GENERAL ANOMALOUS SOUND DETECTION USING SOUND EVENT CLASSIFICATION AND DETECTION

Technical Report

Ying Zeng, Hongqing Liu, and Yi Zhou

School of Communication and Information Engineering,
Chongqing University of Posts and Telecommunications, Chongqing, China
hongqingliu@cqupt.edu.cn

ABSTRACT

This technical report describes our team's submission to DCASE 2023 Task 2. In this report, we utilize sound event classification and detection as an auxiliary task for anomalous sound detection (ASD), and this method only needs to train a general ASD model to detect anomalies, and detects multiple anomalies, and can detect them at the same time. The experimental results show that our ASD model outperforms the official model.

Index Terms— Anomalous sound detection, sound event classification, sound event detection

1. INTRODUCTION

With the developments of society and the progress of technology, machine equipment plays an increasingly important role in the industrial production. However, during the process of machine equipment operations, various factors often cause failures, which can affect production efficiency and equipment performance, even causing serious safety accidents. Anomalous Sound Detection (ASD), as an important part of machine state monitoring tool, involves real-time monitoring and analysis of the sound during the operations of the equipments to detect possible failures of the equipment.

Neural network-based methods have been widely used in ASD problems. These methods train an autoencoder (AE) [1]. AE only needs normal sounds for training. It uses the encoder to compress the data to preserve the most important features, and uses the decoder to reconstruct raw data. Another common approach is to train a classifier as an auxiliary task, based on a realistic basic assumption. Since there are often multiple machines of the same machine type in a factory, there are often some differences in the sounds from different machine IDs. We can distinguish different IDs of the same machine type by training a classifier. In the testing phase, using the negative logarithm of the probability as the anomaly score, anomalous sounds tend to output a smaller probability, thus revealing a higher anomaly score. In the previous challenge [2, 3], training a classifier as an auxiliary task greatly improved the detection performance of each class of machine type.

In the DCASE 2023 Task2 challenge [4, 5], each machine type contains only one section, so the previous method cannot be used directly. This paper studies the classification and detection of sound events to solve the above limitations. By training a sound event classification or detection model as an auxiliary task, we can also use the log likelihood as an anomaly score to detect anomalies. In addition, we also use the embedding extracted by the model to calculate the Mahalanobis distance to measure the abnormality, and the experimental results show that it is better than the official AE method.

Table 1: Model architecture, where N is the number of classes, t indicates the expansion factor, c is the output channels, n denotes the number of Inverted residuals blocks, and s is the stride. The first layer of each sequence has a stride s and others use stride 1.

Operator	t	c	n	s
Conv2d 3x3	-	32	1	2
Blockneck	1	16	1	1
Blockneck	6	24	2	2
Blockneck	6	32	3	2
Blockneck	6	64	4	2
Blockneck	6	96	3	2
Blockneck	6	160	3	1
Blockneck	6	320	1	1
Conv2d 1x1	-	1280	1	-
Linear	-	N	-	-

2. METHOD

Our proposed method is categorized into two stages:

1. Train a sound event classification or detection model as an auxiliary task.
2. Calculating the anomaly score using the embeddings extracted by the model.

In actual scenarios, it is necessary to reduce the complexity of the algorithm and improve the operation speed, so we use MobileNetV2 [6] and shallow ViT [7] as our baseline model.

2.1. MobileNetV2

2.1.1. MODEL PRETRAIN

The model structure is provided in Table 1. Our approach first trains a standard multi-label classification model [8] on the Audioset dataset, which contains 527 labels. Given a training audio clip signal(i.e., of length 10 seconds) and a multi-hot label $\boldsymbol{y}$, the model will output a predicted vector $\hat{\boldsymbol{y}}$. The aim of the training is to optimize the binary cross-entropy (BCE) loss function, which is defined as:

$$\mathcal{L}_{BCE} = -\boldsymbol{y}\log(\hat{\boldsymbol{y}}) + (\mathbf{1}-\boldsymbol{y})\log(\mathbf{1}-\hat{\boldsymbol{y}}) \tag{1}$$

