PATCH-BASED CONTRASTIVE LEARNING WITH LATENT SPACE CLUSTERING FOR UNSUPERVISED SOUND ANOMALY DETECTION

Technical Report

Abhivanth Sivaprakash[1]*, K Krish Sundaresh*[1]*, Ankith Vijayyan*[1]*, Adhithya Srivatsan*[1]*, Chandrakala S*[2]

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

ABSTRACT

This report presents our submission for the DCASE 2025 Challenge Task 2 on first-shot unsupervised anomalous sound detection. We propose a contrastive learning-based framework designed to capture fine-grained patterns from spectrogram representations while adapting to both attribute-rich and attribute-absent machine conditions. The method leverages local feature learning and selectively integrates auxiliary metadata to enhance generalization under domain shifts. Training is performed jointly across all machine types using only normal data. Anomaly scoring is carried out in a learned embedding space using a statistical distance-based method. Our approach outperforms official baselines in both source and target domains on the development dataset, demonstrating strong potential for robust and flexible industrial anomaly detection.

***Index Terms*—** unsupervised anomaly detection, first-shot learning, contrastive representation learning, spectrogram features, attribute-aware modeling, domain adaptation, industrial sound monitoring

1. INTRODUCTION

Anomalous Sound Detection (ASD) is a key component of predictive maintenance in industrial systems. It involves determining whether a sound emitted by a machine is abnormal—an essential task to prevent potential mechanical failures, ensure safety, and reduce downtime. The 2025 edition of the DCASE Challenge Task 2, "First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring," advances this problem setting by introducing more realistic constraints. Participants must detect anomalies across previously unseen machine types, often using only a single ID's normal data and without any anomalous training examples or complete metadata.

This "first-shot" scenario reflects practical challenges in real-world industrial deployment. First, anomalies are inherently rare and difficult to collect, making it impractical to rely on fully supervised models. Second, domain shifts—caused by changes in background noise, operational modes, or recording devices—introduce significant distributional variance between training and testing environments. Third, attribute metadata such as machine ID or load condition may be partially missing or noisy, limiting the ability to condition models on known context. Therefore, the central challenge lies in building models that are both generalizable across domains and robust to incomplete supervision.

In this report, we present a patch-level, metadata-aware contrastive learning framework designed to tackle these challenges. Instead of encoding entire spectrograms globally, our model splits input audio representations into overlapping patches, which are encoded through a shared visual backbone. This enables the learning of localized audio representations that are sensitive to fine-grained anomalies. An attention-based pooling module selectively aggregates patch-level features, modulated by auxiliary metadata when available. We further incorporate attribute embeddings into the contrastive objective through early fusion, enabling the model to adjust dynamically to varying levels of label completeness.

The entire system is trained in a self-supervised manner using a contrastive loss function over pairs of augmented spectrograms. During inference, embeddings are scored using Mahalanobis distance to reference normal distributions derived from training data. This approach enables anomaly detection based purely on distributional shifts in the learned feature space, without requiring anomaly labels or per-machine tuning.

Our experiments on the DCASE 2025 development dataset demonstrate that this method achieves strong generalization across both attribute-rich and attribute-absent conditions, outperforming baseline systems in multiple domains. The combination of local feature modeling, contrastive pretraining, and metadata-aware pooling proves to be a robust and effective strategy for first-shot unsupervised anomaly detection.

2. PROPOSED METHODOLOGY

We propose a self-supervised patch-level representation learning framework, augmented with attribute-conditioned attention and domain-adaptive anomaly scoring. Our pipeline is designed to tackle two central challenges in DCASE2025 Task 2: (i) generalization to unseen target domains, and (ii) handling partial or noisy metadata. The model learns robust localized features from log-Mel spectrograms via contrastive learning over augmented patch views.

