MACHINE ANOMALOUS SOUND DETECTION COMBINING CONVOLUTIONAL AUTO-ENCODER AND CONTRASTIVE LEARNING

Technical Report

Qing Zhou, Sai Wu

Xi'an University of Architecture and Technology
College of Information and Control Engineering, Xi'an, China

ABSTRACT

Machine anomalous sound detection (MASD) under noisy industrial conditions remains challenging due to limited anomalous samples, background noise interference, and domain shift. This paper proposes a multi-task learning framework combining a semi-supervised convolutional auto-encoder (CAE) with self-supervised classification and contrastive learning to address these issues. The core architecture uses a CAE backbone and the encoder output is projected into an audio embedding vector which is later fed into a linear classifier for self-supervised attribute classification (e.g., domain, operational parameters). Crucially, the framework leverages newly available clean machine data and noise-only data through a contrastive loss term. This loss pulls embeddings of noisy and clean machine samples of the same class closer while pushing those of noisy machine samples away from pure noise samples, enhancing noise robustness. The model is optimized jointly with a combined loss function integrating reconstruction, classification, and contrastive objectives. During inference, reconstruction errors and audio embeddings are concatenated as input features for a domain-aware anomaly detector. Evaluated on the DCASE2025 Task 2 dataset, the proposed method achieves a harmonic mean score of 63.80%, significantly outperforming the baseline. Ablation studies confirm each component's contribution, demonstrating the effectiveness of the multi-task strategy in learning discriminative and noise-invariant representations for MASD.

***Index Terms*—** Anomalous sound detection, convolutional aoto-encoder, contrastive learning, multi-task learning

1. INTRODUCTION

Monitoring machine conditions and detecting possible mechanical anomalies through acoustic signals is a key technology for smart factories. The DCASE Challenge Task 2 series [1] addresses several critical challenges in machine anomalous sound detection (MASD). These include: (1) scarcity of real-world and diverse anomalous sound samples, (2) interference from high-level background noise in the surroundings, and (3) domain shift caused by variable working conditions of machines. MASD is typically treated as an unsupervised task, using only normal machine sounds for training. Accurately identify abnormality across different devices and environments in practical factories still remains difficult. To enhance performance, the newly launched DCASE2025 Challenge Task 2 [2] supplements noisy normal machine sounds with additional clean machine data or noise-only recordings, captured during factory idle periods or machine inactivity.

State-of-the-art unsupervised MASD methods fall into two categories: reconstruction-based and self-supervised classification-based. Reconstruction-based approaches use auto-encoders to model the distribution of normal data [1, 3, 4, 5], where test samples with high reconstruction errors are flagged as anomalies. However, these methods exhibit noise sensitivity and limited generalizability, often misclassifying unseen data. Self-supervised classification-based approaches leverage the metadata (e.g., machine type, domain, operational parameters) as labels and use a proxy classification task to learn discriminative audio representations [6, 7, 8]. While these improve separability of normal sounds in the embedding space, they rely on auxiliary classification losses not directly optimized for anomaly detection.

To synergize these approches, this paper develops a framework for MASD combining reconstruction and self-supervised classification training. The backbone neural network is based on a convolutional auto-encoder (CAE) and the output of its encoder block is further projected into an audio embedding vector. A simple linear classifier is attached to this embedding for attribute classification. The model is jointly optimized using reconstruction and classification losses. For anomaly detection, reconstruction errors and audio embeddings are concatenated for modeling normal data distribution. Furthermore, the additional clean machine data and noise-only data are utilized to develop a third contrastive loss. This contrastive loss [9] enforces proximity between embeddings of noisy and clean samples of the same class while distancing embeddings of noisy machine samples from pure noise samples. This promotes noise-invariant embeddings that preserve essential machine characteristics. Experiments on the DCASE2025 Task 2 dataset demonstrate superior performance over baselines and validate the effectiveness of the multi-task learning strategy.

2. PROPOSED METHOD

The proposed framework for MASD integrates CAE reconstruction, self-supervised classification, and noise-aware contrastive learning, as depicted in Fig. 1. Note that separate models are trained per machine type using noisy normal machine samples in addition with clean machine data or noise-only data.

