SCAN: Selective Contrastive Learning Against Noisy Data for Acoustic Anomaly Detection

Zhaoyi Liu, Yuanbo Hou, Wenwu Wang *Senior Member, IEEE*, Sam Michiels and Danny Hughes

***Abstract*—Acoustic Anomaly Detection (AAD) has gained significant attention for the detection of suspicious activities or faults. Contrastive learning-based unsupervised AAD has outperformed traditional models on academic datasets, however, its model training is predominantly based on datasets containing only normal samples. In real industrial settings, a dataset of normal samples can still be corrupted by abnormal samples. Handling such noisy data is a crucial challenge, yet it remains largely unsolved. To address this issue, this paper proposes a Selective Contrastive learning framework Against Noisy data (SCAN) to mitigate the adverse effects of training the AAD model with anomaly-corrupted data. Specifically, SCAN progressively constructs confidence sample pairs based on the Mahalanobis distance, which is derived from the geometric median. These selected pairs are then integrated into the contrastive learning framework to enhance representation learning and model robustness. Extensive experiments under varying levels of label noise (i.e., the proportion of mislabeled abnormal samples in training data) demonstrate that SCAN outperforms state-of-the-art (SOTA) AAD methods on the real-world industrial datasets DCASE2022 and DCASE2024 Task2.**

***Index Terms*—Acoustic anomaly detection, unsupervised learning, contrastive learning, noisy data, confident pairs**

I. Introduction

Faults and failures in industrial machinery can significantly decrease its operational efficiency and product quality [1][2]. Identifying anomalies solely from operational machine sounds without requiring annotated data is known as unsupervised Acoustic Anomaly Detection (AAD) [3][4] and has gained increasing attention in both academic research and industrial applications. Anomalous acoustic signals may indicate system malfunctions or security threats, and early detection helps prevent operational failures and mitigate risks [5][6].

Due to the high cost of collecting comprehensive data for all possible anomalous behaviours, AAD systems are often developed using only normal data within an unsupervised learning framework [7]. Most previous unsupervised AAD methods rely on a well-defined dataset that contains only normal samples to establish a standard distribution, which is then utilized to determine whether a given test sample is normal or not [8]–[10]. However, excessive reliance on such data is risky. If the dataset of normal samples is contaminated by abnormal samples, the decision boundaries given by these models become unreliable, leading to a potential performance degradation in anomaly detection. Thus, the performance of previous mainstream methods is sensitive to noisy data [11][12].

In this paper, the term *noise* refers to label noise, meaning abnormal samples inadvertently present in training data assumed to be normal. In real-world industrial settings, such contamination is often unavoidable due to human misjudgment or distributional shifts over time. This reality challenges the clean-data assumption and affects the practical reliability of existing methods. While prior studies have effectively addressed data corruption from sensor or environmental noise [13]–[15], the specific challenge of label noise within the unsupervised AAD context remains largely unexplored. To address this critical gap, we propose a first fully unsupervised AAD framework that explicitly accounts for label noise without relying on manual data filtration. This design improves the robustness and practicality of anomaly detection in industrial quality inspection, enabling rapid deployment across diverse production lines with minimal human intervention.

Several studies have explored the effectiveness of Contrastive Learning (CL) for AAD, demonstrating its potential to substantially improve latent representations and enhance the accuracy of anomaly detection on benchmark datasets [10], [16]–[18]. However, such methods have not considered the impact of training the AAD model with normal data corrupted by anomalous samples. To fill this gap, we introduce a novel Selective Contrastive learning framework Against Noisy data (SCAN), by designing a *selective* process to identify the high-confidence samples (i.e., the normal samples) from the corrupted dataset and utilizing them during training.

More specifically, at each training epoch, we compute anomaly scores using the Mahalanobis distance [19]–[21], leveraging the geometric median to improve the robustness of the model. Based on these scores, confident pairs are identified from the dataset that may have been corrupted by abnormal samples using a thresholding strategy based on the Chi-squared distribution. Then, we utilize these pairs to facilitate the learning of robust latent representations within the CL framework, thereby reducing the adverse impact of abnormal samples on the AAD model training. We evaluated SCAN on DCASE2022 and DCASE2024 Task2 datasets. Extensive experiments under different levels of label noise demonstrate that the proposed method effectively handles AAD in the presence of partially incorrect or noisy training labels.

