ENHANCED UNSUPERVISED ANOMALOUS SOUND DETECTION USING CONDITIONAL AUTOENCODER FOR MACHINE CONDITION MONITORING

Technical Report

Ronghuan Zhao[1]*, Kelong Ren*[1]*, Liang Zou*[1]*,*

[1] China University of Mining and Technology, Xuzhou, China
{ronghuanzhao, klren, liangzou}@cumt.edu.cn

ABSTRACT

This report outlines our approach to first-shot unsupervised anomalous sound detection for machine condition monitoring, developed for the *DCASE 2024 Challenge Task 2*. Given the constraint of only having normal operational data, our method focuses on leveraging generative models for anomaly detection by employing an Autoencoder (AE).

Key components of our approach include training an AE model on normal sound data to use reconstruction loss for detecting anomalies, transforming sounds into log-mel spectrograms for better feature representation, incorporating attribute or domain labels in a conditional AE to enhance context-specific anomaly detection, normalizing reconstruction losses by domain to address machine variations, and inferring domain categories using classification or clustering when labels are absent. To further improve detection performance, we employ guided diffusion model for data augmentation, enhancing the diversity and robustness of the training data. We also implement custom filtering techniques tailored to sound signals, improving the quality and relevance of the input data. By integrating these advanced techniques, our approach significantly enhances the accuracy and reliability of anomaly detection, providing a robust tool for machine condition monitoring.

Our approach achieved notable performance on the development set, demonstrating its effectiveness. The AUC for the target domain was 61.50% and for the source domain was 60.25%. Additionally, the Partial AUC values (p = 0.1) for the target and source domain was 53.26%. These results underscore the robustness and applicability of our methodology in detecting anomalous sounds in various operational contexts.

Index Terms— first-shot, anomalous sound detection, machine condition monitoring, conditional Autoencoder, guided diffusion model, custom filter

1. INTRODUCTION

In the DCASE 2024 challenge Task 2 – *First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring*, the objective is to detect anomalous sounds in machines. For this task, we need to utilize normal sounds from the training data to detect anomalies in the test data [1]. Moreover, changes in the operational states of machines or environmental noise can cause domain shifts. The system must employ domain generalization techniques to handle frequent or subtle domain shifts. In the DCASE 2024 task, attribute information for some machines is hidden. While additional attribute information can help improve detection performance, we cannot always obtain such information. Therefore, the system must perform well whether attribute information is available or not [2, 3].

Our proposed solution addresses these challenges by employing guided diffusion model for data augmentation, followed by custom filtering for certain machines. We then use a conditional Autoencoder, utilizing attribute or domain information as conditions to detect anomalies, and the Mahalanobis distance normalized according to different domains or attributes is used to calculate anomaly scores. For machines without attribute or domain labels, we set corresponding pseudo labels for them through clustering or classification strategies.

2. METHODOLOGY

2.1. Diffusion Model

Diffusion models are generative models that create new data by progressively adding noise to the original data and then training a model to reverse this process [4]. This approach involves two main phases: the forward diffusion process and the reverse diffusion process. In the forward diffusion process, the model gradually converts the original data (such as images or sounds) into pure noise by incrementally adding small amounts of noise until the data is completely obscured. This can be viewed as a multi-step Markov chain, where each step introduces slight noise perturbations. During the reverse diffusion process, the model learns to reverse this transformation, reconstructing the original data from the noise. A neural network is trained to estimate the reverse transformations at each step, progressively reducing the noise to recover the original data [5].

The **Guided Diffusion Model** enhances diffusion models with guidance mechanisms to steer the generation process [6, 7]. To generate additional data using a guided diffusion model, we model the reverse diffusion process conditioned on some guidance G:

$$\hat{\mathbf{x}}_0 = \mathbf{x}_T + \sum_{t=1}^{T} f_\theta(\mathbf{x}_t, t, G) \cdot \Delta t \tag{1}$$

where:

- $\hat{\mathbf{x}}_0$ is the reconstructed data.
- $\mathbf{x}_T$ is the noisy data at the final diffusion step.
- f_θ is the neural network parameterized by θ, which estimates the reverse process.
- G is the guidance used to control the generation process.

