# Anomalous Sound Detection

**Wenwu Wang**

Centre for Vision, Speech and Signal Processing (CVSSP)
& Surrey Institute for People Centred Artificial Intelligence

**University of Surrey**

United Kingdom

Email: w.wang@surrey.ac.uk
Web: https://personalpages.surrey.ac.uk/w.wang/

Keynote talk on the MAD Workshop, Aalborg, Denmark
02/02/2026

# Many thanks to my collaborators and co-authors:

UNIVERSITY OF SURREY

- Jian Guan
- Zhaoyi Liu
- Feiyang Xiao
- Haiyan Lan
- Hejing Zhang
- Kejia Zhang
- Qiaoxi Zhu
- Youde Liu
- Jiantong Tian
- Qiuqiang Kong
- Sam Michiels
- Danny Hughes
- Haohe Liu
- Xinhao Mei
- Xubo Liu
- Shiheng Zhang
- Yuming Wei
- Wenbo Wang
- …

# Outline

- **Introduction**
  - What is anomalous sound detection (ASD)?
  - Potential applications, data sets, metrics, open problems
- **Example solutions to open problems**
  - Unsupervised learning problem
  - Domain shift problem
  - Few shot problem
  - Noisy data/labels problem
- **Conclusions and future works**

# What is Anomalous Sound Detection?

- **Anomalous Sound Detection** is about detecting **anomalies** (e.g. in a device, equipment, or an environment) through listening to the sounds produced by them or presented in them.

- Examples include detecting **engine faults or failures** through listening to engine sound, **detecting crimes** through listening to screaming and gunshot sounds, and detecting air **intrusions** via listening to drone or UAV sounds.

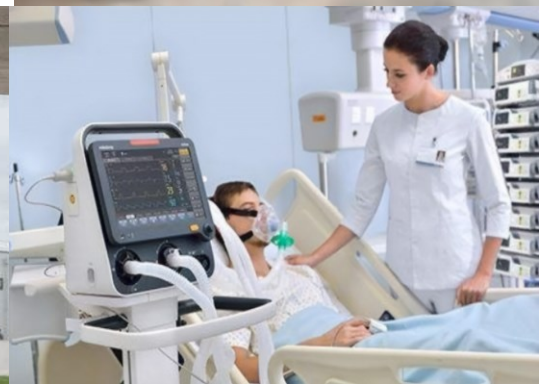
- **Human observers**: time consuming, expensive, relying on expert experience, limited listening duration, and low accuracy

- **Automated ASD via algorithms**: autonomous, highly efficient, cost-effective, continuous monitoring, and mitigating safety risk of human observation in hazardous conditions

# Potential Applications

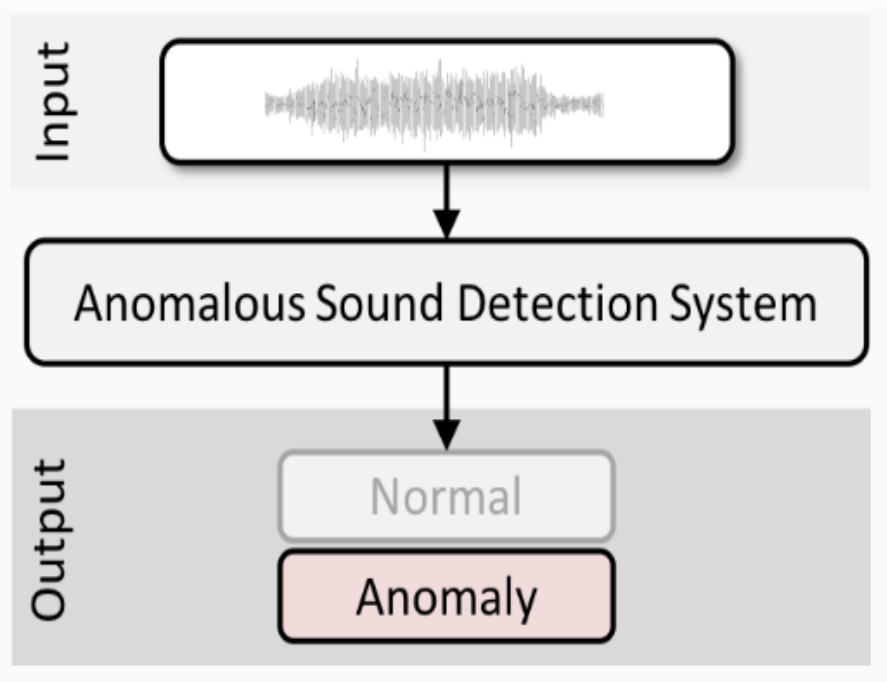
- Industrial manufacturing
- Automotive manufacturing
- Aerospace
- Rail transportation
- Power generation
- Healthcare
- Defense
- Surveillance
- Cyber security
- .....

# Potential Impacts

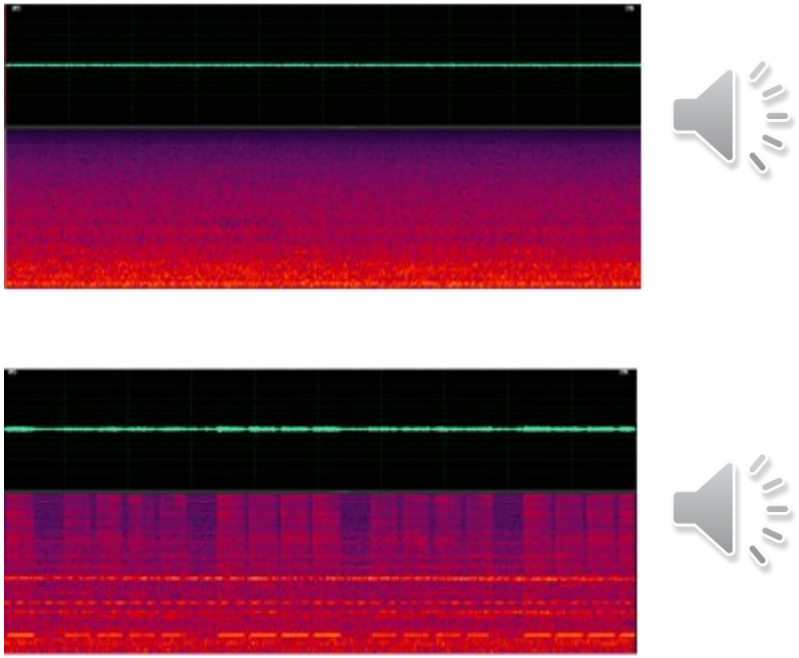


- Improved product quality
- Enhanced equipment reliability
- Reduced failure rates of energy infrastructure
- Strengthened aerospace safety
- Reliable operation of medical equipment
- .....

# An Example




A typical industrial ASD system



Examples of Normal and Abnormal Sounds

# Datasets

**ToyADMOS** (Koizumi et al, 2019):

- 160 hours normal hours & 4000+ anomalous samples, recorded with 4 mics, sampled at 48kHz.
- Three subsets: product inspection (toy car), fault diagnosis of stationary (toy conveyor) and moving machines (toytrain)

**MIMII** (Purohi et al, 2019):

- 26,092 normal and 6,065 anomalous sound segments, each lasting 10 seconds, 8-mics, sampled at 16kHz.
- Four types of industrial machines, e.g., valves, pumps, fans, and slide rails.
- Each with multiple models and operating conditions under both normal and anomalous states (contamination, leakage, rotating unbalance, and rail damage).

# Datasets

**ToyADMOS2** (Harada et al, 2021):

- 27,000 normal and over 8,000 anomalous machine sounds, recorded at 48 kHz using between 5 and 8 microphones.
- Two machine types, i.e., a toy car for product inspection and a toy train for fault diagnosis.
- Four sets of separately recorded background noises, such as outdoor and air-conditioning noise.
- Domain shift, robustness to variability, and ASD performance under diverse operating conditions.

**MIMII-DG** (Dohi et al, 2022):

- 42,000 normal and 6,000 anomalous ten-second machine sounds recorded at 16 kHz.
- Five machine types, i.e., fan, gearbox, bearing, slide rail, and valve.
- Anomalies: gear damage, fan imbalance, valve contamination, and rail defects. Proposed for domain shift.

# Datasets

**ToyADMOS2+** (Harada et al, 2023):

- Four additional machine types: ToyDrone, ToyNscale, Vacuum, and ToyTank.
- 4 to 8 microphones (with ToyDrone using 3) at a 48 kHz sampling rate and later downsampled to 16 kHz for training, with time durations ranging from 6 to 18 seconds.
- Anomalies: damaged propellers in ToyDrone, flat tires in ToyNscale, clogged nozzles in Vacuum, and damaged belts in ToyTank.
- Proposed for few-shot scenario.

# Performance Metrics

The performance of ASD is commonly evaluated using the

- **area under the receiver operating characteristic (ROC) curve (AUC)**

- **partial AUC (pAUC)**

The pAUC represents the AUC computed over a specific portion of the ROC curve within a predefined range of interest. In unsupervised ASD, it is typically calculated over a low false positive rate (FPR) range [0, p]. It aims at maximizing the true TPR within a low FPR range is of particular importance.