Z. Liu, S. Michiels and D. Hughes are with the Distrinet, Computer Science, KU Leuven, Belgium (e-mail: {zhaoyi.liu, sam.michiels, and danny.hughes}@kuleuven.be).

Y. Hou is with the Machine Learning Research Group, Engineering Science, University of Oxford, UK (email: Yuanbo.Hou@eng.ox.ac.uk)

W. Wang is with the Centre for Vision, Speech, and Signal Processing, University of Surrey, Guildford GU2 7XH, UK (e-mail: w.wang@surrey.ac.uk).

Detailed results and visualizations of hidden representations learned by SCAN are available at https://github.com/sherrylzy/SCAN.

This work is supported by VLAIO under grant number HBC.2024.0406.

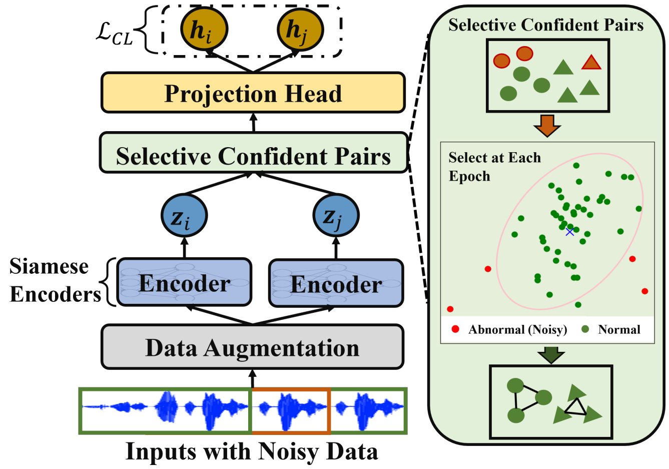


Fig. 1: An overview of SCAN. During the training, SCAN iteratively selects confident pairs to mitigate noisy data. At inference, each test sample's anomaly score is the Mahalanobis distance between its latent representation and the selected confident training samples.

II. Proposed Method

The proposed SCAN is shown in Fig. 1, where a novel selective CL framework is designed to address the adverse impact of noisy data on AAD performance. In this method, the geometric median-based Mahalanobis distance is used to provide a robust reference against variations in the data distribution [22][23]. In training, SCAN progressively selects a set of confident pairs from the training set $\mathcal{S}$, which helps mitigate the adverse impact of noisy data at each epoch, thereby enhancing the learning of robust acoustic representations. During inference, the Mahalanobis distances are used to derive the anomaly scores for the test samples.

A. Contrastive Latent Acoustic Representation Learning

The CL framework is used to derive acoustic representations from unlabeled data by creating multiple variations of an acoustic clip through data augmentation [17], [24]. Consequently, the proposed SCAN firstly augments each acoustic clip with five transformations operating in both the time and frequency domains, as described in [16].

Let $\mathbf{x}_k$ be the k-th audio clip, randomly selected from a training mini-batch of N raw audio clips. In the data augmentation process, we define $\mathcal{C}$ as the set of possible augmentation operators. For each acoustic clip $\mathbf{x}_k$, we randomly draw two operators from $\mathcal{C}$, which we denote as c^1 and c^2. These are then applied to produce a pair of augmented views: $\tilde{\mathbf{x}}^1_{2k-1} = c^1(\mathbf{x}_k)$ and $\tilde{\mathbf{x}}^2_{2k} = c^2(\mathbf{x}_k)$. Following the augmentation step, each augmented view is converted into a two-dimensional mel spectrogram, denoted as $\mathbf{X}_n \in \mathbb{R}^{F\times T}$. Here, the index $n \in \{1, \cdots, 2N\}$ represents elements in the augmented view sets.