2.1. Convolutional Auto-encoder

The CAE performs reconstruction-based anomaly detection using normal training samples. Log-mel magnitude spectrograms are extracted from raw audio signals. Consecutive frames are combined to form a 2D data matrix as inputs. Denote the ith input by $\mathbf{X}_i \in \mathbb{R}^{F \times T}$, where F is the number of mel-scale filter banks and

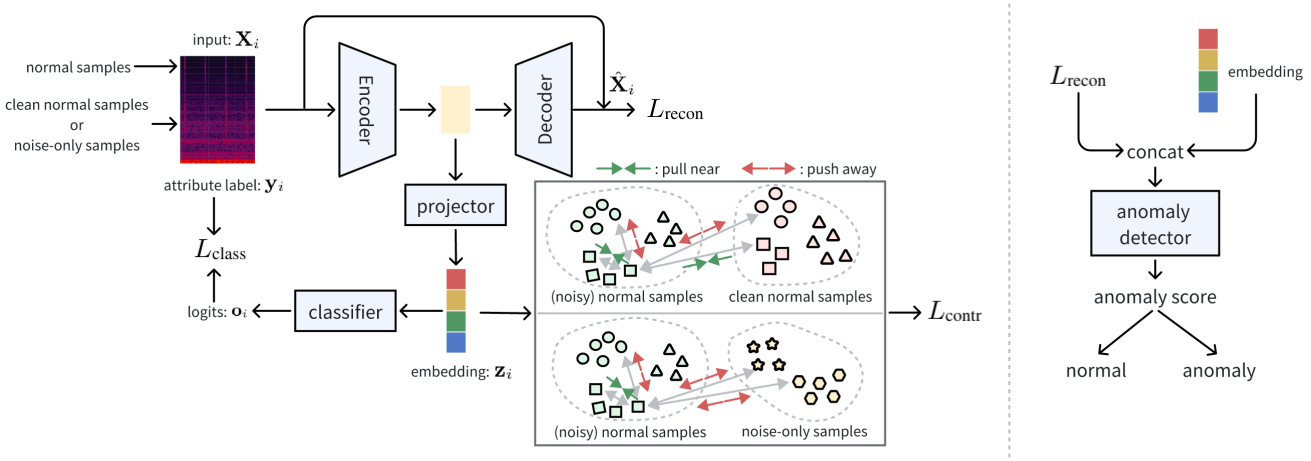

Figure 1: The proposed framework for machine anomalous sound detection

T is the number of time frames within the analysis window. The encoder consists of several convolutional blocks with max-pooling operations and maps $\mathbf{X}_i$ into a lower-dimensional latent representation. The decoder reconstructs the input using deconvolutional layers with up-sampling operations. The CAE network is trained to minimize the l_2-norm difference between the original input $\mathbf{X}_i$ and the reconstructed output $\hat{\mathbf{X}}_i$ by

$$L_{\text{recon}} = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{X}_i - \hat{\mathbf{X}}_i\|_2 \tag{1}$$

where N is the batch size.

After training, abnormal samples would produce large reconstruction errors which suggest high deviation from the training normal samples. Other than basic reconstruction-based learning, the encoder output is further projected into a d-dimensional audio embedding vector through global average pooling, denoted by $\mathbf{z}_i \in \mathbb{R}^d$. This compressed representation is combined with the reconstruction error for anomaly detection in later sections.

2.2. Self-supervised classification

To get discriminative representations, attribute information associated with machine sounds is used for self-supervised classification. For those machine types without attribute information, only domain labels are distinguished. A linear classifier maps the audio embedding $\mathbf{z}_i$ into logits output. A cross entropy (CE) loss is calculated between the groundtruth one-hot attribute label $\mathbf{y}_i$ and the output logits $\mathbf{o}_i$ as

$$L_{\text{class}} = \frac{1}{N}\sum_{i=1}^{N}\text{CE}(\mathbf{y}_i, \mathbf{o}_i) \tag{2}$$

2.3. Contrastive learning

Contrastive learning pulls similar samples within the same class closer together while pushing dissimilar samples from different classes further away [10, 11]. Given a batch of N samples, select the ith audio embedding $\mathbf{z}_i$ as the anchor. Among the remaining $N-1$ embeddings, those that has the same attribute class as $\mathbf{z}_i$ are considered positive samples and the others are considered negative samples. The contrastive loss is calculated to maximize the similarity between $\mathbf{z}_i$ and its positives:

$$L_{\text{contr}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(\text{sim}(\mathbf{z}_i,\mathbf{z}_p)/\tau)}{\sum_{j\neq i}\exp(\text{sim}(\mathbf{z}_i,\mathbf{z}_j)/\tau)} \tag{3}$$

where

$$\text{sim}(\mathbf{z}_i,\mathbf{z}_j) = \frac{\mathbf{z}_i^T * \mathbf{z}_j}{\|\mathbf{z}_i\|_2\|\mathbf{z}_j\|_2} \tag{4}$$

represents the cosine similarity function between two vectors, τ is a temperature scalar to scale the similarity scores, $P(i) = \{p|1 \leq p \leq N, \mathbf{y}_p = \mathbf{y}_i\}$ are the set of indexes of those positive samples of $\mathbf{z}_i$, and $|P(i)|$ is the number of indexes in $P(i)$.