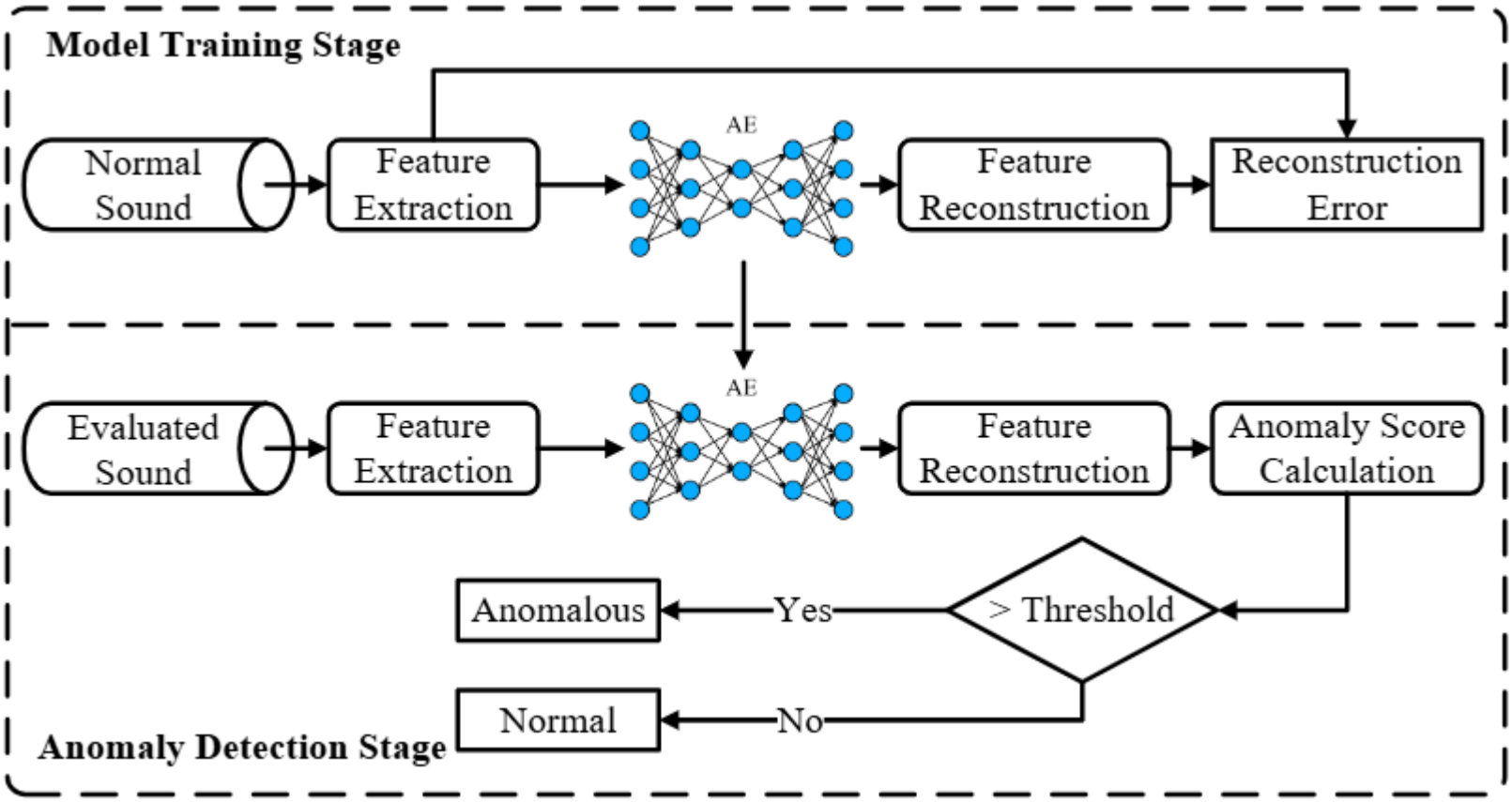
# Open Problems

- **Data scarcity and annotation difficulty**: Anomalous sounds are often rare and highly diverse. Existing models are often trained using only normal data, with model selection and hyperparameter tuning heavily relying on expert experience, leading to limited interpretability and transferability.
- **Domain shift or changes:** Acoustic characteristics vary significantly across operating environments and conditions, causing models trained under specific conditions (source domain) to perform poorly when applied to unseen conditions (target domain).
- **Lack of scenario transferability:** Models trained for specific devices in a single scenario struggle to adapt to diverse production lines and equipment types, limiting their practical applicability in industrial settings, for example, for unseen anomalies.
- **Data or labels are noisy:** Data used for training a model might be contaminated by noise either in the samples or in the labels.

# Solutions to Open Problems

- **Unsupervised Learning**
  - Audio feature reconstruction
  - Label classification
- **Domain Shift**
  - Domain mixing
  - Domain classification
- **First-shot Problem**
  - Data augmentation
  - Pre-trained models
  - Synthetic supervision
- **Noisy Data**
  - Selective contrastive learning

# **Problem 1: Unsupervised Learning** - Feature Reconstruction-Based Methods

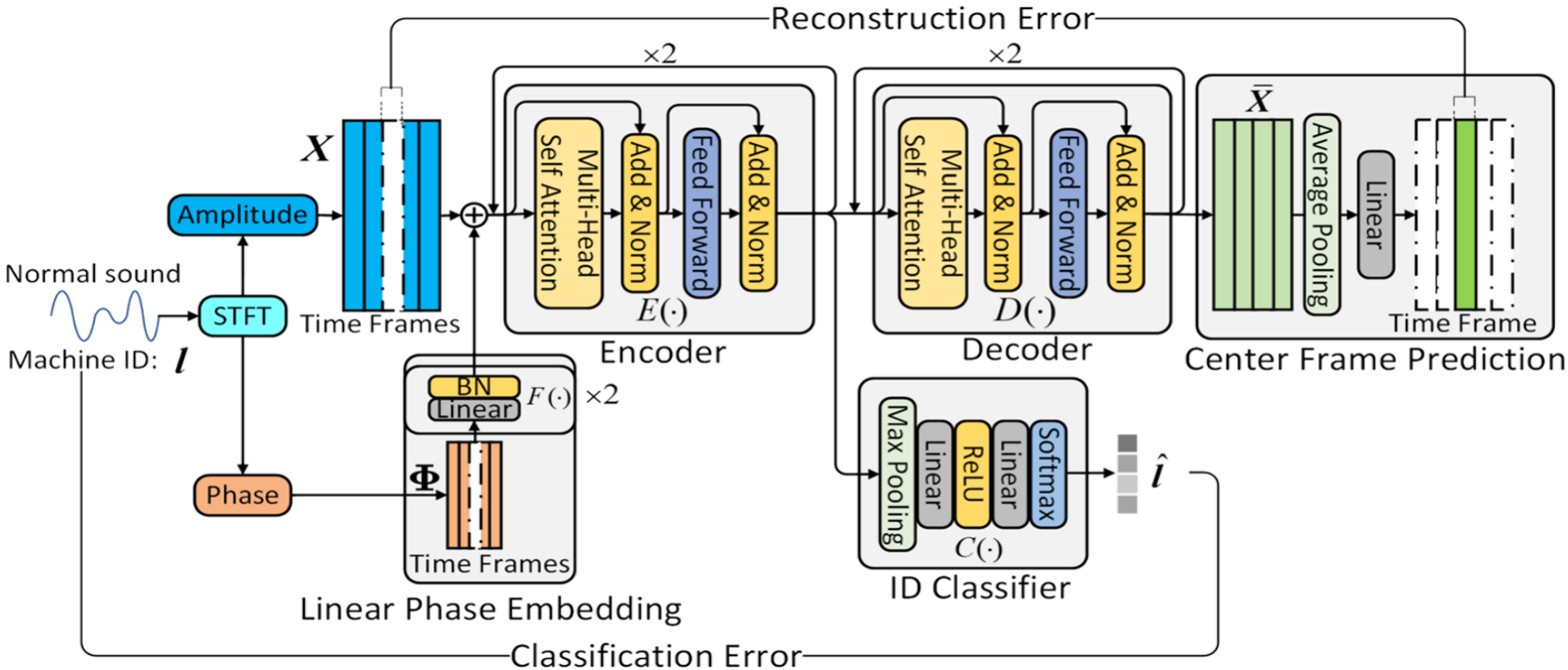


# Problem 1: Unsupervised Learning
## - Feature Reconstruction

- DCASE AE Baseline (Koizumi et al, 2020)
  - DCASE Challenge baseline, MLP-based autoencoder
- IDNN (Suefusa et al, 2020)
  - Predicting and reconstructing intermediate-frame features, enabling the model to capture time-varying characteristics, & in detecting short-term non-stationary signals
- IDC-TransAE (Guan et al, 2023)
  - ID constraint, phase information encoding and centre frame prediction

Y. Koizumi, Y. Kawaguchi, K. Imoto, , et al., "Description and discussion on DCASE2020 challenge task2: unsupervised anomalous sound detection for machine condition monitoring," in Proc. Detect. Classif. Acoust. Scenes Events 2020 Workshop (DCASE), Nov. 2020, pp. 81–85.
K. Suefusa, T. Nishida, H. Purohit, et al., "Anomalous sound detection based on interpolation deep neural network," in Proc. Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2020, pp. 271-275.
J. Guan, Y. Liu, Q. Kong, et al., "Transformer-based autoencoder with ID constraint for unsupervised anomalous sound detection," EURASIP J. Audio Speech Music Process., vol. 2023, no. 42, 2023.

# An Example Method: IDC-TransAE




- ID constraint
- Phase information encoding
- Centre frame prediction
- Weighted anomaly score calculation

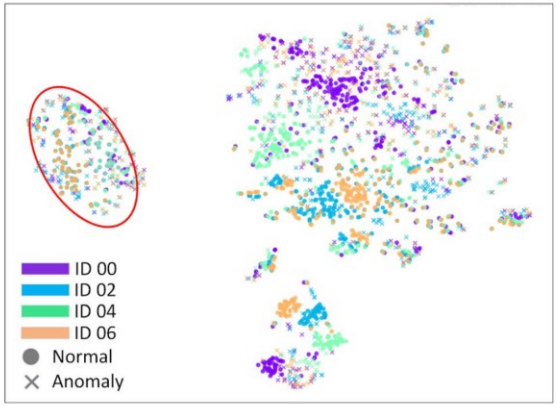
J. Guan, Y. Liu, Q. Kong, et al., "Transformer-based autoencoder with ID constraint for unsupervised anomalous sound detection," EURASIP J. Audio Speech Music Process., vol. 2023, no. 42, 2023.