2.1. Audio Preprocessing

Each 10-second audio clip is resampled to 16 kHz to ensure consistent time resolution across all devices. The waveform $x(t)$ is then transformed into a time-frequency representation using the log-Mel spectrogram, computed with:

- FFT size $n_{\text{fft}} = 1024$
- Hop length $h = 512$
- $M = 128$ Mel filter banks
- Sampling rate: 16kHz

This yields a matrix $S \in R^{128\times T}$, where $T \approx 313$ time frames for a 10 s segment.

We compute the log-compressed Mel energies:

$$\text{logmel}(x) = \log\left(\frac{\text{MelSpec}(x)}{\max(\text{MelSpec}(x))} + \epsilon\right), \quad \epsilon = 10^{-6}$$

This transformation ensures numerical stability while enhancing low-amplitude regions that might encode anomalies. The resulting grayscale spectrogram is duplicated across 3 channels (RGB) and resized to 224×224 via bilinear interpolation to match input requirements of standard CNN backbones.

2.2. Patch-Based Representation Learning

To capture fine-grained anomaly cues, the full spectrogram image $I \in R^{3\times224\times224}$ is divided into overlapping square patches of size 32×32, using a fixed stride of 16. Each patch $p_{i,j}$ corresponds to a localized time-frequency region, allowing the model to reason over temporal and spectral variations independently.

Formally, we define the set of extracted patches as:

$$\mathcal{P}(I) = \left\{ I_{i,j} \mid I_{i,j} \in R^{3\times32\times32},\ (i,j) \in \mathcal{G} \right\}$$

where $\mathcal{G}$ defines the grid of valid patch coordinates. We cap the maximum number of patches per view to 64 for memory efficiency.

Patch-wise representation allows our model to detect highly localized disruptions such as short squeaks or transient distortions, which might be missed by holistic encoders.

2.3. View Generation via Data Augmentation

To enforce representational invariance, we create two stochastic augmentations $x_1 = t_1(x)$ and $x_2 = t_2(x)$ of the same spectrogram. Each transform includes:

- RandomResizedCrop (scale range: 0.8–1.0)
- Horizontal flip
- Color jitter (brightness, contrast, saturation)
- Grayscale conversion (p=0.2)
- Per-channel normalization to zero mean, unit variance

These transformations simulate domain perturbations (e.g., lighting conditions, camera variations) and force the model to learn content-preserving representations. Each view is independently patchified, yielding $\{p_i^{(1)}\}$ and $\{p_i^{(2)}\}$.

2.4. Patch Encoding and Projection

Each patch p_i is resized to 224×224 and passed through a ResNet-34 encoder f_θ pretrained on ImageNet. The output is a feature vector $h_i \in R^{512}$, capturing mid- to high-level visual semantics.

These vectors are projected into a contrastive embedding space via an MLP:

$$z_i = W_2 \cdot \text{ReLU}(W_1 h_i + b_1) + b_2, \quad z_i \in R^{128}$$

This projection is shared across patches and views, ensuring consistent feature alignment. The latent dimension of 128 was empirically chosen to balance expressivity and generalization.

2.5. Attention Pooling with Attribute Conditioning

The patch embeddings $\{z_1, \ldots, z_N\}$ are aggregated into a global vector using an attention mechanism that optionally incorporates device metadata (e.g., RPM, temperature, load).

The attention weights are computed via:

$$\alpha_i = \frac{\exp(w^\top \tanh(W_a z_i))}{\sum_j \exp(w^\top \tanh(W_a z_j))}$$

If attributes $a \in R^{d_a}$ are available, a bias term $\beta = W_{\text{attr}}^\top a$ is added to every score:

$$\alpha_i = \frac{\exp(w^\top \tanh(W_a z_i) + \beta)}{\sum_j \exp(w^\top \tanh(W_a z_j) + \beta)}$$

The final embedding is a weighted sum:

$$z_{\text{pooled}} = \sum_{i=1}^{N} \alpha_i z_i$$

2.5.0.1. *Early Fusion of Attributes:*

To retain device-specific context, the attribute vector is also passed through an MLP to produce $a' \in R^{128}$:

$$a' = \text{MLP}(a), \quad z_{\text{final}} = [z_{\text{pooled}}; a']$$

This design allows for flexible fusion: models trained on datasets without attributes fall back to standard attention, while those with metadata benefit from conditional reasoning.