Given additional clean normal samples or noise-only samples, we extend the concept of contrastive loss to acquire noise-robust representation from noisy training samples, as illustrated in Fig.1. For machine types with clean normal samples available, noisy and clean samples of the same attribute class are considered positive pairs for each other. For those machine types with pure noise samples available, noise samples are considered negatives for noisy machine samples. This strategy minimizes distances of embeddings of noisy and clean machine samples while pushing embeddings of noisy machine samples and pure noise samples further apart, thus guiding the model towards noise-robust representations.

Note that the classification task and contrastive learning task converge much faster than the reconstruction task. A two-stage training procedure is adapted:

Stage 1: Train the CAE model only with L_{recon};

Stage 2: Train the full model with the classifier by a triple loss weighted by empirically chosen parameters as:

$$L = L_{\text{recon}} + \alpha * L_{\text{class}} + \beta * L_{\text{contr}} \tag{5}$$

2.4. Anomaly Detector

The reconstruction error and the audio embedding vector are concatenated into a $d+1$-dimensional input feature to the anomaly detector. To tackle the domain shift issue, the SMOTE technique [12] is employed to over-sample the training vectors of the target

Table 1: Results on DCASE2025 Task 2 development datasets

Method	ToyCar		ToyTrain		bearing		fan		gearbox		slider		valve		hmean
	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	
baseline-MSE	62.28	49.70	59.11	50.19	59.84	61.12	54.85	49.46	57.64	52.49	59.43	52.32	65.35	57.35	56.26
baseline-MAHALA	62.04	49.05	48.51	48.32	61.33	61.86	58.27	50.82	62.43	55.07	62.03	**53.61**	58.61	52.53	55.33
the proposed:															
L_{recon}	60.32	50.37	65.62	48.68	62.96	**60.32**	62.26	57.47	55.21	53.26	61.17	50.47	65.35	63.47	57.79
$L_{recon} + L_{class}$	**65.31**	48.95	69.24	51.89	62.27	59.37	**64.97**	55.74	71.88	54.68	56.00	49.47	71.71	61.58	60.95
$L_{recon} + L_{class} + L_{contr}$	63.90	**53.58**	**71.25**	**52.89**	**63.24**	59.11	63.60	**61.32**	**73.87**	**57.00**	**63.82**	52.16	**79.53**	**68.16**	**63.80**

domain and domain-aware models are trained separately. K-nearest neighbor (KNN) and local outlier factor (LOF) are used for modeling normal data. For KNN, the abnormality is measured by the minimum distance to the cluster centers. For LOF, it is measured by the local density output to its neighbors. The anomaly score is calculated as the minimum of two domain models.

For testing, each query sound sample is segmented with overlap to generate several output representations. Different aggregation skills are employed: (1) feature-based aggregation where features of different segments are averaged and a single anomaly score is generated; (2) score-based aggregation where anomaly scores are calculated per segment and then averaged to a final score.

3. EXPERIMENTS

3.1. Experimental Settings

Experiments were conducted on the DCASE2025 Task 2 development dataset [13, 14] that consists of seven machine types (fan, gearbox, bearing, slider, ToyCar, ToyTrain, and valve). Each machine type contains 1000 normal training audio samples, of which 990 samples are in the source domain and 10 samples are in the target domain. The test data includes 200 samples of mixed normal and anomalous data. All audio samples are embedded in strong environmental noise. In addition, 100 samples of clean normal machine data or noise-only data are provided for each machine type. Four machine types are associated with rich attribute information while the rest three (bearing, slider, ToyTrain) have no attribute information. Each audio clip has a duration of 10s or 12s and a sampling rate of 16kHz.

For audio preprocessing, 128-dimensional log-mel energies were extracted from the raw signal with a window size of 1024 and a hop size of 512. It should be noted that a cut-off frequency of a high-pass filter was set for each machine type to suppress the low-frequency noises. As for the input to the network, 64 time frames (approximately of 2s in length) were combined to form an input feature of dimension $128 * 64$. The projection layer of the model consisted of global average pooling operations and the final audio embedding was reduced to a 128-dimensional vector.