# IDC-TransAE Method

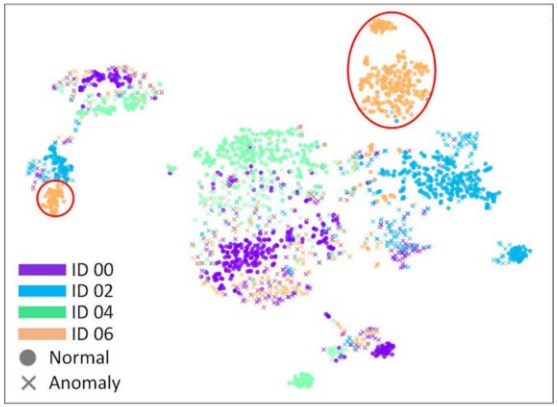
## Effectiveness of ID constraint

| | TransAE/LPE/CFP-W | | IDC-TransAE-W | |
|---|---|---|---|---|
| | AUC | pAUC | AUC | pAUC |
| Fan | 73.91 | 54.14 | **80.44** | **70.21** |
| Pump | 77.31 | 68.96 | **83.41** | **79.24** |
| Valve | 94.52 | 82.33 | **96.20** | **86.38** |
| Slider | **99.68** | **98.31** | 99.60 | 98.29 |
| ToyCar | 80.62 | 72.65 | **93.40** | **87.43** |
| ToyConveyor | 74.32 | 59.80 | 75.69 | 62.96 |
| Average | 83.39 | 72.70 | **88.12** | **80.94** |

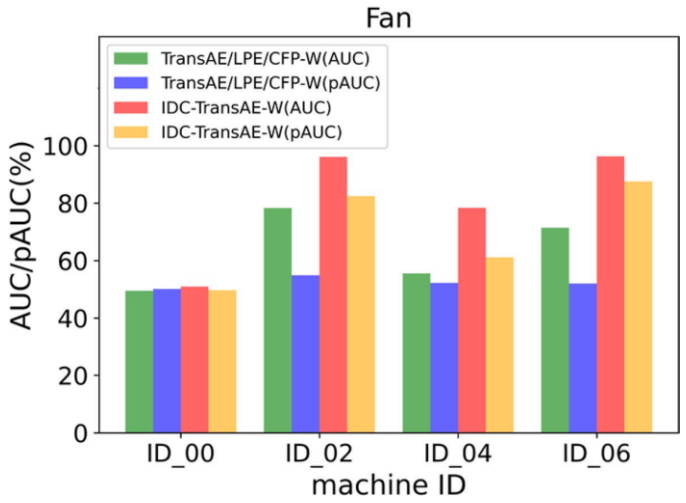
- No ID：TransAE/LPE/CFP-W
- ID：IDC-TransAE-W




(a) TransAE/LPE/CFP




(b) IDC-TransAE



# IDC-TransAE Method

| | Fan | | Pump | | Slider | | Valve | | ToyCar | | ToyConveyor | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | pAUC | AUC | pAUC | AUC | pAUC | AUC | pAUC | AUC | pAUC | AUC | pAUC | AUC | pAUC |
| w/o ID information | | | | | | | | | | | | | | |
| AE baseline [12] | 65.91 | 51.93 | 70.20 | 61.69 | 83.42 | 65.72 | 67.78 | 51.67 | 78.77 | 67.58 | 72.53 | **60.43** | 73.10 | 59.84 |
| IDNN [20] | 65.94 | 52.48 | 74.26 | 62.20 | 84.34 | 65.48 | 83.70 | 62.02 | 77.42 | 62.64 | 69.36 | 58.58 | 75.67 | 60.57 |
| ANP-Boot [21] | 64.80 | 53.00 | 65.50 | 59.00 | **94.90** | 83.10 | 85.20 | 72.00 | 72.90 | 68.10 | 67.10 | 54.20 | 75.07 | 64.90 |
| Group MADE [24] | 68.00 | 53.10 | 74.10 | 66.20 | 94.40 | **83.70** | 95.60 | 85.50 | 79.50 | 68.40 | **74.70** | 60.30 | 81.05 | 69.53 |
| **TransAE-mean** | 73.91 | 54.14 | 77.31 | 68.96 | 91.51 | 74.66 | 96.09 | 84.65 | 80.62 | 72.65 | 74.32 | 59.80 | 82.29 | 69.14 |
| **TransAE-W** | **73.91** | **54.14** | **77.31** | **68.96** | 94.52 | 82.33 | **99.68** | **98.31** | **80.62** | **72.65** | 74.32 | 59.80 | **83.39** | **72.70** |
| w/ ID information | | | | | | | | | | | | | | |
| MobileNetV2 [29] | 80.19 | **74.40** | 82.53 | 76.50 | 95.27 | 85.22 | 88.65 | 87.98 | 87.66 | 85.92 | 69.71 | 56.43 | 84.34 | 77.74 |
| Glow_Aff [37] | 74.90 | 65.30 | 83.40 | 73.80 | 94.60 | 82.80 | 91.40 | 75.00 | 92.20 | 84.10 | 71.50 | 59.00 | 85.20 | 73.90 |
| IDCAE [27] | 77.45 | 70.32 | 77.29 | 70.33 | 80.04 | 68.25 | 78.26 | 55.80 | 78.07 | 74.22 | 70.29 | 59.46 | 76.90 | 66.40 |
| **IDC-TransAE-mean** | 80.44 | 70.21 | 83.41 | 79.24 | 92.17 | 77.10 | 94.04 | 78.94 | 93.17 | 87.43 | 75.69 | 62.96 | 86.49 | 75.98 |
| **IDC-TransAE-W** | **80.44** | 70.21 | **83.41** | **79.24** | **96.20** | **86.38** | **99.60** | **98.29** | **93.40** | **87.43** | **75.69** | **62.96** | **88.12** | **80.94** |

J. Guan, Y. Liu, Q. Kong, et al., "Transformer-based autoencoder with ID constraint for unsupervised anomalous sound detection," EURASIP J. Audio Speech Music Process., vol. 2023, no. 42, 2023.

# Problem 1: Unsupervised Learning
## Label Classification Based Methods

- Examples:
  - MobileNet v2 (Giri et al, 2020)
  - Glow_Aff (Dohi et al 2021)
  - STgram-MFN (Liu et al 2022)
  - ASD-AFPA (Zhang et al, 2023)
  - CLP-SCF (Guan et al, 2023)


R. Giri, S. V. Tenneti, F. Cheng, K. Helwani, U. Isik, and A. Krishnaswamy, "Self-supervised classification for detecting anomalous sounds," in Proc. DCASE Workshop, Tokyo, Japan, Nov. 2020, pp. 46–50.

K. Dohi, T. Endo, H. Purohit, R. Tanabe, and Y. Kawaguchi, "Flow-based self-supervised density estimation for anomalous sound detection," in Proc. ICASSP, 2021, pp. 336–340.

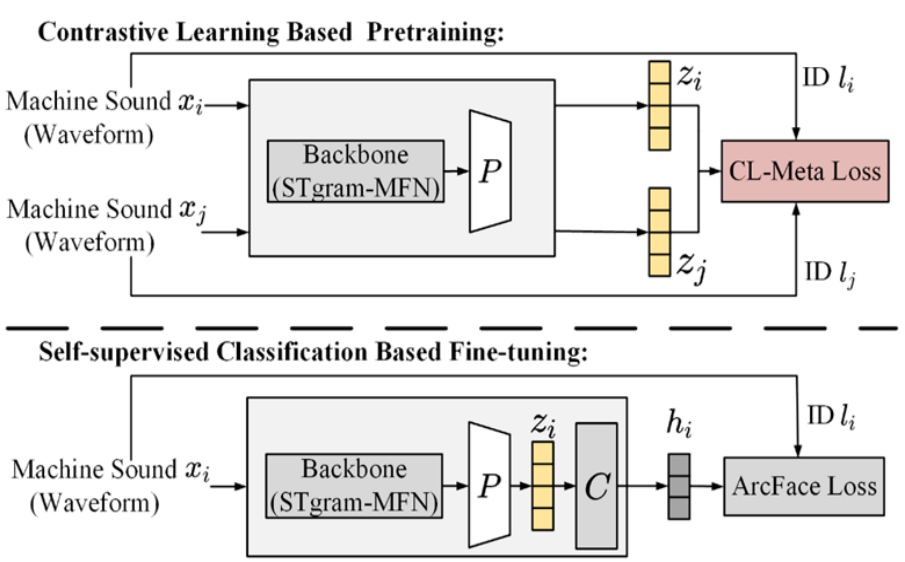
Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous sound detection using spectral-temporal information fusion," in Proc. ICASSP, 2022, pp. 816–820.

H. Zhang, J. Guan, Q. Zhu, F. Xiao, and Y. Liu, "Anomalous sound detection using self-attention-based frequency pattern analysis of machine sounds," in Proc. INTERSPEECH, 2023, pp. 336–340.

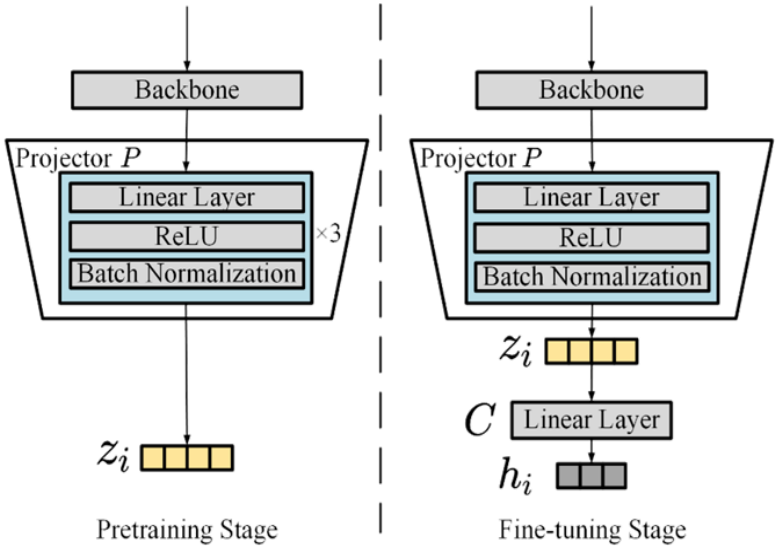
J. Guan, F. Xiao, Y. Liu, Q. Zhu, and W. Wang, "Anomalous sound detection using audio representation with machine ID based contrastive learning pretraining," in Proc. ICASSP, 2023, pp. 1–5.

# An Example Method: CLP-SCF

Using a contrastive learning–based pretraining to exploit **machine ID metadata** & **acoustic features** to improve feature consistency for the same ID while enlarging inter-ID differences, thereby improving the discriminability and stability of anomalous sound detection.




(a) Training procedure of the proposed CLP-SCF method.




(b) Different model structures in two stages.

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. ICASSP, 2023, pp. 1–5.

