# FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION WITH UNKNOWN ANOMALIES ESTIMATED BY METADATA-ASSISTED AUDIO GENERATION

*Hejing Zhang*[1,2], *Qiaoxi Zhu*[3], *Jian Guan*[1,2*], *Haohe Liu*[4], *Feiyang Xiao*[1,2], *Jiantong Tian*[1,2], *Xinhao Mei*[4], *Xubo Liu*[4], *Wenwu Wang*[4]

[1] Group of Intelligent Signal Processing (GISP), College of Computer Science and Technology, Harbin Engineering University, Harbin, China

[2] National Engineering Laboratory for Modeling and Emulation in E-Government, Harbin Engineering University, Harbin, China

[3] Centre for Audio, Acoustics and Vibration (CAAV), University of Technology Sydney, Ultimo, Australia

[4] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK

## ABSTRACT

First-shot (FS) unsupervised anomalous sound detection (ASD) is a brand-new task introduced in DCASE 2023 Challenge Task 2, where the anomalous sounds for the target machine types are unseen in training. Existing methods often rely on the availability of normal and abnormal sound data from the target machines. However, due to the lack of anomalous sound data for the target machine types, it becomes challenging when adapting the existing ASD methods to the first-shot task. In this paper, we propose a new framework for the first-shot unsupervised ASD, where metadata-assisted audio generation is used to estimate unknown anomalies, by utilising the available machine information (i.e., metadata and sound data) to fine-tune a text-to-audio generation model for generating the anomalous sounds that contain unique acoustic characteristics accounting for each different machine type. We then use the method of Time-Weighted Frequency domain audio Representation with Gaussian Mixture Model (TWFR-GMM) as the backbone to achieve the first-shot unsupervised ASD. Our proposed FS-TWFR-GMM method achieves competitive performance amongst top systems in DCASE 2023 Challenge Task 2, while requiring only 1% model parameters for detection, as validated in our experiments.

***Index Terms*—** Unsupervised learning, anomalous sound detection, audio generation, metadata, latent diffusion model

## 1. INTRODUCTION

Anomalous sound detection (ASD) aims to distinguish between the normal and anomalous operating states of a machine based on the sounds emitted from it [1–5]. However, due to the infrequent occurrence and potential diversity of anomalous sound, it is challenging and time-consuming to gather sufficient training data for anomalous sound covering various situations. To mitigate this issue, unsupervised ASD, utilising only normal sounds during training, becomes a desirable, albeit challenging, option.

First-shot (FS) unsupervised ASD has been introduced for the Detection and Classification of Acoustic Scenes and Events (DCASE) 2023 Challenge Task 2 [2, 6], aiming to detect target machine types' anomalous sounds that are unseen in training. There are three sets of information used for training: (1) anomalous and normal sounds from the *reference* machine types, (2) normal sounds from the *target* machine types, (3) metadata, including machine type and attributes on operational and environmental conditions, as the label of each of the above sounds. These target machine types (e.g., Vacuum, ToyTank, ToyNscale, ToyDrone, Bandsaw, Grinder, and Shaker) are entirely distinct from the reference machine types (e.g., Fan, Gearbox, Bearing, Slider, ToyCar, ToyTrain, and Valve). State-of-the-art ASD methods often rely on the availability of normal and abnormal data from the target machines. However, in practice, anomalous sound data for the target machine types may be difficult to capture due to their rare occurrence in practice. This makes it difficult to adapt these existing ASD methods to the first-shot task, as discussed by the DCASE 2023 Challenge Task 2 organisers [2, 7].

To address this challenge, we present a new framework for the first-shot unsupervised ASD with unknown anomalies estimated by metadata-assisted audio generation. Specifically, we use a text-to-audio (TTA) generation model for synthesizing anomalous and normal sounds for the target machine type. We use the state-of-the-art TTA model, i.e. AudioLDM [8], but fine-tuned using all the available data in the first-shot scenario, including the anomalous and normal sounds from the reference machine types, normal sounds from the target machine types, and their corresponding metadata describing the operational and environmental conditions of these machines. The proposed approach is built on the ASD model in our previous study [5], which is a Time-Weighted Frequency domain

*Corresponding author.