For model training, a two-stage training strategy was adapted: the first stage included 100 epochs only with the reconstruction loss in (1) and the second stage included 250 epochs with the triple loss in (5). The learning rate was 0.001 and the batch size was set to 128. Because of the highly unbalanced distribution of source and target samples as well as noisy samples versus clean or noise-only samples in the training set, a batch sampler was realized to control the proportion of different sample types within each batch. Furthermore, the mixup technique [15] was applied to generate more intermediate-domain samples for enriching the training data and specaugment was employed to prevent overfitting. For other parameters, $\tau = 0.07, \alpha = 2, \beta = 1$. The official metrics of AUC, pAUC, and the harmonic mean score of the challenge were used for evaluation.

3.2. Results

Table 1 show the evaluation results of the proposed method compared to the official baseline system of the challenge. For ablation study, different training settings were investigated to verify the effectiveness of the strategies mentioned in Section 2:

1. training only with L_{recon};
2. training with L_{recon} and L_{class} in two stages;
3. training with the triple loss in (5) in two stages. These three systems were trained for the same total number of epochs for fair comparison.

As a result, the proposed method with the triple loss achieved the best performance with a harmonic mean score of 63.80%. Through ablation experiments, it can be seen that both the supervised classification and contrastive learning strategies contributed to the improvement of the performance. This demonstrated the effectiveness of multi-task learning attached to the reconstruction-based network. Specifically, integrating with the metadata for classification exhibited an overall performance increase except for a slight degradation on machine types (ToyTrain, bearing, slider) with no attribute information available. The contrastive learning displayed a consistent improvement over all machine types which demonstrated its ability of enhanced noise-robust representation learning.

For challenge evaluation, 4 systems were submitted which differed in the anomaly scoring settings. Specifically, system 1 used KNN and feature-based aggregation; system 2 used KNN and score-based aggregation; system 3 used LOF and feature-based aggregation; system 4 used LOF and score-based aggregation.

4. CONCLUSION

This paper presented a robust MASD framework by integrating reconstruction-based learning, self-supervised classification, and contrastive learning within a multi-task CAE architecture. The self-supervised classifier leverages metadata (domain/attributes) to learn discriminative embeddings, synergizing with the reconstruction objective. The contrastive task exploits auxiliary clean/noise-only data to explicitly enforce noise-invariant representations. Experiments on the challenging DCASE2025 Task 2 dataset validated the framework's effectiveness, achieving superior performance over the official baseline. Ablation studies confirmed each component's contribution: classification enhances discriminability (when metadata available), while contrastive learning consistently improves noise robustness across all machine types.

5. REFERENCES

[1] 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.

[2] http://dcase.community/challenge2025/.

[3] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 271–275.

[4] Y. Wang, Q. Zhang, W. Zhang, and Y. Zhang, "A lightweight framework for unsupervised anomalous sound detection based on selective learning of time-frequency domain features," *Applied Acoustics*, vol. 228, p. 110308, 2025.

[5] J. Guan, Y. Liu, Q. Kong, F. Xiao, Q. Zhu, J. Tian, and W. Wang, "Transformer-based autoencoder with id constraint for unsupervised anomalous sound detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, p. 42, 2023.

[6] A. Jiang, B. Han, Z. Lv, Y. Deng, W.-Q. Zhang, X. Chen, Y. Qian, J. Liu, and P. Fan, "Anopatch: Towards better consistency in machine anomalous sound detection," *arXiv preprint arXiv:2406.11364*, 2024.

[7] X.-M. Zeng, Y. Song, Z. Zhuo, Y. Zhou, Y.-H. Li, H. Xue, L.-R. Dai, and I. McLoughlin, "Joint generative-contrastive representation learning for anomalous sound detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[8] J. Guan, F. Xiao, Y. Liu, Q. Zhu, and W. Wang, "Anomalous sound detection using audio representation with machine id based contrastive learning pretraining," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[9] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3875–3879.

[10] H. Hojjati and N. Armanfard, "Self-supervised acoustic anomaly detection via contrastive learning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3253–3257.

[11] E. Zahedi, M. Saraee, F. S. Masoumi, and M. Yazdinejad, "Regularized contrastive masked autoencoder model for machinery anomaly detection using diffusion-based data augmentation," *Algorithms*, vol. 16, no. 9, p. 431, 2023.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[13] 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.

[14] 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.

[15] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.