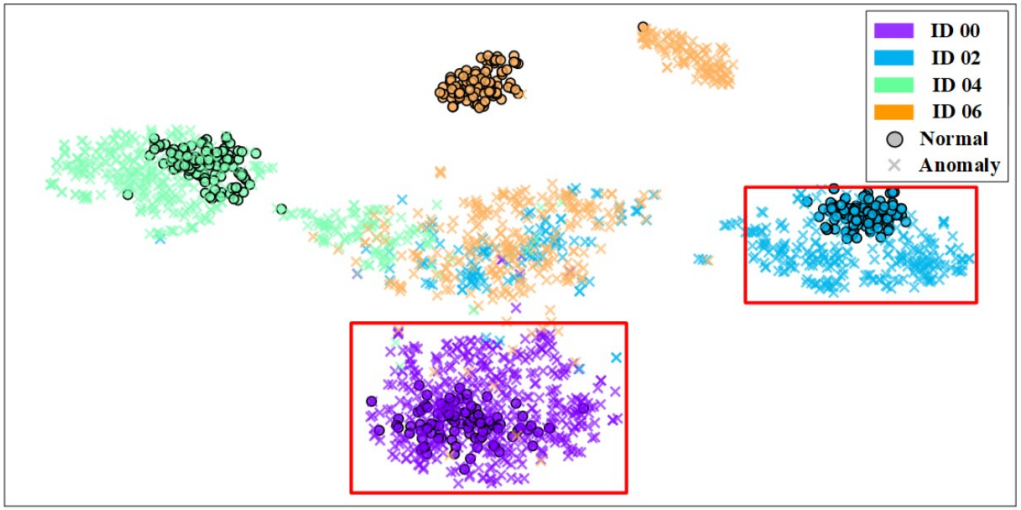
# CLP-SCF

| Methods | Fan | | Pump | | Slider | | Valve | | ToyCar | | ToyConveyor | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | pAUC | AUC | pAUC | AUC | pAUC | AUC | pAUC | AUC | pAUC | AUC | pAUC | AUC | pAUC |
| IDNN [2] | 67.71 | 52.90 | 73.76 | 61.07 | 86.45 | 67.58 | 84.09 | 64.94 | 78.69 | 69.22 | 71.07 | 59.70 | 76.96 | 62.57 |
| MobileNetV2 [3] | 80.19 | 74.40 | 82.53 | 76.50 | 95.27 | 85.22 | 88.65 | 87.98 | 87.66 | 85.92 | 69.71 | 56.43 | 84.34 | 77.74 |
| Glow_Aff [4] | 74.90 | 65.30 | 83.40 | 73.80 | 94.60 | 82.80 | 91.40 | 75.00 | 92.20 | 84.10 | 71.50 | 59.00 | 85.20 | 73.90 |
| STgram-MFN (ArcFace) [9] | 94.04 | 88.97 | 91.94 | 81.75 | 99.55 | 97.61 | 99.64 | 98.44 | 94.44 | 87.68 | 74.57 | **63.60** | 92.36 | 86.34 |
| AADCL [13] | 85.27 | 68.93 | 86.75 | 70.85 | 77.74 | 61.62 | 68.62 | 55.03 | 88.79 | 75.95 | 71.26 | 57.40 | 79.74 | 64.96 |
| **CLP-SCF** | **96.98** | **93.23** | **94.97** | **87.39** | **99.57** | **97.73** | **99.89** | **99.51** | **95.85** | **90.19** | **75.21** | 62.79 | **93.75** | **88.48** |

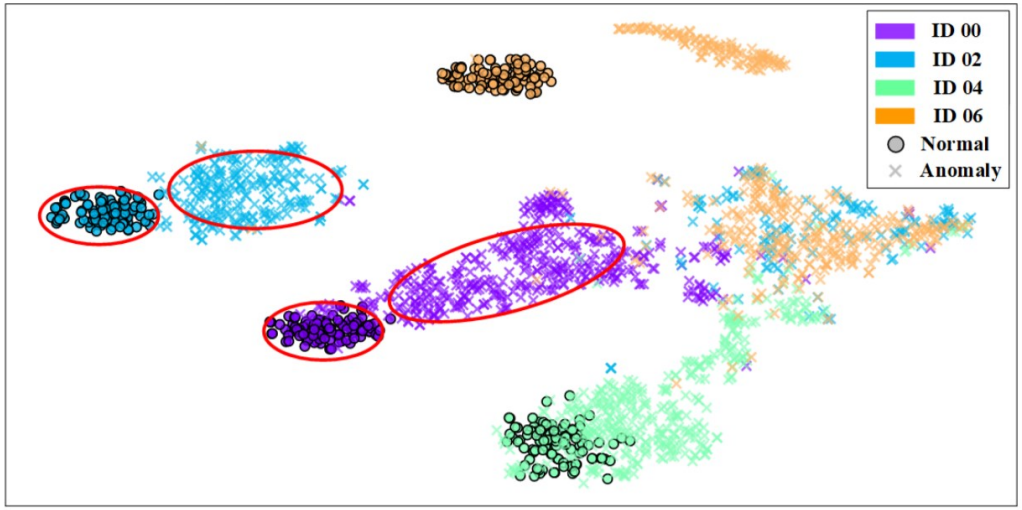
| Methods | STgram-MFN (ArcFace) [9] | **CLP-SCF** |
|---|---|---|
| Fan | 81.39 | **88.27** |
| Pump | 83.48 | **87.27** |
| Slider | 98.22 | **98.28** |
| Valve | 98.83 | **99.58** |
| ToyCar | 83.07 | **86.87** |
| ToyConveyor | 64.16 | **65.46** |
| Average | 84.86 | **87.62** |

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. ICASSP, 2023, pp. 1–5.

# CLP-SCF




(a) STgram-MFN (ArcFace) [9]




(b) Proposed CLP-SCF

Fan type t-SNE visualisation

# Problem 2: Domain Shift

**Possible solutions to mitigate domain shift:**

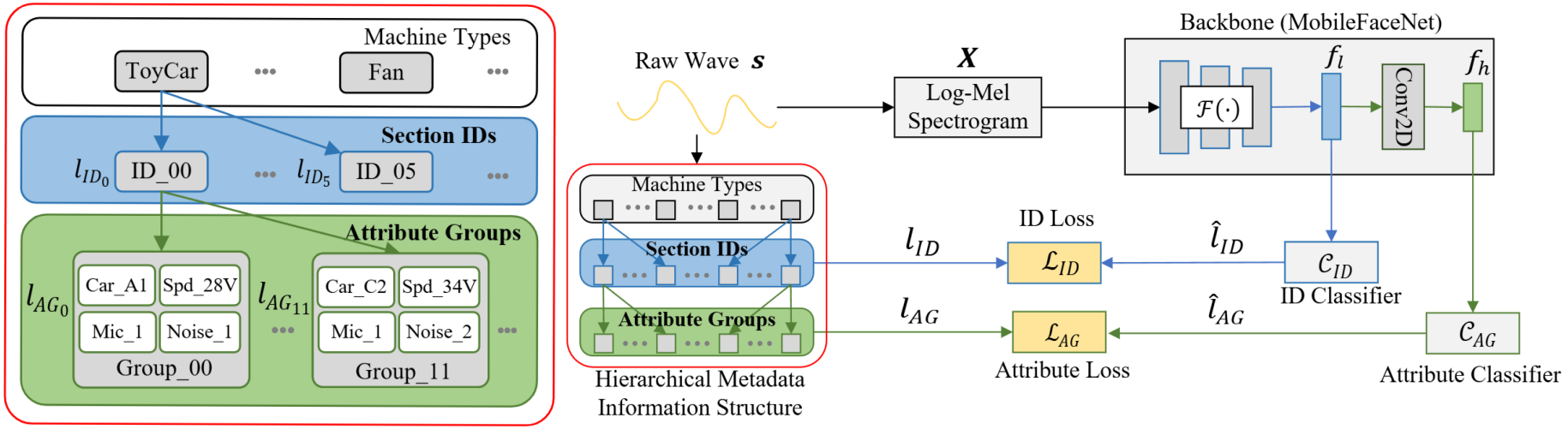
- **Domain mixing**

➢ **Idea**: Mixing data from source and target domain for training the ASD models.
➢ **Methods**: Wilkinghoff, 2022, Verbitskiy et al 2022.
➢ **Pros**: Easy to implement, useful for small distribution gaps.
➢ **Cons**: Problematic for large distribution gaps, deviations from normal data distributions, and lack theoretical foundations.

- **Domain classification**

➢ **Idea**: Leverages metadata carried by machine sound data (e.g., operating speed, factory environment, machine type, section IDs) as self-supervised labels to model distribution differences across training samples, thereby improving detection performance.
➢ **Methods**: Kuroyanagi et al 2022, Xiao et al, 2022, Venkatesh et al, 2022, Lan et al 2024, Zhang et al 2025, Bian and Chen 2025.
➢ **Pros**: Improving detection performance of self-supervised techniques.
➢ **Cons**: Requiring metadata; often relying on known anomalous samples from development data for hyperparameter tuning and model selection; may overlook the common properties between domains.

# An Example Method: HMIC-AGC

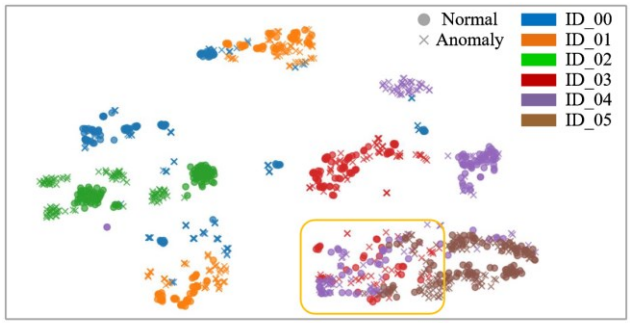



- HMI loss: $\mathcal{L}_{total} = \lambda \mathcal{L}_{ID} + (1-\lambda)\mathcal{L}_{AG}$
  - Section loss: $\mathcal{L}_{ID} = CE(l_{ID}, \hat{l}_{ID})$
  - Attributes loss: $\mathcal{L}_{AG} = CE(l_{AG}, \hat{l}_{AG})$
- Attribute group clustering
  - Attribute group centroid: $\boldsymbol{c}_m = \frac{1}{N}\sum_{n=1}^{N} \hat{\boldsymbol{f}}_{h_n}$
  - Anomaly score: $\mathcal{A} = \min_{m \in [1,M]} \sqrt{(\bar{\boldsymbol{f}} - \boldsymbol{c}_m)^T \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{f}} - \boldsymbol{c}_m)}$

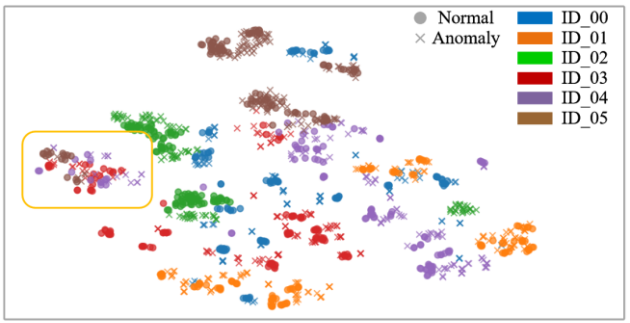
H. Lan, Q. Zhu, J. Guan, Y. Wei, and W. Wang, "Hierarchical metadata information constrained self-supervised learning for anomalous sound detection under domain shift," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2024, pp. 7670–7674.