A classical unsupervised CL framework [25] is subsequently employed to enhance acoustic representation learning by maximizing the similarity between its augmented versions from different perspectives while minimizing the similarity between distinct acoustic clips. Following the SimCLR [25], during SCAN training, we apply two stochastic augmentations to the N acoustic clips within a mini-batch, generating $2N$ spectrograms. These spectrograms are then fed into Siamese encoders [26], denoted as $\mathcal{E}(\cdot)$, which transform the input into lower-dimensional representations: $\mathbf{z}_n = \mathcal{E}(\mathbf{X}_n)$. Note that $\mathcal{E}(\cdot)$ comprises two identical neural networks that share weights. To further enhance the representations, a projection layer $g(\cdot)$ maps the latent representation $\mathbf{z}_n$ into a subspace to compute the sample-wise contrastive loss. The output of the projection head is represented as $\mathbf{h}_n = g(\mathbf{z}_n)$.

Each acoustic clip's augmented versions are considered positive pairs, while all other clips in the mini-batch are treated as negative pairs. The contrastive loss function promotes the alignment of latent representations for positive pairs while enforcing separation among negative pairs. The subscript (i, j) is solely used to index positive pair where $i = 2k - 1$ and $j = 2k$ for $k \in \{1, \cdots, N\}$ within the current mini-batch. Then, the loss function for a positive pair (i, j) is

$$\ell_{(i,j)} = -log\frac{\exp(sim(i,j)/t))}{\sum_{n=1}^{2N}\mathbb{1}_{[n\neq i]}\exp(sim(i,n)/t)}, \qquad (1)$$

where $\mathbb{1}_{[n\neq i]}$ is an indicator function, t denotes a temperature parameter, and $sim(i, j) = (\mathbf{h}_i \cdot \mathbf{h}_j)/(\|\mathbf{h}_i\|\|\mathbf{h}_j\|)$ represents cosine similarity among hidden acoustic representation. The final loss $\mathcal{L}_{\text{CL}}$ for the entire mini-batch is the average of this term computed symmetrically across all N positive pairs, defined as [27]:

$$\mathcal{L}_{CL} = \textstyle\sum_{k=1}^{N}(\ell_{(2k-1,2k)} + \ell_{(2k,2k-1)})/2N. \qquad (2)$$

B. Selecting Confident Pairs

To enhance robustness in contrastive learning under noisy label conditions, our approach progressively selects more confident pairs $\mathcal{S}$ out of noisy pairs to perform unsupervised CL. To achieve this, we estimate the anomaly scores using the Mahalanobis distance [19]–[21], based on the geometric median [22][23], which provides a robust and stable reference under distributional variation. We further introduce a thresholding strategy based on the Chi-squared distribution to identify noisy clips, by computing adaptive confidence thresholds at each epoch. The threshold increases progressively using a logarithmic schedule, starting conservatively and adjusted as representation learning improves.

Specifically, in each epoch e, for every mini-batch of N acoustic clips, we form a temporary set of $2N$ latent representations $\mathcal{Z}^{(e)} = \{\mathbf{z}^{(e)}_n\}^{2N}_{n=1}$. This batch-specific set is then used to compute anomaly scores and select confident pairs for that training step. Here, d denotes the dimension of the latent representation. Then, the geometric median $\boldsymbol{\mu}^{(e)} \in \mathbb{R}^d$ of $\mathcal{Z}^{(e)}$ is calculated as:

$$\boldsymbol{\mu}^{(e)} = \arg\min_{\mathbf{y}\,\in\,\mathbb{R}^d} \sum_{\mathbf{z}\,\in\,\mathcal{Z}^{(e)}} \|\mathbf{z} - \mathbf{y}\|, \qquad (3)$$

where $\mathbf{y}$ is a candidate center in $\mathbb{R}^d$. Here, Weiszfeld's approximation [28] is used to estimate the geometric median.

The anomaly score $s^{(e)}_n$ of a spectrogram $\mathbf{X}_n$ at epoch e is computed using the Mahalanobis distance, based on its corresponding latent representation $\mathbf{z}^{(e)}_n$, as follows:

$$s^{(e)}_n = (\mathbf{z}^{(e)}_n - \boldsymbol{\mu}^{(e)})^T(\boldsymbol{\Sigma}^{(e)})^{-1}(\mathbf{z}^{(e)}_n - \boldsymbol{\mu}^{(e)}), \qquad (4)$$

where $\Sigma^{(e)}$ is the covariance of the training data at epoch e.