This work was partly supported by the GHfund under Grant No. 202302026860.

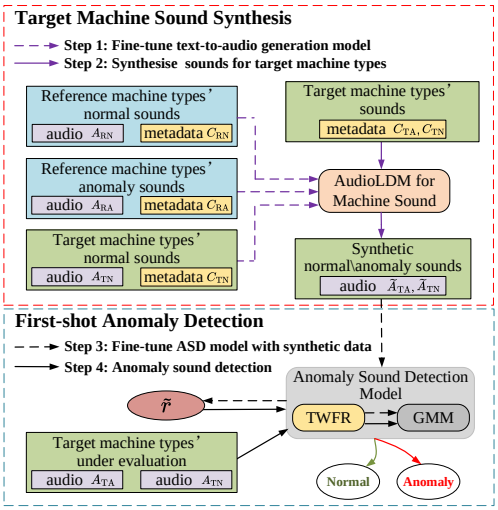


**Fig. 1**. The proposed first-shot unsupervised ASD method using TWFR-GMM as the backbone. For target machine sound synthesis (steps 1-2), we use normal and anomalous sounds from reference machine type(s) and normal sounds from the target machine type(s), along with corresponding metadata detailing machine operating and monitoring conditions. For first-shot anomaly detection (step 3-4), we identify the unseen anomalous sounds from the target machine type(s).

audio Representation (TWFR) with Gaussian Mixture Model (GMM). For this reason, we abbreviate the proposed first-shot approach as FS-TWFR-GMM. In this model, we use a hyperparameter $r$ to highlight the important information of the audio representation in the time domain that may differ between machine types. It is determined for the target machine type by fine-tuning using synthesised normal and anomalous sounds rather than using real normal and anomalous sounds as in [5].

Our method is the first to estimate unknown anomalies in unsupervised ASD using machine information for audio generation. Existing anomaly synthesis methods (e.g., [9–12]) only use normal sound data, but not abnormal sounds of the reference machines or metadata. In contrast, our method exploits all available audio data and metadata, giving improved quality in the synthesised machine sounds, which captures the features of actual sounds from the corresponding machine type. Furthermore, our approach is versatile in accommodating a wide range of machine metadata, regardless of type and format, thus alleviating challenges posed by diverse machine attributes and label formats frequently encountered in real-world scenarios.

Experiments show that the proposed method provides competitive performance for the first-shot unsupervised ASD. The preliminary version [13] forms the core of the system ranked the 7th in DCASE 2023 Challenge Task 2. Our improved version in this paper is now between 3rd and 4th place, with only $1.34\%$ in the AUC metric and $2.27\%$ in the pAUC metric lower than the top method. Notably, our approach requires vastly reduced resources due to a non-deep-learning design, i.e., only $1.1\%$ of the number of parameters as used in the top method in DCASE 2023 Challenge. Thus, the proposed method is promising for practical applications with computing resource constraints.

## 2. PROPOSED METHOD

Fig. 1 outlines the proposed FS-TWFR-GMM method. First, we synthesize the sounds of target machine types with a fine-tuned text-to-audio generation model, utilising normal and anomalous sounds and their metadata from the reference machine types and normal sound data from the target machine type, as detailed in Section 2.1. Then, the ASD model tuned for each target machine type with the synthetic machine sounds is used to detect the unseen anomalous sounds, as detailed in Section 2.2. Note that the backbone ASD model TWFR-GMM can be replaced with other ASD models, such as [3, 12, 14].

### 2.1. Metadata-Assisted Machine Sound Synthesis

#### *2.1.1. Metadata based Machine Sound Captioning*

In the training dataset for ASD, the label of machine sounds contains related metadata, e.g., machine operating status. In our method, we generate captions based on the metadata, for example, using captions to describe the machine's operating status. We then use captions as textual prompt, and generate synthetic sound using a TTA model. First, we convert the metadata to captions as

$$c = \boldsymbol{F}_c(l) \tag{1}$$

where $\boldsymbol{F}_c(\cdot)$ denotes the captioning function. It converts the label $l$ of a machine sound to caption $c$, with a predefined descriptive text template for each different machine type, with examples illustrated in Table 1.