2.6. Contrastive Learning Objective

We use the NT-Xent loss to train embeddings that pull together views of the same sample and push apart other instances:

$$\mathcal{L}_i = -\log \frac{\exp(\text{sim}(z_i^{(1)}, z_i^{(2)})/\tau)}{\sum_{k=1}^{2N} 1_{[k\neq i]} \exp(\text{sim}(z_i^{(1)}, z_k)/\tau)}$$

$$\text{sim}(u, v) = \frac{u^\top v}{\|u\|\|v\|}, \quad \tau = 0.1$$

The loss is averaged across all N instances in the batch. This formulation effectively separates normality clusters across domains without using labels, facilitating unsupervised pretraining.

2.7. Domain-Adaptive Inference

To tackle domain shift, we employ a two-step statistical modeling approach:

1. Compute domain-specific means μ_s, μ_t from normal samples in the source and target domains.
2. Estimate a shared empirical covariance Σ using the pooled embeddings.

At evaluation time, for each test embedding z we compute two Mahalanobis distances:

$$d_s(z) = (z - \mu_s)^\top \Sigma^{-1} (z - \mu_s), \tag{1}$$

$$d_t(z) = (z - \mu_t)^\top \Sigma^{-1} (z - \mu_t). \tag{2}$$

We then define the final anomaly score as

$$\text{score}(z) \;=\; \min\big\{\, d_s(z),\, d_t(z) \big\}.$$

Finally, we form the threshold τ as the 92%-percentile of the union of all normal distances:

$$\tau \;=\; \text{Percentile}_{0.92}\Big(\{d_s(x)\}_{x\in\mathcal{N}} \;\cup\; \{d_t(x)\}_{x\in\mathcal{N}}\Big),$$

and declare z anomalous if $\text{score}(z) > \tau$.

This allows the system to detect outliers relative to both seen and unseen domains, adapting to varied operating environments.

2.8. Implementation Details

- **Backbone:** ResNet-34, pretrained on ImageNet
- **Patch size:** 32×32, stride = 16
- **Max patches:** 64 per view
- **Embedding dim:** 128 (256 if attributes are concatenated)
- **Temperature:** 0.1 (for NT-Xent loss)
- **Training:** 500 epochs, Adam optimizer with initial LR = 2×10^{-4}
- **LR scheduler:** ReduceLROnPlateau (patience=10, factor=0.5)
- **Early stopping:** patience = 25 epochs
- **Batch size:** 256
- **Training split:** joint across all machine types, with per-sample attributes padded to a common size

3. RESULTS AND DISCUSSION

We evaluated the performance of our system on the development dataset using AUC and partial AUC (pAUC, FPR 0.1) as evaluation metrics. Our method is compared with the official DCASE2025 baselines: the MSE-based autoencoder and the Mahalanobis distance model. Table 1 presents the average performance across source and target domains for all machine types.

Table 1: AUC and pAUC (%) comparison on development set (Ave. across 3 runs)

Machine	Metric	MSE	MAHALA	Ours
ToyCar	AUC(src)	66.98	63.01	**62.12**
	AUC(tgt)	33.75	37.35	**64.36**
	pAUC	48.77	51.04	**49.05**
ToyTrain	AUC(src)	76.63	61.99	**61.36**
	AUC(tgt)	46.92	39.99	**64.16**
	pAUC	47.95	48.21	**54.32**
bearing	AUC(src)	62.01	54.43	**60.88**
	AUC(tgt)	61.40	51.58	**69.40**
	pAUC	57.58	58.82	**58.32**
fan	AUC(src)	70.96	77.99	**59.72**
	AUC(tgt)	38.75	38.56	**55.44**
	pAUC	49.46	50.82	**51.58**
gearbox	AUC(src)	70.40	81.32	**67.12**
	AUC(tgt)	69.34	74.35	**59.64**
	pAUC	55.65	55.74	**54.74**
slider	AUC(src)	66.51	75.35	**68.88**
	AUC(tgt)	56.01	68.11	**59.40**
	pAUC	51.77	49.05	**53.05**
valve	AUC(src)	51.07	55.69	**94.16**
	AUC(tgt)	46.25	53.61	**68.16**
	pAUC	52.42	51.26	**63.16**