Since the true Chi-squared distribution [29] can be approximated by Mahalanobis distances, we adopt the Chi-squared-based threshold to detect noise. In particular, given a confidence level α, the upper threshold $\Theta^{(e)}$ is determined from the Chi-squared distribution χ^2_p, where p is the dimension of $\mathbf{z}_n$, equivalent to d, as follows:

$$\Theta^{(e)} = \chi^2_p(1-\alpha). \tag{5}$$

The acoustic clip with anomaly scores $s_n^{(e)}$ exceeding a threshold $\Theta^{(e)}$ are treated as noisy and excluded from gradient propagation during epoch e to reduce the influence of uncertain or potentially mislabeled data.

During the early stages of training, the CL model may have limited capability to learn robust latent representations, leading to less reliable anomaly detection based solely on Mahalanobis distance scores. To address this, we gradually adjust the confidence level for detecting noise during training, ensuring that the threshold value steadily approaches its upper limit as training progresses, by employing a logarithmic function as the balancing factor as follows,

$$\Theta_p^{(e)} = \Theta^{(e)} \times \left(1 - 1/\log(e+\epsilon)\right), \tag{6}$$

where ϵ is a small constant for numerical stability.

Then, acoustic clips with scores exceeding $\Theta_p^{(e)}$ are considered noisy (i.e., potentially anomalous) and excluded from training. Only acoustic clips with scores below the threshold are retained, and the contrastive loss is computed using pairs where both scores fall below $\Theta_p^{(e)}$ to ensure robust representation learning. The set of confident pairs is denoted as $\mathcal{S}$:

$$\mathcal{S} := \left\{ \left(\mathbf{z}_i^{(e)}, \mathbf{z}_j^{(e)}\right) \mid \forall\, v \in \{i,j\},\ \mathrm{score}(\mathbf{z}_v^{(e)}) \le \Theta_p^{(e)} \right\}. \tag{7}$$

To mitigate the impact of noisy data, the contrastive loss is computed only on selected confident pairs in $\mathcal{S}$. That is, in Eq. 1 and Eq. 2, the loss calculation is restricted entirely to the set of confident pairs $\mathcal{S}$ to reduce the influence of potential noisy data. During inference, the Mahalanobis distance as in Eq. 4 is computed, where the geometric median $\boldsymbol{\mu}$ and the covariance $\boldsymbol{\Sigma}$ are estimated from the final epochs using the selected confident training samples. This distance serves as the anomaly score, measuring the deviation of the latent representation of the query sample $\mathbf{z}_{\text{test}} \in \mathbb{R}^d$ from the pre-estimated geometric median, which is compared to the threshold learned during training to determine whether the sample is anomalous.

III. EXPERIMENTS

A. Experimental Setting

1) Dataset: We evaluate SCAN on industrial datasets from DCASE2022 [3] and DCASE2024 Task2 [4], using real-world machine types to assess robustness under diverse conditions. The datasets differ as follows: (a) DCASE2022 Task2 includes five machine types—bearing, fan, gearbox, slider, and valve—with domain shift between training and test data. Each machine type has 3000 training samples in the source domain and only 30 in the target domain, with variation caused by changes in operator settings or environmental factors. (b) DCASE2024 Task2 adds complexity by introducing unseen machine types in the test set, including 3D printer, air compressor, brushless motor, hair dryer, hovering drone, robotic arm, scanner, and toothbrush. Each unseen machine type has 200 test samples. Despite these differences, both datasets aim to distinguish normal from abnormal acoustic samples, regardless of domain variations or machine types.