# HMIC-AGC

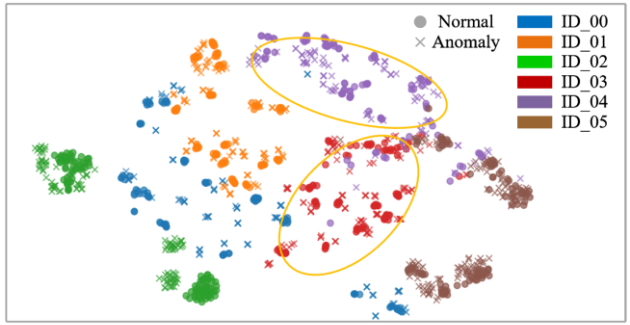
| Methods | ToyCar | | ToyTrain | | Bearing | | Fan | | Gearbox | | Slider | | Valve | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | pAUC | AUC | pAUC | AUC | pAUC | AUC | pAUC | AUC | pAUC | AUC | pAUC | AUC | pAUC | AUC | pAUC |
| AE [18] | 61.18 | 60.21 | 43.14 | 49.36 | 59.93 | 53.95 | 41.16 | 50.12 | 61.92 | 51.95 | 58.95 | 54.16 | 54.26 | 51.30 | 53.01 | 52.80 |
| MobileNetV2 [18] | 42.79 | 53.44 | 51.22 | 50.98 | 58.23 | 52.16 | 50.34 | **55.22** | 51.34 | 48.49 | 62.42 | 53.07 | 72.77 | 65.16 | 54.19 | 53.67 |
| Attribute-only [9] | 87.61 | 73.12 | 56.64 | 52.60 | **73.92** | 58.77 | 52.69 | 49.79 | 74.11 | 59.96 | 73.39 | 59.51 | 78.14 | 69.26 | 67.68 | 59.47 |
| Domain-only [8] | 77.15 | 67.47 | 55.92 | 51.53 | 71.91 | **60.74** | 54.52 | 53.86 | 78.75 | 53.30 | 78.87 | **59.56** | 85.60 | 78.59 | 69.51 | 59.56 |
| **HMIC-DC** | 82.44 | 71.92 | 57.88 | 52.75 | 67.45 | 59.14 | 56.55 | 53.03 | 77.22 | 59.74 | 80.59 | 58.75 | 89.70 | **82.69** | 70.20 | 61.15 |
| **HMIC-AGC** | **87.91** | **77.51** | **59.10** | **52.83** | 68.14 | 59.41 | **57.63** | 53.25 | **79.78** | **61.29** | **80.76** | 58.29 | **89.87** | 82.30 | **71.79** | **61.91** |




(a) Domain-only [8]




(b) Attribute-only [9]




(c) HMIC (proposed)

- ToyCar in DCASE 2022 challenge Task 2 as an example, section ID 00 contains four attributes (i.e., “car model”, “speed”, “microphone number”, and “noise number”) with different attribute values.
- Here, “car model” has the value of A1, C2, etc., and “noise number” has the value of 1, 2, etc. By grouping these attributes in terms of their values, we obtain 12 AGs for section ID 00, and a total number of 44 AGs for the ToyCar.
- Thus, the DCASE 2022 dataset of 7 machine types each with 6 section IDs, becomes 250 AGs under 42 section IDs, each with 990 and 10 audio clips from the source and target domain, respectively.

# Problem 3: Few Shot Problem

**Aim**: adapting ASD methods to previously unseen machine types (i.e. anomalous samples are entirely unavailable during training)

**Possible solutions:**

- **Data augmentation**

- **Idea**: generating more data based on existing training data, using data augmentation techniques such as Mixup, StatEx, and FeatEx.
- **Methods**: Zhang et al 2024, Wilkinghoff et al 2025, etc.
- **Pros**: Easy to implement, improve detection accuracy, noise robustness and generalisation ability.
- **Cons**: Lack of realism, risk of overfitting and misinterpretation, hyperparameter sensitivity

Y. Zhang, J. Liu, Y. Tian, H. Liu, and M. Li, “A Dual-path Framework with Frequency-and-time Excited Network for Anomalous Sound Detection,” in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, 2024, pp. 1266–1270.
K. Wilkinghoff, H. Yang, J. Ebbers, F. G. Germain, G. Wichern, andJ. L. Roux, “Local Density-based Anomaly Score Normalization for Domain Generalization,” arXiv preprint arXiv:2509.10951, 2025.

# Few Shot Problem

- **Pre-trained model**

➢ **Idea**: use pretrained models, such as Wav2Vec 2.0, HuBERT, and BEATs, to improve audio representations (e.g. richer and more discriminative feature representations).

➢ **Methods**: Jiang et al 2023, Zhang et al 2025, Fan et al 2025, etc

➢ **Pros**: Strong general audio representation.

➢ **Cons**: Reliance on anomaly-dependent tuning, sensitivity to domain shift, overfitting under limited data, and high computational cost.

A. Jiang, Q. Hou, J. Liu, P. Fan, J. Ma, C. Lu, Y. Zhai, Y. Deng, and W.-Q. Zhang, “THUEE System for First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring,” DCASE2023Challenge, Tech. Rep., June 2023.
P. Fan, A. Jiang, S. Zhang, Z. Lv, B. Han, X. Zheng, W. Liang, J. Li,W.-Q. Zhang, Y. Qian et al., “Fisher: A Foundation Model for Multi-modal Industrial Signal Comprehensive Representation,” arXiv preprintarXiv:2507.16696, 2025.
Y. Zhang, J. Liu, and M. Li, “ECHO: Frequency-aware Hierarchical En-coding for Variable-length Signals,” arXiv preprint arXiv:2508.14689,2025.

# Few Shot Problem

- **Synthetic supervision with generative models**

➢ **Idea**: turn metadata into text prompts, and then use text to audio generation models to generate both the normal and anomalous sounds of the target machine types, which are then used for hyperparameter tuning during model training.

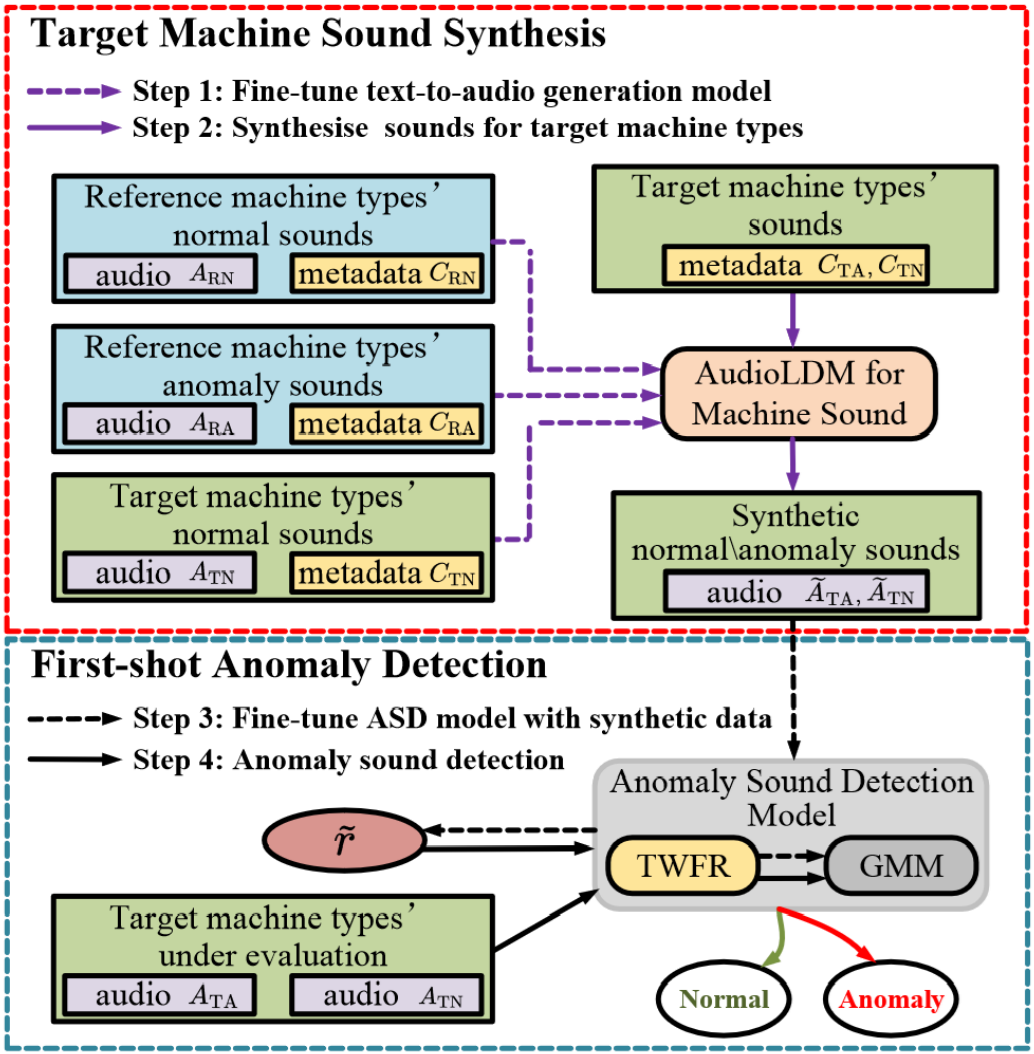
➢ **Methods**: Zhang et al 2024, Niu et al 2025, Purohit et al 2025, etc.

➢ **Pros**: Controllable and interpretable creation of anomalies in terms of the simulated machine failure conditions (e.g., squeaking, rattling, grinding, humming, whistling), via transforming normal sound into anomalous sound. Text prompts reflect semantic information related to the machine type, which is different from conventional data augmentation. Alleviate data scarcity.

➢ **Cons**: Potentially introduces bias toward artificial anomaly patterns, amplifies domain mismatch, and relies on heuristic design choices, limiting robustness and generalization in real-world anomalous sound detection.

H. Zhang, Q. Zhu, J. Guan, H. Liu, F. Xiao, J. Tian, X. Mei, X. Liu, and W. Wang, “First-shot Unsupervised Anomalous Sound Detection with Unknown Anomalies Estimated by Metadata-assisted Audio Generation,” in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, 2024, pp.1271–1275.
B. Niu, Y. Wei, G. Yang, Y. Wang, and S. Yu, “StarGAN-Aug: A Cross-domain Fault Audio Generation Method for High-performance Fault Diagnosis of Power Transformers,” in Proceedings of the Annual Conference of the International Speech Communication Association(INTERSPEECH), 2025, pp. 3399–3403.
H. Purohit, T. Nishida, K. Dohi, T. Endo, and Y. Kawaguchi, “MIMII-Agent: Leveraging LLMs with Function Calling for Relative Evaluation of Anomalous Sound Detection,” arXiv preprint arXiv:2507.20666,2025.