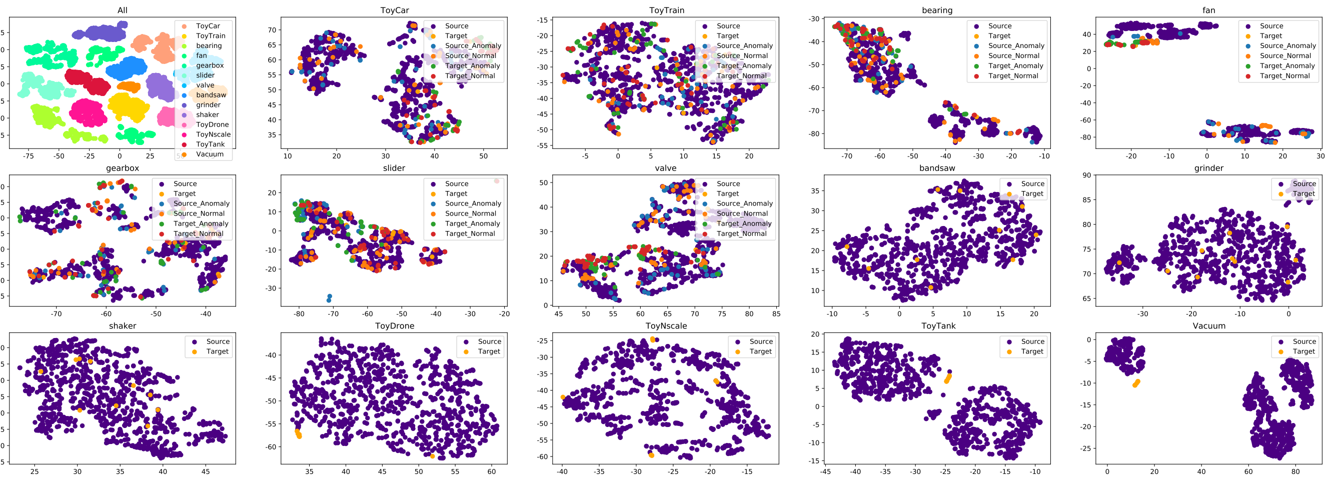

Figure 1: The visualization of the embeddings of the MV2_SEC by t-SNE.

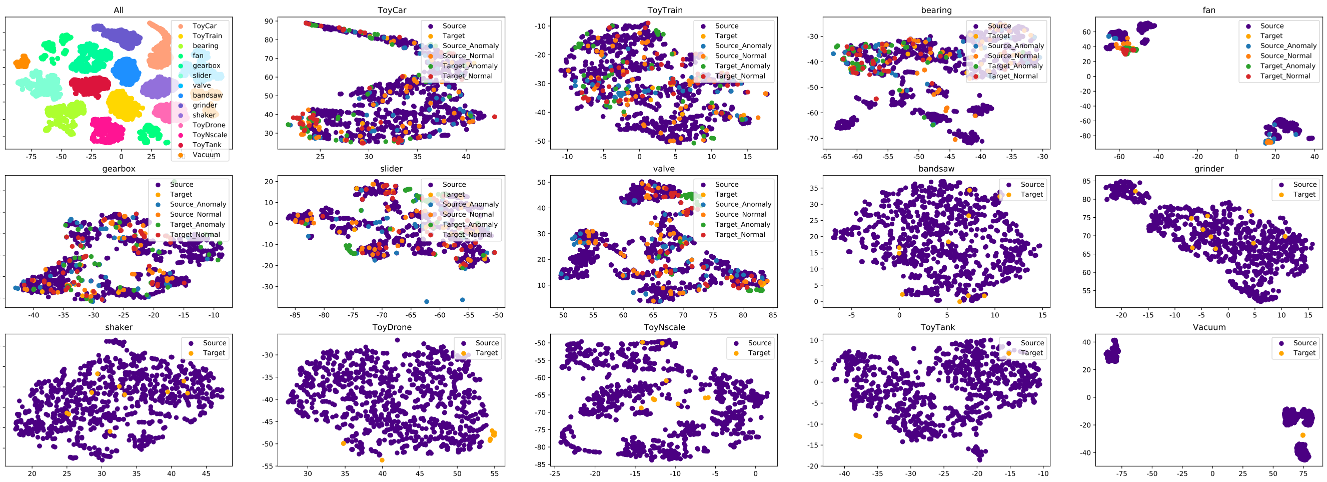

Figure 2: The visualization of the embeddings of the MV2_SED by t-SNE.