2) Noisy Data Settings: The number of normal samples is the same as that in the noiseless setting. By injecting varying proportions of anomalous samples, we generate noisy datasets labelled as "noise-*r*%," where *r*% denotes the noise ratio. Anomalous samples, randomly selected from the DCASE2022 Task2 test set, are added to the normal training set as noise. These injected anomalies remain in the test set, creating label conflicts where they are treated as normal during training but anomalous during inference. This creates a more challenging yet realistic AAD benchmark under noisy conditions. Due to the limited number of abnormal samples in DCASE2022, the noise ratio is capped at 8% in our experiments.

3) Implementation Details: The input feature is the mel spectrogram, computed using a Hann window of length 2048 with 50% overlap and 128 mel filter banks. The encoder $\mathcal{E}(\cdot)$ is based on the ResNet-18 architecture [30], generating a 512-dimensional linear output vector. To further refine the latent space, a projection head, implemented as a Multi-Layer Perceptron (MLP), consists of a 512-unit hidden layer followed by a 128-unit output layer, producing the final representation vectors. The confidence level is set to $\alpha = 0.05$, while the temperature parameter t is empirically chosen as 0.007. Optimization is performed using the AdamW optimizer [31] with a batch size of 128. The initial learning rate is selected from four logarithmically spaced values between 0.0005 and 0.01 and is adjusted using cosine annealing over 100 training epochs. The model is trained for a total of 400 epochs. Each experiment was repeated 10 times.

B. Results and Analysis

The SCAN-based model is trained on DCASE2022 Task2 training data, sharing the same machine types as DCASE2024 Task2 training set. We compare SCAN with fully unsupervised classic AAD methods and SOTA CL-based approaches without relying on meta-information like machine ID. Classic methods include AE-AAD [3], MobileNetV2 [4], and IDNN [32], alongside CLF-AIAD [16] and AADCL [17] as advanced CL-based frameworks. Additionally, we benchmark SCAN against the top five AUC-ranked methods from DCASE2022 Task2 [33]–[37] and DCASE2024 Task2 [38]–[42], adopting their official leaderboard AUC scores as the leading baseline results.

1) Robustness with Varying Noise Ratios: To analyze how different methods respond to increasing noise levels, we conduct experiments on DCASE2022 Task2 using varying noise ratios $\{0,\ 2,\ 4,\ 6,\ 8\}(\%)$, following the settings in Section III-A2. As shown in Fig. 2, fully unsupervised AAD methods without noise handling strategies exhibit a significant decrease in performance as the noise ratio increases. In contrast, the proposed SCAN demonstrates greater robustness, consistently achieving the most robust performance across varying noise levels and different domains. Furthermore,

TABLE I: Comparison between presented and 5 top-performing systems in Task2 of DCASE2022 and DCASE2024.

Method	AE_Baselines	Top-k AUCs					Proposed SCANs AUCs				
		Top-1	Top-2	Top-3	Top-4	Top-5	SCAN-0%	SCAN-2%	SCAN-4%	SCAN-6%	SCAN-8%
DCASE2022	61.5	77.13	74.87	73.72	69.7	68.22	79.78	78.56	78.22	77.86	77.02
DCASE2024	60.6	73.98	72.81	69.04	61.35	61.09	74.53	74.11	73.52	72.98	71.63

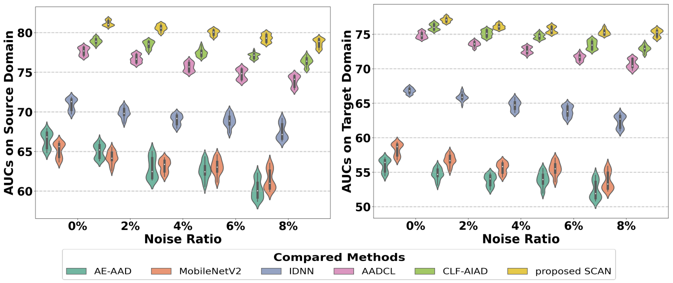
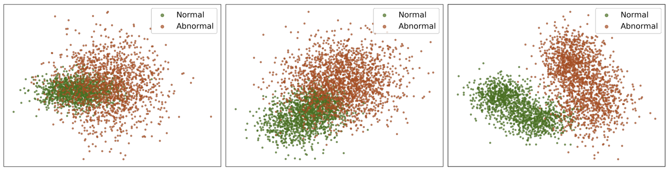


Fig. 2: Comparison of performance trends on source and target domains in the DCASE 2022 for the proposed SCAN, baselines, and advanced methods with increasing noise ratios (0% to 8%).