The guided diffusion model involves using guidance to achieve specific goals, such as generating images conditioned on certain in-

puts or following a particular distribution. This approach allows for more precise and controllable outputs, making guided diffusion model powerful tools for tasks like image synthesis, inpainting, and conditional generation.

In our approach, guided diffusion model are specifically employed for data augmentation in sound anomaly detection, providing a convenient way to generate data for specific domains or attributes. This technique increases the diversity and robustness of the training data, thereby enhancing the model's ability to detect anomalies under various operating conditions and environments.

2.2. Custom Filtering

In this task, the provided machine sound clips contain substantial noise, making direct training with these raw clips often ineffective. To address this issue, we employed Short-Time Fourier Transform (STFT) for signal processing [8]. STFT transforms time-domain signals into time-frequency domain representations, allowing us to analyze the spectral characteristics of machine sounds in different frequency bands. Through this spectral analysis, we observed significant differences in the energy distribution between machine sounds and noise across various frequency bands. Specifically, we first applied STFT to the sound signals, decomposing them into different frequency bands.

$$\mathbf{X}(t,f) = \sum_{n=0}^{N-1} x[n] \cdot w[n-t] \cdot e^{-j2\pi f n/N} \tag{2}$$

where:

- $\mathbf{X}(t,f)$ is the STFT of the signal.
- $x[n]$ is the time-domain signal.
- $w[n-t]$ is the window function.

We then performed filtering on each band, focusing on identifying those bands with higher noise energy [9]. For high-pass filtering, we apply a filter $H(f)$:

$$\mathbf{X}_{\text{filtered}}(t,f) = \mathbf{X}(t,f) \cdot H(f) \tag{3}$$

where:

- $H(f)$ is a high-pass filter function.

Through detailed band analysis and filtering experiments, we established an effective noise reduction strategy. Notably, we discovered that for certain machines, applying a high-pass filter at 1500 Hz effectively preserves the intrinsic machine sounds while significantly reducing noise interference. This STFT-based band analysis and high-pass filtering approach significantly improved the signal-to-noise ratio of our training data, thereby enhancing the performance of the anomaly sound detection model. This method not only effectively reduces the impact of noise on model training but also retains crucial machine sound features, ultimately improving the accuracy and robustness of the model in detecting anomalous sounds.

2.3. Conditional Autoencoder

Autoencoder (AE) detects anomalous sounds based on reconstruction loss [10]. Specifically, the encoder component maps the input feature vector to a low-dimensional latent representation, and the decoder component attempts to reconstruct the original input signal from this latent representation. The reconstruction loss is defined as the difference between the original input feature vector and the output vector produced by the AE [11]. For samples not present in the training set (i.e., anomaly samples), the reconstruction loss of the AE will increase significantly, allowing them to be identified as abnormal.

In terms of data processing, we convert the filtered STFT into log-Mel spectrogram for better feature representation [12]. To convert the filtered STFT to a log-Mel spectrogram, we apply a Mel filter bank $M(f)$ and take the logarithm:

$$\mathbf{S}_{\text{Mel}}(m,t) = \log\left(\sum_{f=0}^{F-1} |\mathbf{X}_{\text{filtered}}(t,f)|^2 \cdot M(m,f)\right) \tag{4}$$

where:

- $\mathbf{S}_{\text{Mel}}(m,t)$ is the log-Mel spectrogram.
- $M(m,f)$ is the Mel filter for the m-th Mel frequency bin.

In the training phase, we use machine attributes or domain information as conditions, encode this information and input it into the AE together with the audio features to enhance anomaly detection in a specific context [13]. In the testing phase, we reconstruct the samples using conditional AE corresponding to the specific machine, and calculate the Mahalanobis distance [14] for the source and target domains separately, taking the minimum value as the anomaly score. The specific process is shown in the following formula:

$$\hat{\mathbf{S}} = f_{\text{AE}}(\mathbf{S}_{\text{Mel}}, \mathbf{c}) \tag{5}$$

where:

- $\mathbf{S}_{\text{Mel}}$ is the log-Mel spectrogram.
- $\mathbf{c}$ is the condition vector.
- f_{AE} is the AE's reconstruction function.