3.1. Discussion

Our method outperforms the baselines in several challenging machine types, particularly in the target domain where domain shift is more significant. Notable improvements are observed in machines like *valve*, *ToyTrain*, and *ToyCar*, demonstrating the robustness of patch-wise contrastive representation learning combined with attribute conditioning.

Despite slightly lower performance in the *fan* machine compared to Mahalanobis, our method shows more balanced detection across domains. This indicates that the proposed attention and metadata-aware fusion is effective in handling partial attribute availability.

The strong performance in the *valve* class suggests the model's strength in modeling aperiodic, impulse-heavy signals via localized patches, while the improvements in *ToyTrain* confirm the effectiveness of domain-aware Mahalanobis scoring.

Further ablation could study the individual contributions of attention pooling, metadata fusion, and patch-level representation in isolation.

4. CONCLUSION

In this report, we proposed a novel patch-wise contrastive learning framework for first-shot unsupervised anomalous sound detection under domain-shift and partial attribute conditions, as required by the DCASE2025 Task 2 challenge. Our system introduces attribute-conditioned attention pooling and a metadata-aware fusion mechanism that enables robust representation learning even when attribute labels are partially missing.

By leveraging ResNet-based encoders with strong data augmentation and contrastive objectives, the model learns localized anomaly-sensitive embeddings. At inference time, domain-specific Mahalanobis scoring allows for effective domain adaptation without explicit labels from the target distribution.

Experimental results on the DCASE2025 development dataset show that our method outperforms standard MSE and Mahalanobis baselines across multiple machine types, particularly under challenging target domain settings. The proposed system achieves strong generalization with minimal supervision, making it a promising approach for real-world industrial condition monitoring.

Future work includes extending this architecture with diffusion-based augmentation, transformer-based encoders, and a deeper analysis of patch-level anomaly localization.

5. REFERENCES

[1] Tomoya Nishida, Noboru Harada, Daisuke Niizumi, Davide Albertini, Roberto Sannino, Simone Pradolini, Filippo Augusti, Keisuke Imoto, Kota Dohi, Harsh Purohit, Takashi Endo, and Yohei Kawaguchi, "Description and discussion on DCASE 2025 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2506.10097*, 2025.

[2] Noboru Harada, Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Masahiro Yasuda, and Shoichiro Saito, "ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," in *Proc. DCASE Workshop*, pp. 1–5, Barcelona, Spain, 2021.

[3] Kota Dohi, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, Masaaki Yamamoto, Yuki Nikaido, and Yohei Kawaguchi, "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," in *Proc. DCASE Workshop*, Nancy, France, 2022.

[4] Noboru Harada, Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, and Masahiro Yasuda, "First-shot anomaly detection for machine condition monitoring: A domain generalization baseline," in *Proc. 31st European Signal Processing Conference (EUSIPCO)*, pp. 191–195, 2023.

[5] Md Moin, Abdullah H. Abdullah, Md Mehedi Hasan, and Abdun Naser Abdalla, "Self-supervised anomalous sound detection with statistical clustering and contrastive learning," *arXiv preprint arXiv:2210.09884*, 2022. (Authors affiliated with Universiti Kebangsaan Malaysia and Taibah University)

[6] Jonghoon Lee, Donghyeon Kim, Jinwoo Choi, and Heung-Il Kim, "Two-stage contrastive learning for anomalous sound detection," in *Proc. Detection and Classification of Acoustic Scenes and Events (DCASE)*, Tampere, Finland, 2023. (Korea Advanced Institute of Science and Technology, South Korea)