(i) AE-AAD (ii) CLF-AIAD (iii) Proposed SCAN

Fig. 3: t-SNE visualization of extracted features for bearing test data in the DCASE 2022 under 8% noise.

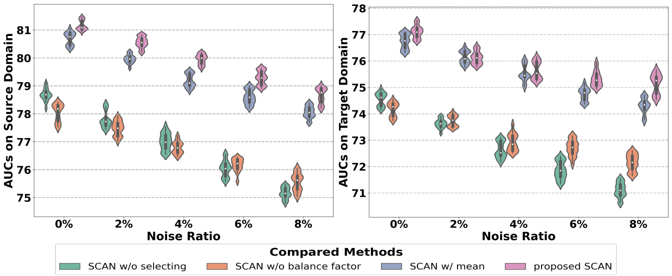


Fig. 4: Comparison of performance trends on source and target domains in the DCASE 2022 for the proposed SCAN and its variations for increasing noise ratios (0% to 8%).

TABLE II: Comparison of SCAN variants with 8% noise.

Dataset	DCASE 2022	DCASE 2024
SCAN w/o Selecting	73.19±1.09	69.97±1.11
SCAN w/o Balance Factor	73.66±1.10	70.14±1.05
SCAN w/ Mean	75.14±1.24	71.11±1.16
Proposed SCAN	**77.02±0.77**	**71.63±0.86**

CL-based methods (proposed SCAN, AADCL, and CLF-AIAD) exhibit stronger stability and higher performance than reconstruction-based approaches (AE-AAD and IDNN), which are more sensitive to input noisy data. However, as the noise level increases, the performance gap between the proposed SCAN and SOTA CL-based methods (AADCL and CLF-AIAD) becomes more noticeable, demonstrating the stronger resistance of SCAN against noise. Additionally, methods without noise-mitigation strategies suffer more severe performance degradation in the target domain as the noise level increases. This issue is particularly evident in real-world industrial environments, where external factors such as temperature fluctuations, varying loads, and operational inconsistencies further amplify the discrepancies in abnormality labels or domain alignments between initial and final calibrations. Building on this trend analysis, we further examine the feature extraction capabilities of AE-AAD, CLF-AIAD, and the proposed SCAN using t-SNE for clustering and visualization, as shown in Fig. 3. The proposed SCAN produces more distinct features, indicating its superior ability to capture latent representations under noisy data conditions.

2) Comparison to Top-performing Systems: The SCAN framework is evaluated against five top-performing systems from the DCASE2022 and DCASE2024 Challenges. In Table I, SCAN consistently outperforms baselines and top systems with higher AUC scores across all noise levels, even without noise (SCAN-0%). This demonstrates the benefits of selecting confident pairs to improve the CL performance. In DCASE2024, when the noise ratio exceeds 8%, the performance gap between SCAN and the top-1 system remains small, demonstrating its strong robustness even when it is trained on noisy data.

3) Ablation Study: To evaluate the effectiveness of the proposed method, ablation experiments were conducted on three SCAN variations: (1) SCAN w/o selecting, which removes the selecting confident pairs module; (2) SCAN w/o balance factor, which applies the final threshold value in Eq. 5 at the start of training; and (3) SCAN w/ mean, which uses the mean vector instead of the geometric median for Mahalanobis distance calculation. Fig. 4 and Table II present the AUC results of SCAN and its variations for different noise ratios and different industrial challenge datasets under a high noise level, respectively. First, SCAN w/o selecting experiences a sharp performance decrease as noise increases, underscoring the importance of noise mitigation strategies. Second, SCAN without balance factor further deteriorates under high noise conditions and across different datasets, emphasizing the need for progressively increasing threshold confidence to enhance acoustic representation learning and model stability. Lastly, SCAN with mean underperforms (Fig. 4) the geometric median, which offers better robustness in real-world scenarios.