The reconstruction loss L_{recon} is calculated as:

$$L_{\text{recon}} = \|\mathbf{S}_{\text{Mel}} - \hat{\mathbf{S}}\|^2 \tag{6}$$

During training, we store the score distributions for both the source and target domains. During testing, we compute the Mahalanobis distance D_{Maha} using these distributions:

$$\begin{aligned} D_{\text{Maha}} = \min\{&(\|\mathbf{S}_{\text{Mel}} - \hat{\mathbf{S}}\|^2)^T \boldsymbol{\Sigma}_{\text{source}}^{-1} (\|\mathbf{S}_{\text{Mel}} - \hat{\mathbf{S}}\|^2), \\ &(\|\mathbf{S}_{\text{Mel}} - \hat{\mathbf{S}}\|^2)^T \boldsymbol{\Sigma}_{\text{target}}^{-1} (\|\mathbf{S}_{\text{Mel}} - \hat{\mathbf{S}}\|^2)\} \end{aligned} \tag{7}$$

where:

- $\boldsymbol{\Sigma}_{\text{source}}$ and $\boldsymbol{\Sigma}_{\text{target}}$ are the covariance matrices of the score distributions for the source and target domains, respectively.

Additionally, we perform score normalization across source and target domains to address machine variations, we apply:

$$L_{\text{anomaly}} = \frac{D_{\text{Maha}} - \mu_d}{\sigma_d} \tag{8}$$

where:

- μ_d are the mean of the Mahalanobis distance for domain d (source or target).
- σ_d are the standard deviation of the Mahalanobis distance for domain d (source or target).

Table 1: DCASE 2024 Task 2 experimental results on development dataset (%). The value in the row "Total Score" represents the harmonic mean of the AUC and pAUC scores over all the machine types and domains.

		Baseline (MSE)	Baseline (MAHALA)	Our System
ToyCar	AUC(source)	66.98	63.01	53.76
	AUC(target)	33.75	37.35	55.98
	pAUC	48.77	51.04	48.32
ToyTrain	AUC(source)	76.63	61.99	58.28
	AUC(target)	46.92	39.99	54.66
	pAUC	47.95	48.21	48.58
bearing	AUC(source)	62.01	54.43	50.82
	AUC(target)	61.4	51.58	55.99
	pAUC	57.58	58.82	58.8
fan	AUC(source)	67.71	79.37	61.68
	AUC(target)	55.24	42.7	68.34
	pAUC	57.53	53.44	56.8
gearbox	AUC(source)	70.4	81.32	78.94
	AUC(target)	69.34	74.35	75.84
	pAUC	55.65	55.74	58.21
slider	AUC(source)	66.51	75.35	78.42
	AUC(target)	56.01	68.11	75.64
	pAUC	51.77	49.05	54.42
valve	AUC(source)	51.07	55.69	52.3
	AUC(target)	46.25	53.61	53.26
	pAUC	52.42	51.26	50.05
All	AUC(source)	65.00	65.77	60.25
	AUC(target)	50.28	49.51	61.50
	pAUC	52.84	52.28	53.26

Our network architecture is a convolutional AE, utilizing 128-dimensional log-Mel spectrogram features as input. The training batch size is set to 256, and the model is trained using the Adam optimizer with a learning rate of 0.001. When labels are missing, we use attribute classification or clustering strategies to generate pseudo labels for conditional AE.