#### *2.1.2. Fine-tuning AudioLDM for Machine Sound Synthesis*

We use the AudioLDM algorithm to synthesise machine sounds related to a specific target machine type in terms of the captions generated in Section 2.1.1. The AudioLDM algorithm uses the contrastive language-audio pretraining (CLAP) [15] to build a shared latent space between text embeddings of the captions and audio embeddings of the sounds, and uses the latent diffusion model (LDM) [8] on a continuous audio representation for text-to-audio generation, conditioned on the caption. To tailor this model for our task, we fine-tune a pre-trained AudioLDM [8], using the machine sound and caption pairs, as follows,

$$\mathcal{G} \leftarrow P(\boldsymbol{A}|\boldsymbol{C}) \tag{2}$$

where $P$ denotes the pre-trained AudioLDM model, and $\mathcal{G}$ is the AudioLDM model fine-tuned by a set of machine audios $\boldsymbol{A}$, with the corresponding set of captions $\boldsymbol{C}$ as the condition.

**Table 1**. Examples of available metadata within the descriptive text captions for machine sounds from audio labels, including the normal and anomalous sound of the reference machine type, i.e., ToyCar, and the normal sound of the target machine type, i.e., Grinder.

| Machine type | Example of the label (metadata) | Caption for text-to-audio generation |
|---|---|---|
| **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** | 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** | 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***. |

$$\begin{cases} \boldsymbol{A} = \{\boldsymbol{A}_{\text{RN}}, \boldsymbol{A}_{\text{RA}}, \boldsymbol{A}_{\text{TN}}\} \\ \boldsymbol{C} = \{\boldsymbol{C}_{\text{RN}}, \boldsymbol{C}_{\text{RA}}, \boldsymbol{C}_{\text{TN}}\} \end{cases} \tag{3}$$

where $\boldsymbol{A}_{\text{RN}}$, $\boldsymbol{A}_{\text{RA}}$ and $\boldsymbol{A}_{\text{TN}}$ respectively represent the sets of audio signals for the reference machine type's normal sounds, the reference machine type's anomalous sounds, and the target machine type's normal sounds, with corresponding sets of captions $\boldsymbol{C}_{\text{RN}}$, $\boldsymbol{C}_{\text{RA}}$, and $\boldsymbol{C}_{\text{TN}}$.

To synthesise sounds for the target machine types, we use the corresponding captions as the condition and gradually denoise from the fine-tuned LDM distribution to estimate the true data distribution and generate audio, that

$$\begin{cases} \widetilde{\boldsymbol{A}}_{\text{TN}} = \mathcal{G}(\boldsymbol{C}_{\text{TN}}) \\ \widetilde{\boldsymbol{A}}_{\text{TA}} = \mathcal{G}(\boldsymbol{C}_{\text{TA}}) \end{cases} \tag{4}$$

where $\widetilde{\boldsymbol{A}}_{\text{TN}}$ and $\widetilde{\boldsymbol{A}}_{\text{TA}}$ denote the sets of synthetic audios for target machine types' normal sounds and anomalous sounds, respectively.

In first-shot unsupervised ASD, the audios and captions for target machine types' anomalous sounds do not exist in the training stage. Therefore, we obtain the captions set $\boldsymbol{C}_{\text{TA}}$ for target machine types' anomalous sound generation by replacing the word "normal" in $\boldsymbol{C}_{\text{TN}}$ with "anomaly".

### 2.2. First-Shot Unsupervised ASD Using Synthesised Sounds

With the synthetic normal and anomaly sounds accounting for characteristics of specific target machine types, we can train the ASD model for the first-shot scenario by optimising audio feature representations to distinguish between normal and abnormal sounds effectively. In this paper, we adapt TWFR-GMM for the first-shot scenario, resulting in FS-TWFR-GMM.