IV. Conclusions

To handle noisy data in real-world anomaly detection, we introduced SCAN, a selective CL framework. SCAN effectively identifies potential noise through the Mahalanobis distance with the geometric median, enabling a progressive construction of confident pairs to learn robust latent representations within the CL framework. Our extensive experiments on multiple industrial datasets under noisy conditions have demonstrated the SOTA performance of SCAN. A potential future work is to enhance the framework's scalability.

References

[1] G.-Q. Zeng, Y.-W. Yang, K.-D. Lu, G.-G. Geng, and J. Weng, "Evolutionary adversarial autoencoder for unsupervised anomaly detection of industrial internet of things," *IEEE Trans. Reliab.*, pp. 1–15, 2025.

[2] N. Chander and M. U. K., "Enhanced pelican optimization algorithm with ensemble-based anomaly detection in industrial internet of things environment," *Clust. Comput.*, vol. 27, no. 5, pp. 6491–6509, 2024.

[3] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Description and discussion on dcase 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," in *Proc. Detect. Classif. Acoust. Scenes Events 2022 Workshop (DCASE2022)*, Nov. 2022, pp. 1–5.

[4] 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. Eur. Signal Process. Conf. (EUSIPCO)*, 2023, pp. 191–195.

[5] V. Zavrtanik, M. Marolt, M. Kristan, and D. Skočaj, "Anomalous sound detection by feature-level anomaly simulation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024, pp. 1466–1470.

[6] R. Ikegami, R. Kainuma, and S. Yano, "Anomalous sound detection system in manufacturing industry using unsupervised learning," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, 2024, pp. 1–4.

[7] K. Wilkinghoff and F. Kurth, "Why do angular margin losses work well for semi-supervised anomalous sound detection?" *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 608–622, 2023.

[8] Z. Liu, H. Tang, S. Michiels, W. Joosen, and D. Hughes, "Unsupervised acoustic anomaly detection systems based on gaussian mixture density neural network," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2022, pp. 259–263.

[9] K. Wilkinghoff, "Self-supervised learning for anomalous sound detection," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024, pp. 276–280.

[10] 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 *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023, pp. 1–5.

[11] J. Zhang, Z. Liu, C. Chatzichristos, S. Michiels, W. V. Paesschen, D. Hughes, and M. D. Vos, "Select for better learning: Identifying high-quality training data for a multimodal cyclic transformer," *J. Neural Eng.*, 2025.

[12] C. Wang, X. Jiang, B.-B. Gao, Z. Gan, Y. Liu, F. Zheng, and L. Ma, "Softpatch+: Fully unsupervised anomaly classification and segmentation," *Pattern Recognit.*, vol. 161, p. 111295, 2025.

[13] Y. Zhang, Y. Chen, J. Wang, and Z. Pan, "Unsupervised deep anomaly detection for multi-sensor time-series signals," *IEEE Trans. Knowl. Data Eng. (TKDE)*, vol. 35, no. 2, pp. 2118–2132, 2023.

[14] Y. Zhang, J. Wang, Y. Chen, H. Yu, and T. Qin, "Adaptive memory networks with self-supervised learning for unsupervised anomaly detection," *IEEE Trans. Knowl. Data Eng. (TKDE)*, vol. 35, no. 12, pp. 12 068–12 080, 2023.

[15] M. Yao, D. Tao, P. Qi, and R. Gao, "Rethinking discrepancy analysis: Anomaly detection via meta-learning powered dual-source representation differentiation," *IEEE Trans. Autom. Sci. Eng. (T-ASE)*, vol. 22, pp. 8579–8592, 2025.

[16] Z. Liu, Y. Hou, H. Tang, Á. López-Chilet, S. Michiels, D. Botteldooren, J. A. Gómez, and D. Hughes, "Clf-aiad: A contrastive learning framework for acoustic industrial anomaly detection," in *Proc. Int. Conf. Neural Inf. Process. (ICONIP)*, 2024, pp. 125–137.

[17] H. Hojjati and N. Armanfard, "Self-supervised acoustic anomaly detection via contrastive learning," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2022, pp. 3253–3257.