# An Example: **FS-TWFR-GMM**




**FS-TWFR-GMM:** By exploring the intrinsic relationship between **metadata** and **acoustic features**, this method leverages device-specific metadata to guide generative models to synthesize anomalous sounds for unseen machines, enabling model selection and overcoming the lack of anomaly samples in fully unsupervised settings

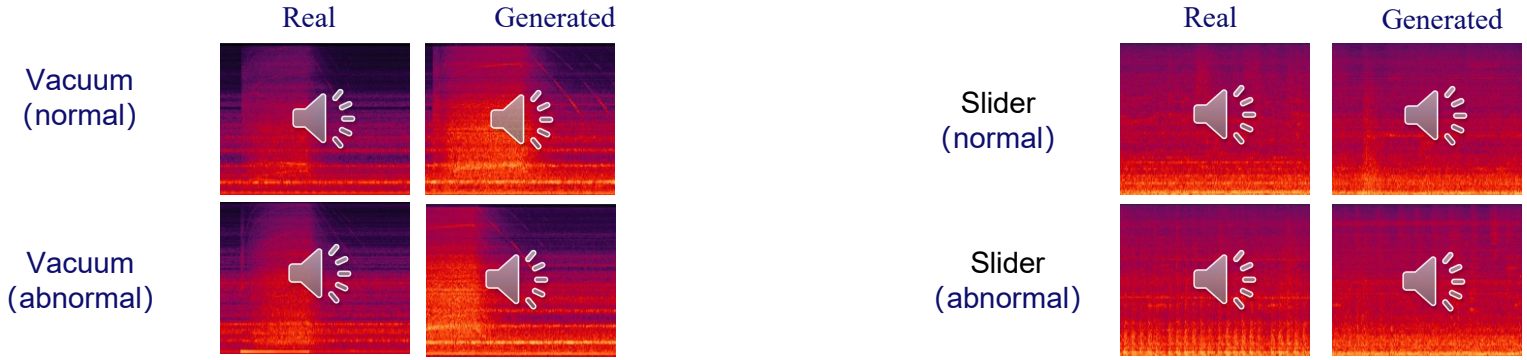
H. Zhang, Q. Zhu, J. Guan, H. Liu, F. Xiao, J. Tian, X. Mei, X. Liu, and W. Wang, "First-shot unsupervised anomalous sound detection with unknown anomalies estimated by metadata-assisted audio generation," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2024, pp. 1271–1275.

# Meta data -> Caption-> Sound

**Metadata** captures critical physical and operational characteristics of machine sounds, such as **machine type**, **status**, **operating conditions** and **environmental noise**, which directly influence acoustic properties

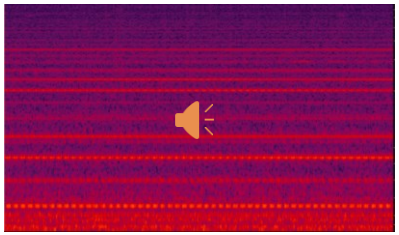
| Machine Type | Original Filename | Metadata Caption |
|---|---|---|
| **ToyCar** | ToyCar/section_00_source_test_***normal***_0001_car_***B2***_spd_***31V***_mic_***1***.wav | This is the ***normal*** sound of a ***toy car*** with model ***B2*** and speed ***31V***, recorded by a microphone placed at the position ***1***. |
| **ToyCar** | ToyCar/section_00_source_test_***anomaly***_0001_car_***B2***_spd_***31V***_mic_***1***.wav | This is the ***anomaly*** sound of a ***toy car*** with model ***B2*** and speed ***31V***, recorded by a microphone placed at the position ***1***. |
| **Grinder** | Grinder/section_00_source_train_***normal***_0000_grindstone_***2***_plate_***2***.wav | This is the ***normal*** sound of a ***grinding*** machine with grindstones ***2*** and metal plates ***2***. |

**Caption to sound generation** can be achieved using text to audio (TTA) models, such as AudioLDM (Liu et al 2023), AudioLDM2 (Liu et al 2024), and WavJourney (Liu et al 2025).

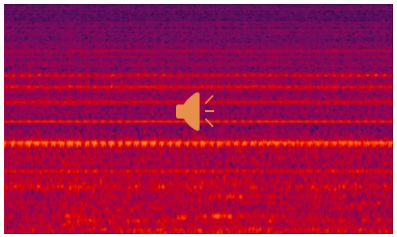



H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, M. D. Plumbley, "AudioLDM: text-to-audio generation with latent diffusion models," in Proc. IEEE International Conference on Machine Learning (ICML 2023), Hawaii, USA, 23-29 July, 2023.
H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, M. D. Plumbley, "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining," IEEE/ACM Transactions on Audio Speech and Language Processing, vol. 32, pp. 2871-2883, 2024.
X. Liu, Z. Zhu, H. Liu, Y. Yuan, Q. Huang, M. Cui J. Liang, Y. Cao, Q. Kong, M. D. Plumbley, and W. Wang, "WavJourney: Compositional Audio Creation with Large Language Models," IEEE Transactions on Audio Speech and Language Processing, vol. 33, pp. 2830 - 2844, June 2025.

# Impact of Synthetic Samples

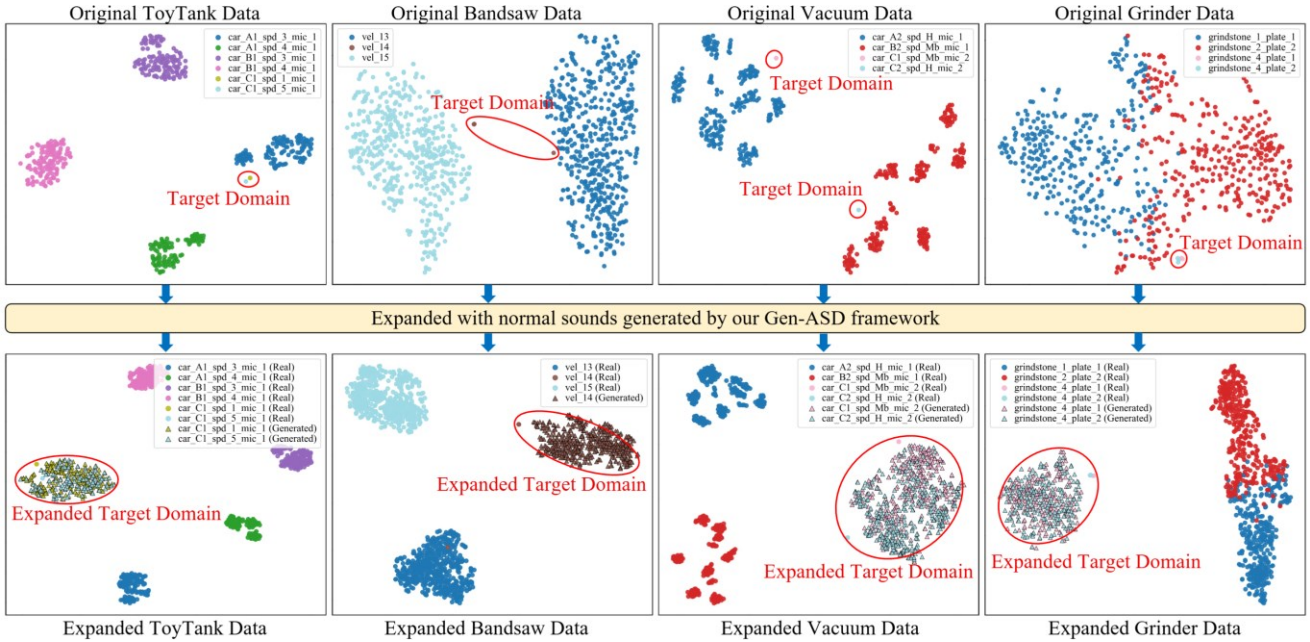


source_train_normal_0016_vel_17_loc_A.wav



source_train_normal_0017_vel_21_loc_A.wav

Real audio of Bearing with different operation status




T-SNE visualizations before and after using Gen-ASD

F. Xiao, J. Guan, K. Zhang, Q. Zhu, H. Liu, S. Zhang, and W. Wang, "First-Shot Unsupervised Anomalous Sound Detection with Metadata-Assisted Audio Generation" IEEE TASLP, 2025, submitted.

# Results

Performance comparison with DCASE 2023 Challenge Task 2 top submissions

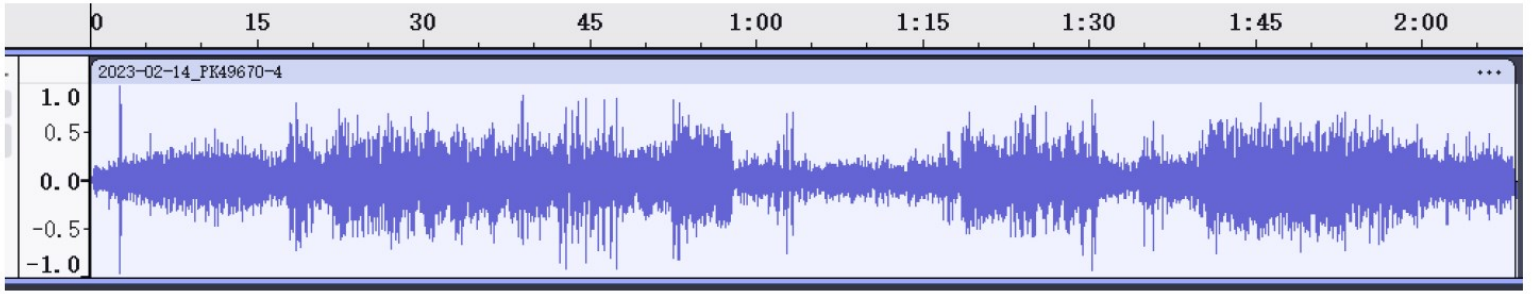
| Method | Ranking | ToyDrone | | ToyNscale | | ToyTank | | Vacuum | | Bandsaw | | Grinder | | Shaker | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUC | pAUC | AUC | pAUC | AUC | pAUC | AUC | pAUC | AUC | pAUC | AUC | pAUC | AUC | pAUC | AUC | pAUC |
| Jie_IESEFPT | 1 | 58.03 | 51.58 | 89.03 | 77.74 | 60.33 | 61.53 | 96.18 | 85.32 | 65.66 | 53.35 | 66.63 | 62.45 | 68.08 | 55.97 | 69.75 | 62.03 |
| Lv_HUAKONG | 2 | 54.84 | 49.37 | 82.71 | 57.00 | 74.80 | 63.79 | 93.66 | 87.42 | 58.48 | 50.30 | 66.69 | 61.22 | 74.24 | 65.24 | 70.05 | 60.11 |
| Jiang_THUEE | 3 | 55.83 | 49.74 | 73.44 | 61.63 | 63.03 | 59.74 | 81.98 | 76.42 | 71.10 | 56.64 | 62.18 | 62.41 | 75.99 | 64.68 | 68.03 | 60.71 |
| **FS-TWFR-GMM (Proposed)** | – | 56.28 | 50.89 | 64.33 | 54.16 | 62.60 | 57.47 | 82.75 | 75.84 | 78.31 | 61.62 | 61.75 | 54.98 | 83.39 | 71.32 | 68.41 | 59.76 |
| Wilkinghoff_FKIE | 4 | 53.90 | 50.21 | 87.14 | 76.58 | 63.43 | 62.21 | 83.26 | 74.00 | 66.06 | 52.87 | 67.10 | 62.11 | 65.91 | 50.24 | 67.95 | 59.58 |
| Bai_JLESS | 5 | 51.44 | 50.89 | 59.85 | 51.16 | 70.05 | 59.58 | 81.46 | 69.47 | 74.51 | 55.65 | 67.07 | 63.03 | 78.30 | 63.37 | 68.95 | 59.02 |
| Zhou_SHNU | 6 | 61.10 | 55.74 | 62.23 | 52.11 | 68.66 | 59.53 | 77.05 | 63.53 | 69.13 | 51.99 | 69.04 | 61.51 | 68.83 | 55.94 | 68.01 | 57.19 |
| Guan_HEU | 7 | 62.93 | 52.05 | 68.94 | 54.21 | 66.41 | 60.63 | 79.47 | 72.47 | 57.22 | 50.76 | 62.38 | 54.96 | 78.46 | 61.47 | 67.12 | 57.32 |
| Du_NERCSLIP | 8 | 58.06 | 51.47 | 86.84 | 65.37 | 61.29 | 57.58 | 82.05 | 60.84 | 47.58 | 49.92 | 49.06 | 49.21 | 93.24 | 80.78 | 68.30 | 59.31 |
| DCASE2023_Baseline | 9 | 58.93 | 51.42 | 50.73 | 50.89 | 57.89 | 53.84 | 86.84 | 65.32 | 69.10 | 57.54 | 60.19 | 59.55 | 72.28 | 62.33 | 63.41 | 56.82 |