The TWFR-GMM algorithm [5] obtains the time-weighted frequency domain audio representation (TWFR) $R(\mathbf{X}) \in \mathbb{R}^M$, by incorporating a hyperparameter $r$ for each machine type, as follows

$$R(\mathbf{X}) = Ranking(\mathbf{X}) \cdot \left[\frac{r^0}{z(r)}, \frac{r^1}{z(r)}, ..., \frac{r^{N-1}}{z(r)}\right]^\top \tag{5}$$

where $\mathbf{X} \in \mathbb{R}^{M \times N}$ is the log-mel spectrogram of an audio signal with $M$ mel-bins and $N$ time frames. $Ranking(\cdot)$ denotes the operation of re-arranging $\mathbf{X}$ in descending order for the energy values over time frames for time weight calculation, following [5]. Here, $r$ determines the weight assigned to each time frame, and $z(r) = \sum_{n=1}^{N} r^{n-1}$ is for weight normalisation. $\top$ denotes the transposition operation.

The audio representation $R(\cdot)$ is trained using normal sounds, but fine-tuned with anomalous sounds for each machine type's dynamic and unique sound characteristics, which is then used for audio feature extraction in the detection stage with GMM to achieve anomaly detection in [5]. However, it is not applicable in the first-shot scenarios, as there are no abnormal sounds existing for the target machine types.

In this paper, to adapt TWFR-GMM for the first-shot scenario, the hyperparameter $r$ is estimated by optimising the following cost with the synthesised machine sounds,

$$\widetilde{r} = \underset{r}{\text{argmax}} \left\{ E(r, \widetilde{\boldsymbol{A}}_{\text{TA}}, \widetilde{\boldsymbol{A}}_{\text{TN}}) \right\} \tag{6}$$

where $\widetilde{r}$ denotes the estimated value of $r$, and $E(\cdot)$ is evaluation metric for ASD following [5]. We set the selection range $r \in [0, 1.10]$, and the selection interval is $0.01$. When $r$ exceeds one, it gives greater weights to the time frames with lower energy, whereas when $r$ is less than one, it gives higher weights to time frames with higher energy. This approach considers the diverse audio patterns observed among different types of machines. For instance, some machines produce loud anomalous sounds, while others exhibit short-term stalls caused by extraneous object interference, as discussed in [16].

The synthesised normal and anomalous sounds ($\widetilde{\boldsymbol{A}}_{\text{TN}}$ and $\widetilde{\boldsymbol{A}}_{\text{TA}}$) are employed to optimise the hyperparameter $r$ in TWFR. In contrast, the GMM uses real normal sounds of the target machine type ($a_{\text{TN}}$). This approach helps minimise the potential bias introduced by the sounds generated by AudioLDM. Other detailed implementations of TWFR can be found in [5].

## 3. EXPERIMENTS

### 3.1. Experimental Setup

**Dataset:** We use the DCASE 2023 Challenge Task 2 dataset [2], including 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.

**Table 2**. 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 [17] | 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 [18] | 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 [19] | 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 [9] | 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 |
| Guan_HEU [13] | 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 |
| DCASE2023_Baseline [6] | 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 |

**Table 3**. Comparison of the number of parameters.

| Method | Ranking | Training (one-off cost) | Detection |
|---|---|---|---|
| Jie_IESEFPT [17] | 1 | 3M | 3M |
| Lv_HUAKONG [18] | 2 | 300M | 300M |
| Jiang_THUEE [19] | 3 | 6M | 6M |
| **FS-TWFR-GMM (Proposed)** | – | 33K+792M (AudioLDM) | **33K** |
| Wilkinghoff_FKIE [9] | 4 | 34M | 34M |
| Guan_HEU [13] | 7 | 33K+792M (AudioLDM) | **33K** |
| DCASE2023_Baseline [6] | 9 | 267K | 267K |

