, and $\mathbf{a}$ denotes the machine attribute embedding. The attention mechanism assigns higher weights to informative patches while suppressing background regions and non-discriminative acoustic content.

2.4.3. Temporal Feature Encoding

While spatial attention identifies informative regions, temporal modeling captures the evolution of acoustic events across patches. A temporal attention layer first reweights patch embeddings according to their sequential importance. The reweighted sequence is subsequently processed using stacked one-dimensional convolutions and adaptive average pooling to obtain a fixed-length temporal representation. This pathway captures periodic machine behavior and transient acoustic deviations that may indicate anomalous operation.

2.4.4. Feature Fusion

The spatial and temporal representations are concatenated and ℓ_2 normalized to form the final embedding used for training and inference:

$$\mathbf{z}_{\text{final}} = \frac{[\mathbf{z}_{\text{spatial}} \,||\, \mathbf{z}_{\text{temp}}]}{\|[\mathbf{z}_{\text{spatial}} \,||\, \mathbf{z}_{\text{temp}}]\|_2}. \tag{3}$$

where $\mathbf{z}_{\text{spatial}}$ and $\mathbf{z}_{\text{temp}}$ denote the spatial and temporal feature representations, respectively. Feature fusion combines complementary information from both pathways, allowing the model to jointly capture local spectral structures and longer-range temporal dependencies. When machine attributes are available, an attribute embedding is concatenated before normalization.

2.5. Confidence based Contrastive Learning

To reduce the influence of noisy or unreliable samples, an exponential moving average (EMA) center is maintained throughout training. Samples closer to the center are assigned higher confidence weights, while distant samples are down-weighted. The overall contrastive objective combines confidence-weighted NT-Xent loss with center-pull and boundary-expansion regularization:

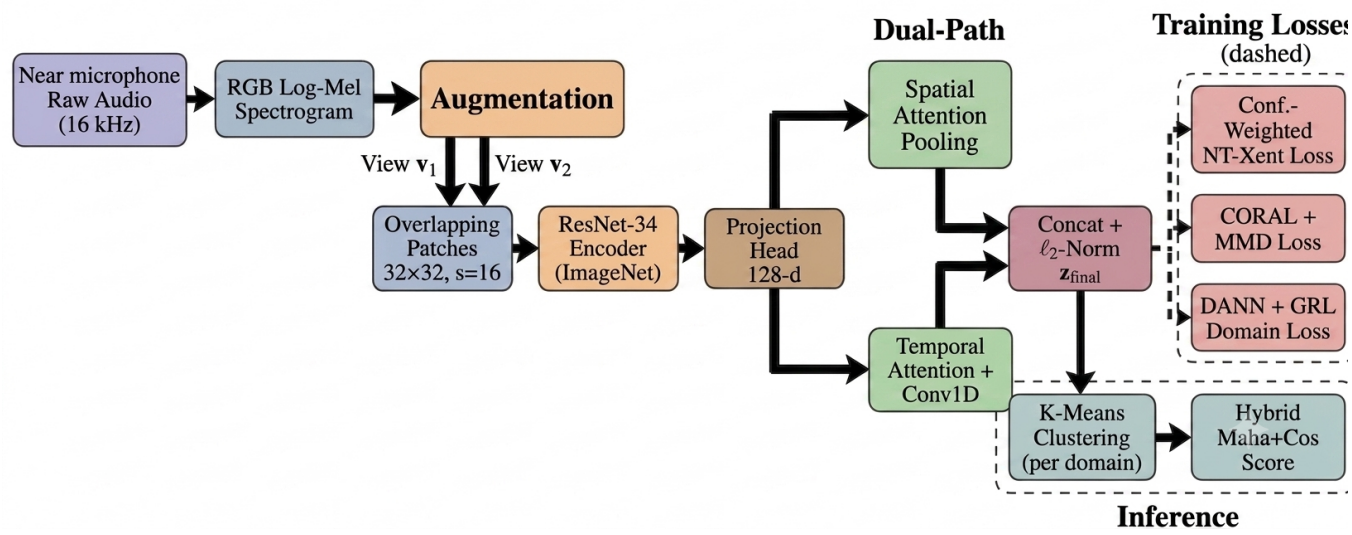

Figure 1: Overview of the proposed anomalous sound detection framework.