[18] Z. Liu, Á. López-Chilet, D. Chong, S. Michiels, J. A. Gómez, F. Wolf-Monheim, D. Newton, and D. Hughes, "Srad-clf: Squeak and rattle anomaly detection via contrastive learning framework on real industrial noise recordings," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2024, pp. 181–185.

[19] P. C. Mahalanobis, "On the generalized distance in statistics," *Sankhya: Indian J. Stat.*, vol. 80, pp. S1–S7, 2018.

[20] M. Yao, D. Tao, P. Qi, and R. Gao, "Scalable large model for unlabeled anomaly detection with trio-attention u-transformer and manifold-learning siamese discriminator," *IEEE Trans. Serv. Comput.*, vol. 18, no. 2, pp. 1012–1025, March-April 2025.

[21] Y. Yang, C. Zhang, T. Zhou, Q. Wen, and L. Sun, "Dcdetector: Dual attention contrastive representation learning for time series anomaly detection," in *Proc. ACM SIGKDD Conf. Knowl. Discov. Data Min. (KDD)*. New York, NY, USA: ACM, 2023, pp. 3033–3045.

[22] A. Acharya, I. S. Dhillon, and S. Sanghavi, "Geometric median matching for robust data pruning," in *ICML Workshop Found. Models Wild*, 2024.

[23] E. Cabana, R. E. Lillo, and H. Laniado, "Multivariate outlier detection based on a robust mahalanobis distance with shrinkage estimators," *Stat. Pap.*, vol. 62, pp. 1583–1609, 2021.

[24] X. Wang and G.-J. Qi, "Contrastive learning with stronger augmentations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5549–5560, 2023.

[25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1597–1607.

[26] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 15 750–15 758.

[27] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Adv. Neural Inf. Process. Syst.*, vol. 29, 2016.

[28] E. Weiszfeld and F. Plastria, "On the point for which the sum of the distances to n given points is minimum," *Ann. Oper. Res.*, vol. 167, pp. 7–41, 2009.

[29] D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt, "Unsolved problems in ml safety," in *arXiv preprint arXiv:2109.13916*, 2021.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[32] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020, pp. 271–275.

[33] I. Kuroyanagi, T. Hayashi, K. Takeda, and T. Toda, "Two-stage anomalous sound detection systems using domain generalization and specialization techniques," Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge, Tech. Rep., 2022.

[34] Y. Zeng, H. Liu, L. Xu, Y. Zhou, and L. Gan, "Robust anomaly sound detection framework for machine condition monitoring," Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge, Tech. Rep., 2022.

[35] F. Xiao, Y. Liu *et al.*, "Anomalous sound detection with self-supervised attribute classification and gmm-based clustering," Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge, Tech. Rep., 2022.

[36] Y. Deng, J. Liu, and W.-Q. Zhang, "Aithu system for unsupervised anomalous detection of machine working status via sounding," Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge, Tech. Rep., 2022.

[37] S. Venkatesh *et al.*, "Disentangled surrogate task learning for improved domain generalization in unsupervised anomalous sound detection," Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge, Tech. Rep., 2022.

[38] Z. Lv, A. Jiang, B. Han, Y. Liang *et al.*, "Aithu system for first-shot unsupervised anomalous sound detection," Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge, Tech. Rep., 2024.

[39] A. Jiang, X. Zheng, Y. Qiu, W. Zhang, B. Chen *et al.*, "Thuee system for first-shot unsupervised anomalous sound detection," Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge, Tech. Rep., 2024.

[40] R. Zhao, K. Ren, and L. Zou, "Enhanced unsupervised anomalous sound detection using conditional autoencoder for machine condition monitoring," Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge, Tech. Rep., 2024.

[41] J. Yang, "Adaptive framework for first-shot unsupervised anomalous sound detection in industrial machine monitoring," Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge, Tech. Rep., 2024.

[42] L. Wang, "Two-step anomaly detection: Integrating attribute classification and generative modeling with attribute inference for diverse machine types," Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge, Tech. Rep., 2024.