3. RESULT

Table 1 presents the results of our system. Compared to the baseline [15], our improved Conditional AE, along with the score normalization scheme for Mahalanobis distance-based anomaly scores across source and target domains, showed slightly lower performance in the source domain but significantly better performance in the target domain. Overall, our Conditional AE-based anomaly sound detection model demonstrated notable improvements over the baseline, enhancing the detection performance. We submitted four systems for Task 2 of the DCASE 2024 Challenge, all of which have the same processing pipeline except:

1. The conditional Autoencoder that uses conditional inputs for improved anomaly detection.
2. The Autoencoder without conditional inputs.
3. A system with 256-sized log-Mel features for higher resolution analysis.
4. A Fully Connected Network mimicking the baseline structure with fully connected layers instead of convolutional layers.

4. CONCLUSION

In this technical report, we present our submission for Task 2 of the DCASE 2024 Challenge. We propose a Conditional AE-based anomaly detection system designed to accurately identify anomalous sounds. To enhance the robustness of our model, we utilize a guided diffusion model for data augmentation, thereby increasing the diversity and robustness of the training data. This improvement enables our model to detect anomalies effectively across various operational conditions and environments. For feature extraction, we address the issue of significant noise in the machine sound clips, which can obscure the machine's intrinsic sounds. We first apply Short-Time Fourier Transform (STFT) to the sound signals to decompose them into different frequency bands. We then filter these STFT results to reduce noise and preserve critical machine sound features. Finally, we convert the filtered STFT into log-Mel spectrograms for detailed analysis. This approach significantly enhances the accuracy of our anomaly detection model by providing a clearer and more informative representation of the machine sounds. For the final anomaly scores, we normalize them according to their domain or attribute labels to obtain better detection performance.

Furthermore, when attribute and domain information are available, we integrate these directly into the network with the AE to achieve comprehensive anomaly monitoring. In cases where attribute or domain information is missing, we train classifiers or use clustering techniques to infer this information and incorporate it into the model. Our results on the development dataset demonstrate that our approach yields a significant improvement in anomaly sound detection accuracy compared to the baseline.

5. REFERENCES

[1] T. Nishida, N. Harada, D. Niizumi, D. Albertini, R. Sannino, S. Pradolini, F. Augusti, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *In arXiv e-prints: 2406.07250*, 2024.

[2] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Barcelona, Spain, November 2021, pp. 1–5.

[3] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

[4] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 850–10 869, 2023.

[5] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[6] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[7] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*. PMLR, 2021, pp. 8162–8171.

[8] H. Tao, P. Wang, Y. Chen, V. Stojanovic, and H. Yang, "An unsupervised fault diagnosis method for rolling bearing using stft and generative neural networks," *Journal of the Franklin Institute*, vol. 357, no. 11, pp. 7286–7307, 2020.

[9] H. Schroter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, "Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7407–7411.

[10] P. Li, Y. Pei, and J. Li, "A comprehensive survey on design and application of autoencoder in deep learning," *Applied Soft Computing*, vol. 138, p. 110176, 2023.

[11] H. Zeng, N. Xia, M. Tao, D. Pan, H. Zheng, C. Wang, F. Xu, W. Zakaria, and G. Dai, "Dcae: A dual conditional autoencoder framework for the reconstruction from eeg into image," *Biomedical Signal Processing and Control*, vol. 81, p. 104440, 2023.

[12] A. Meghanani, C. Anoop, and A. Ramakrishnan, "An exploration of log-mel spectrogram and mfcc features for alzheimer's dementia recognition from spontaneous speech," in *2021 IEEE spoken language technology workshop (SLT)*. IEEE, 2021, pp. 670–677.

[13] L. Yang and Z. Zhang, "A conditional convolutional autoencoder-based method for monitoring wind turbine blade breakages," *IEEE transactions on industrial informatics*, vol. 17, no. 9, pp. 6390–6398, 2020.

[14] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The mahalanobis distance," *Chemometrics and intelligent laboratory systems*, vol. 50, no. 1, pp. 1–18, 2000.

[15] N. Harada, D. Niizumi, Y. Ohishi, D. Takeuchi, and M. Yasuda, "First-shot anomaly sound detection for machine condition monitoring: A domain generalization baseline," in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 191–195.