$$\mathcal{L}_{\text{CL}} = \frac{1}{B}\sum_{i=1}^{B} \omega_i \ell_{\text{NT-Xent}}^{(i)} + \beta_c \mathcal{L}_{\text{center}} + \beta_b \mathcal{L}_{\text{boundary}}. \quad (4)$$

where ω_i denotes the confidence weight of sample i, $\ell_{\text{NT-Xent}}$ is the contrastive loss, and β_c and β_b control the contributions of center-pull and boundary regularization terms, respectively. Confidence weighting reduces the influence of unreliable samples and stabilizes representation learning under noisy training conditions. This strategy encourages compact normal clusters while preserving sufficient feature diversity for anomaly discrimination.

2.6. Domain Alignment

To improve robustness against domain shifts, three complementary alignment objectives are employed.

2.6.1. CORAL Alignment

CORAL minimizes covariance discrepancies between source and target embeddings:

$$\mathcal{L}_{\text{CORAL}} = \frac{1}{4D^2} \|C_S - C_T\|_F^2 . \quad (5)$$

where C_S and C_T represent the covariance matrices of source and target embeddings, respectively, and $\|\cdot\|_F$ denotes the Frobenius norm. Minimizing this objective encourages the second-order statistics of both domains to become similar.

2.6.2. MMD Alignment

Maximum Mean Discrepancy (MMD) aligns source and target feature distributions through kernel mean matching, reducing first-order distributional differences.

2.6.3. Adversarial Domain Learning

A domain discriminator is trained through a Gradient Reversal Layer (GRL) to encourage domain-invariant embeddings. The final training objective combines contrastive learning and domain-alignment losses:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{CL}} + 0.015\,\mathcal{L}_{\text{CORAL}} + 0.005\,\mathcal{L}_{\text{MMD}} + 0.02\,\mathcal{L}_{\text{adv}}. \quad (6)$$

2.7. Inference and Anomaly Scoring

During inference, embeddings are reduced using Principal Component Analysis (PCA) and clustered using K-means to model multiple modes of normal machine behavior. For each cluster, covariance statistics are estimated using Ledoit-Wolf estimators. The anomaly score combines Mahalanobis distance and cosine distance:

$$\text{score}(\mathbf{x}) = 0.7\, z(m_{\min}) + 0.3\, z(c_{\min}), \quad (7)$$

where $m_{\min}$ and $c_{\min}$ denote the minimum Mahalanobis and cosine distances across clusters, respectively. This hybrid formulation captures both distributional deviation and angular dissimilarity, resulting in robust anomaly detection under domain-shift conditions.

Table 1: Performance comparison on the DCASE 2026 Task 2 development dataset.

Machine Type	Method	AUC [%]		pAUC [%]
		Source	Target	
Fan	MSE	61.45	46.94	53.33
	MAHALA	60.00	45.09	52.29
	OURS	**63.96**	**59.16**	**54.37**
Bearing (Emu)	MSE	62.34	59.56	59.85
	MAHALA	**65.92**	**62.28**	**60.42**
	OURS	63.40	59.92	51.47
Gearbox (Emu)	MSE	68.23	49.78	52.94
	MAHALA	**74.48**	52.74	53.97
	OURS	70.80	**59.48**	**55.58**
Slider (Emu)	MSE	**67.25**	45.05	50.38
	MAHALA	66.36	49.18	50.36
	OURS	62.40	**59.96**	**53.47**
Valve (Emu)	MSE	67.74	68.78	55.08
	MAHALA	56.60	56.50	50.20
	OURS	**87.60**	**80.12**	**63.32**
ToyCar	MSE	75.62	37.87	54.03
	MAHALA	77.28	53.17	**58.25**
	OURS	**82.36**	**73.48**	50.53
ToyCar (Emu)	MSE	69.62	61.20	55.89
	MAHALA	**69.49**	66.62	53.47
	OURS	63.08	**85.32**	**59.16**