- **Seven reference machine types** (Fan, Gearbox, Bearing, Slider, ToyCar, ToyTrain, Valve) and **seven target machine types** (Vacuum, ToyTank, ToyNscale, ToyDrone, Bandsaw, Grinder, Shaker).
- In the training set, for each reference machine type, there are 1100 normal sound clips and 100 abnormal sound clips, while for each target machine type, there are 1000 normal sound clips.
- In the evaluation set, there are 200 sound clips with unknown conditions (normal or abnormal) for each target machine type.

# Results

| Method | DCASE Ranking | Parameters for Training (one-off cost) | Parameters for Detection | Average AUC |
|---|---|---|---|---|
| Jie_IESEFPT [48] | 1 | 3M | 3M | 69.75 |
| Lv_HUAKONG [49] | 2 | 300M | 300M | 70.05 |
| Jiang_THUEE [50] | 3 | 6M | 6M | 68.03 |
| Wilkinghoff_FKIE [51] | 4 | 34M | 34M | 67.95 |
| Guan_HEU [39] | 7 | 33K (TWFR-GMM) + 792M (AudioLDM) | **33K** | 67.12 |
| AnoPatch [27] | – | 91M | 91M | 71.96 |
| SSL4ASD [21] | – | 34M | 34M | 72.60 |
| **Gen-TWFR-GMM (Ours)** | – | 33K (TWFR-GMM) + 1.1B (AudioLCM) | **33K** | 70.80 |
| **Gen-AnoPatch (Ours)** | – | 91M (AnoPatch) + 1.1B (AudioLCM) | 91M | **73.88** |

# Problem 4: Noisy Labels





Disagreement between annotators due to various factors:

- Low SNR & strong background interference;
- faint short, and intermittent events;
- nonstationary normal and abnormal sound;
- missing context or knowledge about sound events

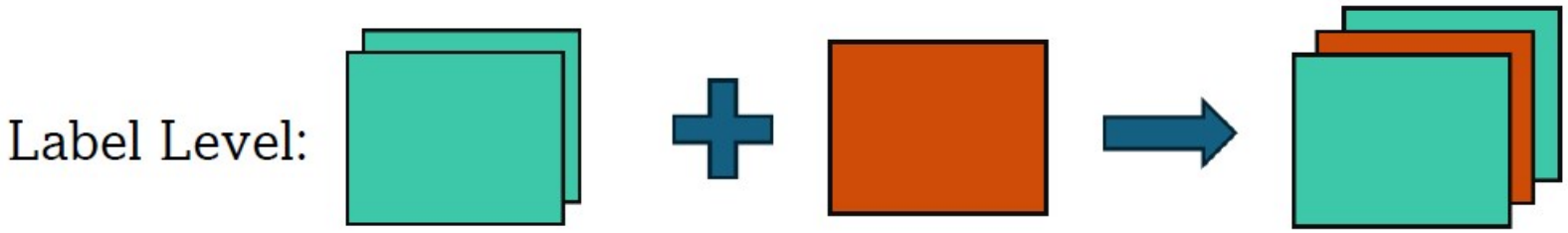
**This can lead to noisy or incorrect labels.**

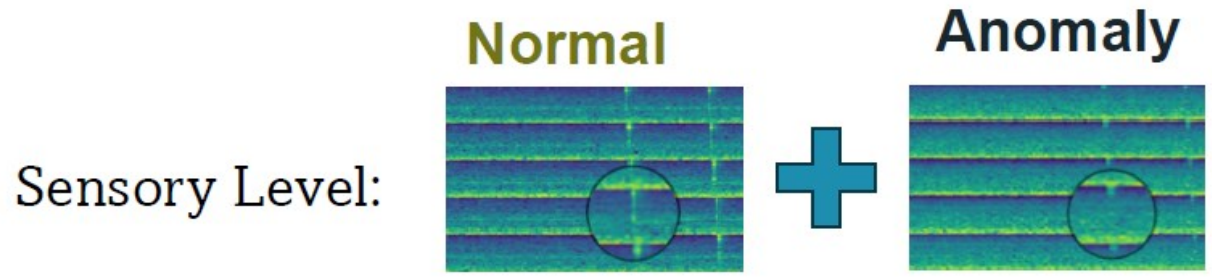
# Problem 4: Noisy Labels

A common assumption in unsupervised AAD: **all training data are normal.**

**Question:**

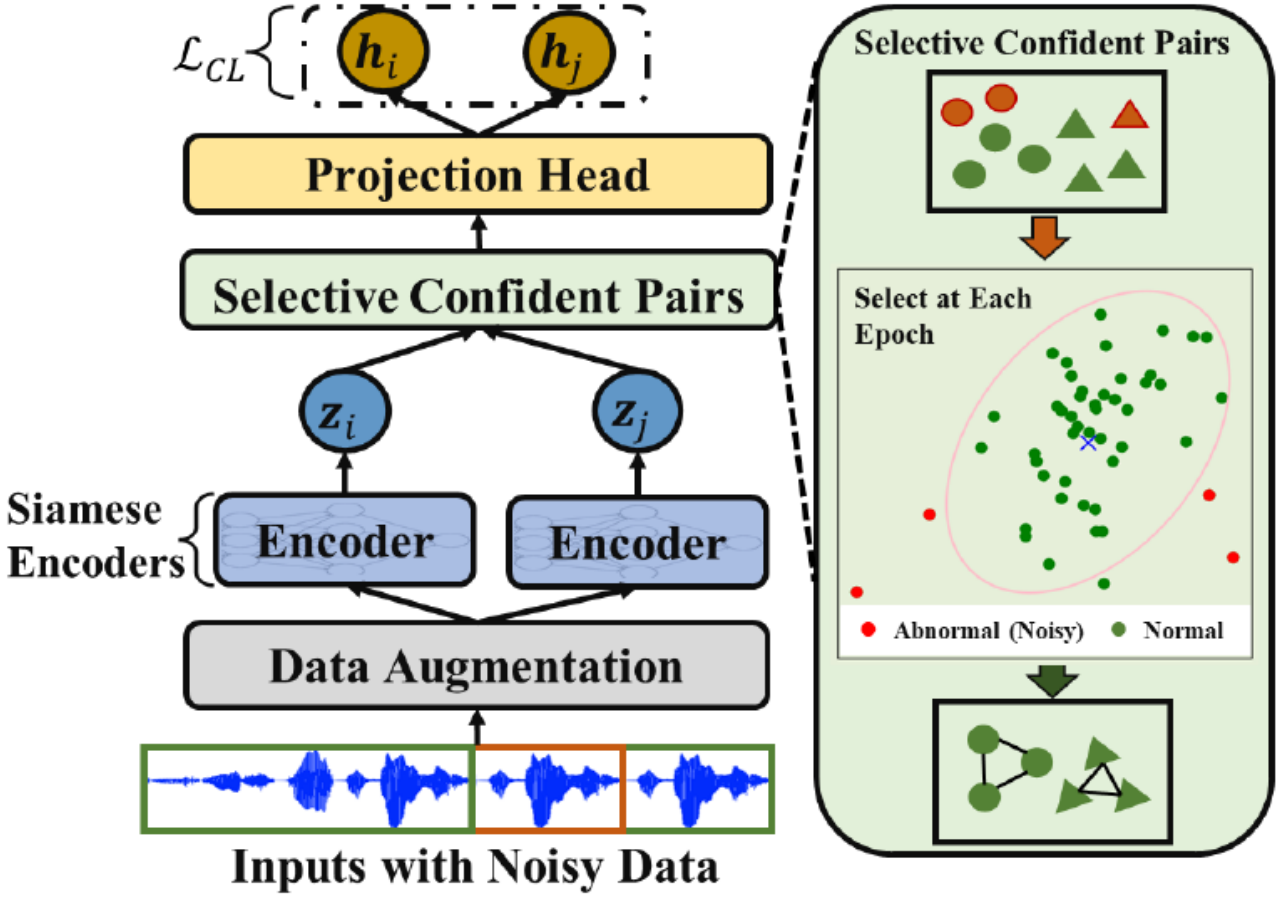
How can an unsupervised AAD model reliably learn from a 'normal' dataset that is contaminated with unlabelled anomalies?



# SCAN

## Selective Contrastive Learning Against Noisy Data






Z. Liu, Y. Hou, W. Wang, S. Michiels, and D. Hughes, “SCAN: Selective Contrastive Learning Against Noisy Data for Acoustic Anomaly Detection”, IEEE Signal Processing Letter, pp. 1–5, 2025.