**Table 4**. Comparison of single systems, namely FS-TWFR-GMM and its initial version (System 2 of the ensemble system [13]). With or without extended $r$ range refers to selecting $r$ from $[0, 1.1]$ or $[0, 1]$. With or without RS refers to removing or keeping the silence part in the synthetic machine sounds.

| Method | Extended $r$ range | RS | AUC | pAUC |
|---|---|---|---|---|
| System 2 of Guan_HEU [13] | ✗ | ✗ | 65.07 | 57.69 |
| **FS-TWFR-GMM** | ✓ | ✗ | 65.22 | 57.72 |
| | ✗ | ✓ | 67.90 | 59.19 |
| | ✓ | ✓ | **68.41** | **59.76** |

**Table 5**. Comparison of using hyperparameter $r$ from no training data, generated, or real machine sound data.

| Methods | Training data | AUC | pAUC | Average |
|---|---|---|---|---|
| $r = 0$ | None | 56.44 | 54.04 | 55.24 |
| $r = 1$ | None | 66.78 | **60.28** | 63.53 |
| **FS-TWFR-GMM** | Synthetic | **68.41** | 59.76 | **64.08** |
| TWFR-GMM | Real | 71.55 | 61.62 | 66.59 |

**Table 6**. Selected $r$ from synthetic or real machine sounds.

| Training data | ToyDrone | ToyNscale | ToyTank | Vacuum | Bandsaw | Grinder | Shaker |
|---|---|---|---|---|---|---|---|
| Synthetic | 1.02 | 1.00 | 0.99 | 0.84 | 1.03 | 0.96 | 1.01 |
| Real | 1.01 | 1.00 | 0.87 | 0.94 | 1.02 | 0.99 | 1.02 |
| *Difference* | 0.01 | 0.00 | 0.12 | 0.10 | 0.01 | 0.03 | 0.01 |

**Evaluation metrics:** The area under the receiver operating characteristic curve (AUC) and the partial-AUC (pAUC) are commonly used for performance evaluation [1, 4, 20, 21], where pAUC represents the AUC over a low false-positive-rate range $[0, 0.1]$ [1]. A larger value indicates better anomalous sound detection performance.

### 3.2. Results

Tables 2 and 3 show that the proposed FS-TWFR-GMM has a significant advantage in the number of parameters (33k) required for the detection stage and achieves competitive performance ranking between the 3rd and 4th places amongst top systems in the DCASE 2023 Challenge Task 2 on first-shot unsupervised ASD, with only $1.34\%$ in AUC and $2.27\%$ in pAUC lower than the 1st placed method.

The initial version of FS-TWFR-GMM, i.e., System 2 of the ensemble system [13] achieved the 7th place in DCASE 2023 Challenge Task 2. In comparison, the FS-TWFR-GMM version proposed in this paper optimises $r$ over an extended range $[0, 1.1]$, and is shown to be more effective, as shown in Table 4.

Table 5 shows that the proposed method fine-tuned from synthetic data is generally better than blindly setting the hyperparameter ($r = 0$ for max pooling, or $r = 1$ for average pooling), which represents the straightforward approach due to the unavailability of anomaly data in first-shot ASD. Furthermore, the performance of the proposed unsupervised method is only $3.14\%$ in AUC and $1.94\%$ in pAUC lower than the performance achieved by the fully supervised approach employing real abnormal and normal data directly from the evaluation set to optimise $r$ in TWFR-GMM.

Table 6 shows $r$ selected in terms of the synthetic sounds or real anomalous sounds in the evaluation set with nearly no difference. Moreover, it is adapted to the unique sound patterns of each machine type, through prioritising lower or higher energy time frames or treating all time frames equally ($r > 1$, $r < 1$, or $r = 1$).

## 4. CONCLUSION

We have presented a new framework for the first-shot unsupervised anomalous sound detection using a text-to-audio generation model to synthesise normal and abnormal machine sounds while leveraging all available training data. With our approach, unseen anomalies in new machine types can be estimated. As a result, it becomes easier to distinguish between normal and unknown anomaly sounds. The first-shot unsupervised method FS-TWFR-GMM implements the proposed framework on the time-weighted frequency domain audio representation with the Gaussian mixture model. It performs similarly to the state-of-the-art first-shot unsupervised ASD methods. Furthermore, the proposed framework can be used with other ASD systems for the first-shot scenarios.