2.8. Implementation Details

The system is implemented in PyTorch and explicitly optimized for constrained hardware, with training conducted on an NVIDIA RTX 5060 Laptop GPU. To maintain computational efficiency, input Log-Mel spectrograms are partitioned into 32x32 patches with a stride of 16 (maximum 32 patches per sample). The network is trained jointly across all machine types for 150 epochs using the Adam optimizer (initial learning rate 2×10^{-4}, batch size 96) and regulated by a ReduceLROnPlateau scheduler (patience 10, decay factor 0.5). For the contrastive learning formulation, the NT-Xent loss temperature is fixed at $\tau = 0.05$, with center representations for confidence weighting maintained via an Exponential Moving Average (momentum $\gamma = 0.9$). Following an 8-epoch warmup phase, the objective is augmented with boundary ($\beta_{\text{boundary}} = 0.02$) and center-pull ($\beta_{\text{center}} = 0.05$) penalties, alongside dynamic confidence weighting ($w_{\text{hard}} \rightarrow 1.25$, $w_{\text{soft}} \rightarrow 0.9$), to progressively refine the representation space.

3. RESULTS AND DISCUSSION

We evaluated the proposed system on the DCASE 2026 Task 2 development dataset using AUC as the primary evaluation metric. Performance was measured separately for the source and target domains to assess robustness under domain-shift conditions. Since DCASE 2026 focuses on noise-aware anomalous sound detection using recordings captured under realistic acoustic environments, the evaluation also reflects the ability of the proposed framework to maintain discriminative representations in the presence of environmental noise and microphone-related variability. The proposed framework was compared against the official DCASE 2026 Mahalanobis-distance baseline [5]. Table 1 summarizes the obtained AUC scores for all machine types. The results demonstrate that the proposed method consistently improves target-domain performance while maintaining competitive source-domain performance, indicating enhanced generalization to unseen operating conditions.

3.1. Discussion

The proposed framework demonstrates strong robustness under domain-shift conditions, particularly for *Valve (Emu)*, *ToyCar*, and *ToyCar (Emu)*. These machine types benefit from the dual-path spatial-temporal encoder, which captures both localized spectral characteristics and longer-term temporal dependencies.

Performance improvements on *Gearbox (Emu)* and *Slider (Emu)* further indicate the effectiveness of attention-based feature aggregation in highlighting informative spectro-temporal regions. While the official baseline achieves slightly higher performance on *Bearing (Emu)*, the proposed method provides more consistent target-domain performance across most machine categories.

Overall, the combination of confidence-aware contrastive learning and domain-alignment objectives enables the model to learn robust and transferable representations for first-shot anomalous sound detection under challenging domain-shift conditions.

4. CONCLUSION

This paper presented a domain-generalized anomalous sound detection framework for DCASE 2026 Task 2. The proposed system combines a dual-path spatial-temporal encoder with confidence-aware contrastive learning to learn robust representations of normal machine behavior. To improve generalization under domain shifts, CORAL, MMD, and adversarial domain adaptation objectives are jointly incorporated during training. During inference, a hybrid Mahalanobis-cosine scoring mechanism models multiple modes of normal operation and enables robust anomaly detection across varying acoustic conditions. Experimental results on the DCASE

2026 development dataset demonstrate consistent improvements in target-domain performance and strong robustness to domain shifts, highlighting the effectiveness of the proposed representation learning and domain-alignment strategy for industrial anomalous sound detection.

5. REFERENCES

[1] S. Ahn *et al.*, "Unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE Workshop*, 2021.

[2] Y. Tagawa *et al.*, "Acoustic anomaly detection of machine sounds based on image transfer learning," in *Proc. DCASE Workshop*, 2021.

[3] Y.-C. Lo, T.-L. Tsai, C.-W. Yang, and A.-Y. Wu, "An efficient anomalous sound detection system for microcontrollers," *Sensors*, vol. 24, p. 7478, 2024.

[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, "First-shot anomaly detection for machine condition monitoring: A domain generalization baseline," in *Proc. EUSIPCO*, 2023, pp. 191–195.

[6] 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, Nov. 2022.

[7] 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, Nov. 2021, pp. 1–5.

[8] K. Wilkinghoff, T. Fujimura, K. Imoto, J. L. Roux, Z.-H. Tan, and T. Toda, "Handling domain shifts for anomalous sound detection: A review of DCASE-related work," *arXiv e-prints: 2503.10435*, 2025.

[9] H. Hojjati and N. Armanfard, "Self-supervised acoustic anomaly detection via contrastive learning," in *Proc. ICASSP*, 2022.