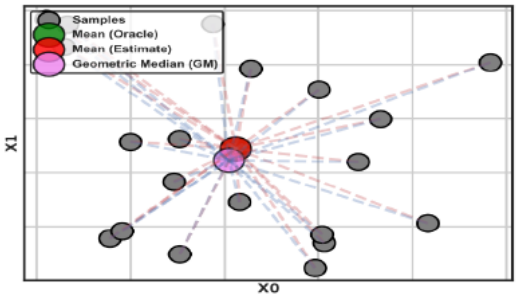
# SCAN

**Selecting Confident Pairs:**

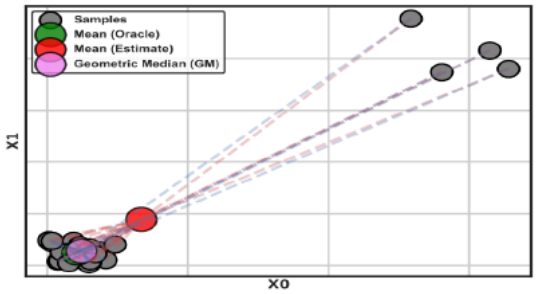
- Robust Scoring: Uses Mahalanobis Distance to compute anomaly scores.
- Robust Center: Employs the Geometric Median (replaces mean)

***Why?***

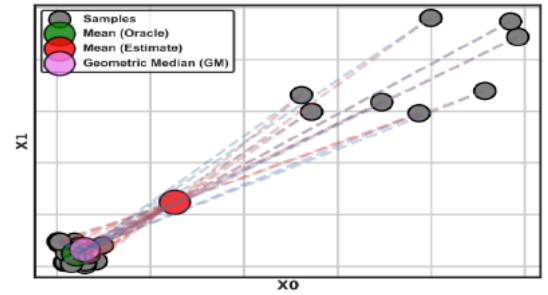
- The “Mean” is sensitive to outliers (noise) and gets "pulled" away.
- The Geometric Median is determined by the bulk of the data, making it robust to outliers and a more stable center.




(a) **No Corruption** ($\psi = 0$)




(b) **20% Corruption** ($\psi = 0.2$)




(c) **40% Corruption** ($\psi = 0.4$)

# SCAN

***Challenge:*** Anomaly scores are unreliable in early training epochs.

- **Handling Early-Stage Unreliability**
  - A Progressive Balancing Factor (logarithmic function) gradually adjusts a Chi-squared-based threshold.
  - Starts conservatively and becomes stricter as representations improve.
  - Final Loss: The contrastive loss InfoNCE is computed only on the selected confident pairs (positive).

Chi-square bound on Mahalanobis distance for “normal” samples.

$$\Theta^{(e)} = \chi_p^2(1-\alpha)$$

Epoch-dependent scaling that gradually increases this bound: Epochs ↑, the threshold ↑→more confident samples pass the filter.

$$\Theta_p^{(e)} = \Theta^{(e)}\left(1 - \frac{1}{\log(e+\epsilon)}\right)$$

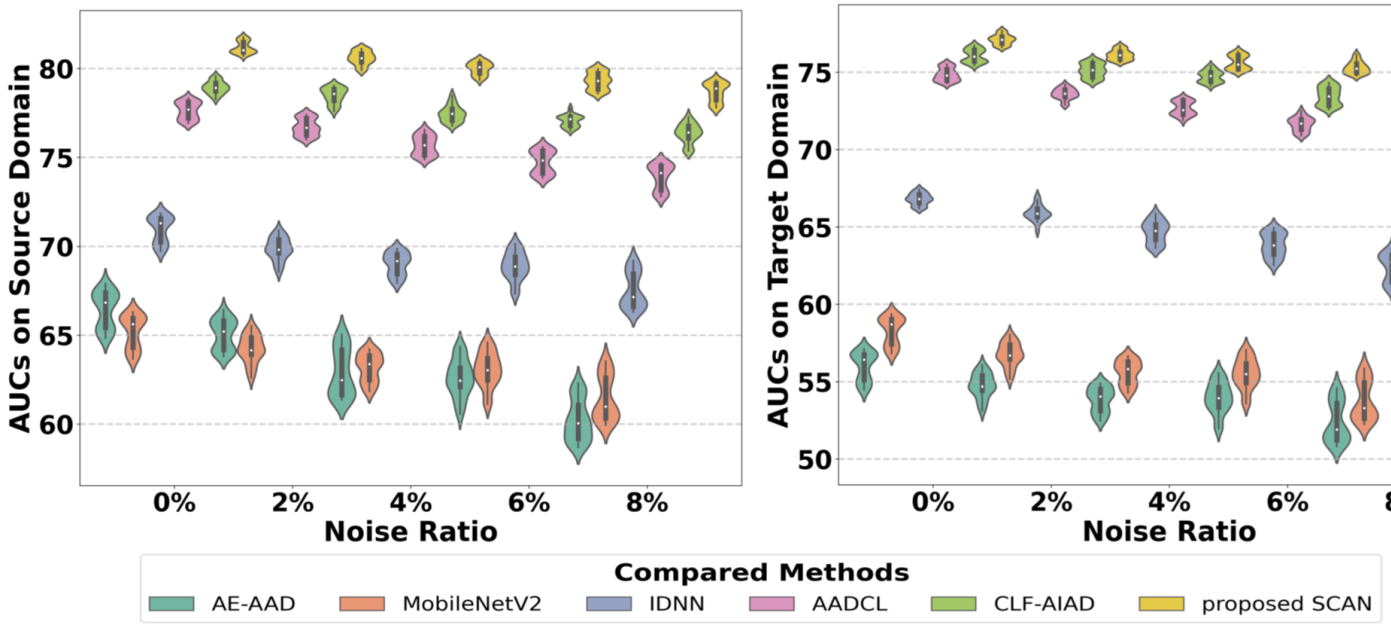
# SCAN - Results

**Noisy Data Settings:**

- Sample **r%** anomalous clips from the MIMII DG test anomaly set (DCASE2022 dataset).
- Inject them into the normal training set as “normal” → create **noise-r%** settings.
- Keep test labels unchanged → injected clips remain anomalies at inference (**label conflict**).

**Feasibility and Effectiveness:**

SCAN vs. Baselines & SOTA: Reliability under Increasing Noise (0–8%)



# Conclusions & Future Directions

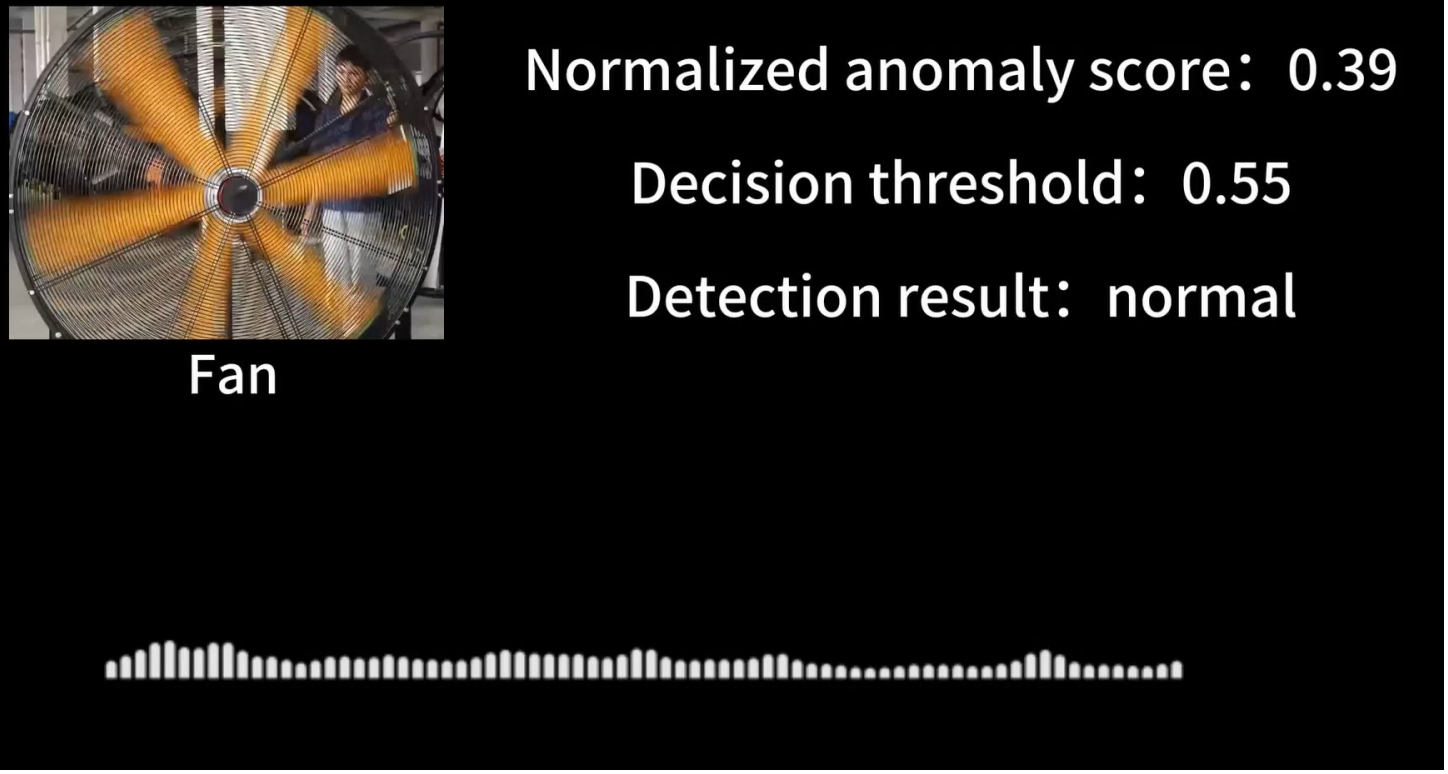
**Conclusions:**

- We have discussed four typical open problems & potential solutions in anomalous sound detection with examples and results: unsupervised learning, domain shift, few shot, and noisy labels.

**Future directions:**

- Continuous developments for new datasets, in particular, anomalous sounds from real industrial environments.
- Further improving latent representations for both normal and anomalous sounds with better discriminative capability and robustness.
- Specialised ASD with domain knowledge
- Further exploitation of metadata for improved ASD

# Demos


Fan
Normalized anomaly score：0.39
Decision threshold：0.55
Detection result：normal

# Take Away

**SCAN**:
Paper/Code/Demo:
at https://github.com/sherrylzy/SCAN

**IDC-TransAE**:
Code/Demo:
at https://github.com/liuyoude/TransAE

**My contact:**
Email: w.wang@surrey.ac.uk
Web: https://personalpages.surrey.ac.uk/w.wang/