## 5. REFERENCES


[1] Y. Koizumi *et al*., "Description and discussion on DCASE2020 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE Workshop*, 2020, pp. 81–85.

[2] K. Dohi *et al*., "Description and discussion on DCASE 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE Workshop*, 2023, pp. 31–35.

[3] Y. Zeng, H. Liu, L. Xu, Y. Zhou, and L. Gan, "Robust anomaly sound detection framework for machine condition monitoring," DCASE2022 Challenge, Tech. Rep., 2022.

[4] Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous sound detection using spectral-temporal information fusion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2022, pp. 816–820.

[5] J. Guan, Y. Liu, Q. Zhu, T. Zheng, J. Han, and W. Wang, "Time-weighted frequency domain audio representation with GMM estimator for anomalous sound detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023, pp. 1–5.

[6] N. Harada, D. Niizumi, Y. Ohishi, D. Takeuchi, and M. Yasuda, "First-shot anomaly sound detection for machine condition monitoring: A domain generalization baseline," *arXiv preprint arXiv:2303.00455*, 2023.

[7] K. Dohi *et al*., "Description and discussion on DCASE 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," in *Proc. DCASE Workshop*, 2022, pp. 1–5.

[8] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023.

[9] K. Wilkinghoff, "Fraunhofer FKIE submission for task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," DCASE2023 Challenge, Tech. Rep., 2023.

[10] Z. Esmaeil, S. Mohamad, M. Fatemeh, Sadat, and Y. Mohsen, "Regularized contrastive masked autoencoder model for machinery anomaly detection using diffusion-based data augmentation," *Algorithms*, vol. 16, p. 431, 2023.

[11] T. Inoue, P. Vinayavekhin, S. Morikuni, S. Wang, T. Hoang Trong, D. Wood, M. Tatsubori, and R. Tachibana, "Detection of anomalous sounds for machine condition monitoring using classification confidence," in *Proc. DCASE Workshop*, 2020, pp. 66–70.

[12] H. Chen, Y. Song, Z. Zhuo, Y. Zhou, Y.-H. Li, H. Xue, and I. McLoughlin, "An effective anomalous sound detection method based on representation learning with simulated anomalies," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023, pp. 1–5.

[13] J. Tian, H. Zhang, Q. Zhu, F. Xiao, H. Liu, X. Mei, Y. Liu, W. Wang, and J. Guan, "First-shot anomalous sound detection with GMM clustering and finetuned attribute classification using audio pretrained model," DCASE2023 Challenge, Tech. Rep., 2023.

[14] K. T. Mai, T. Davies, L. D. Griffin, and E. Benetos, "Explaining the decision of anomalous sound detectors," in *Proc. DCASE Workshop*, 2022.

[15] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP learning audio concepts from natural language supervision," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023, pp. 1–5.

[16] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection," in *Proc. IEEE Workshop Appl. Signal Process. Audio and Acoust. (WASPAA)*, 2019, pp. 313–317.

[17] J. Jie, J. Wang, S. Chen, Y. Sun, and M. Liu, "Anomaly sound detection system based on multi-dimensional attention module," DCASE2023 Challenge, Tech. Rep., 2023.

[18] Z. Lv, B. Han, Z. Chen, Y. Qian, J. Ding, and J. Liu, "Unsupervised anomalous detection based on unsupervised pretrained models," DCASE2023 Challenge, Tech. Rep., 2023.

[19] 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," DCASE2023 Challenge, Tech. Rep., 2023.

[20] K. Dohi, T. Endo, H. Purohit, R. Tanabe, and Y. Kawaguchi, "Flow-based self-supervised density estimation for anomalous sound detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2021, pp. 336–340.

[21] 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.