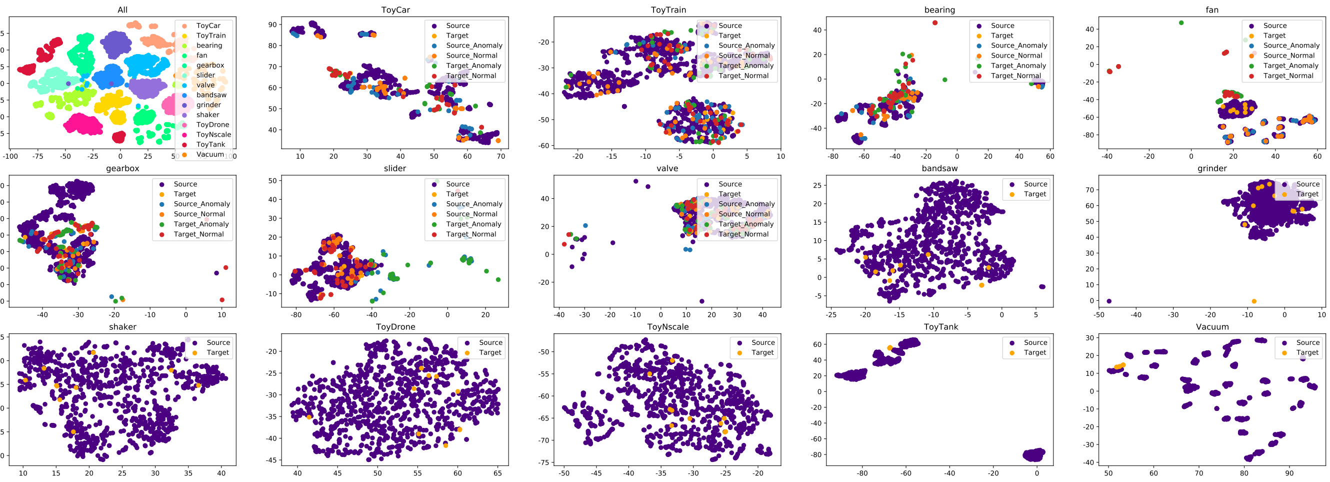

Figure 3: The visualization of the embeddings of the ViT_SEC by t-SNE.

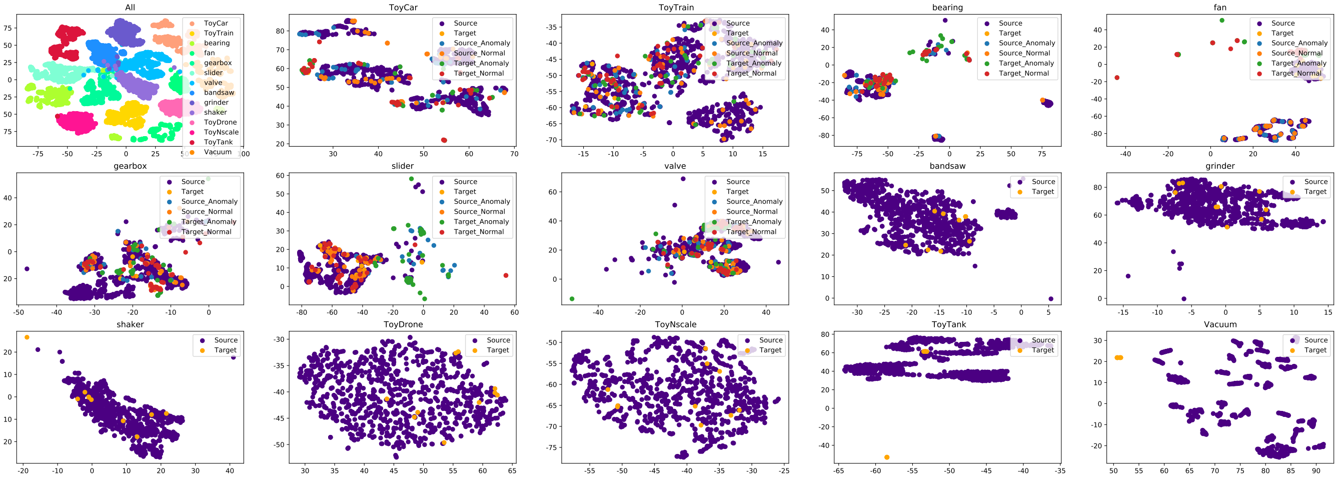

Figure 4: The visualization of the embeddings of the ViT_SED by t-SNE.

Table 2: Training Configurations for Different Models.

	Train clip	Inference clip	Inference stride	Iterations	lr
MV2 Pretrain	10s	-	-	50000000	0.001
MV2 Finetune	3s	3s	0.5s	40000	0.001
ViT	3s	3s	0.5s	40000	0.001

2.1.2. *MODEL FINETUNE*

In the fine-tuning stage, we used the pre-trained model for model parameter initialization to train the sound event classification (SEC) and sound event detection (SED) models. The aim of the SEC training is to optimize the cross-entropy (CE) loss function, which is defined as:

$$\mathcal{L}_{CE} = -\boldsymbol{y}\log(\hat{\boldsymbol{y}}) \tag{2}$$

The aim of the SED training is also to optimize the BCE loss function. It is worth noting that the output time frame number of SED is 1. During the SED training process, several events will be simulated simultaneously with a certain probability. Compared with the SEC model, the advantage of this setting is that it can prevent the machine from misjudgement when it is disturbed by the sound of other machines, because the SEC's output probability will be affected when multiple events occur at the same time.

To reduce the calculation complexity of the subsequent process, the output channel of the Conv2d 1x1 layer is set to 128, which means that the dimension of the embeddings output by the model is 128.

2.2. ViT

The main idea of ViT [7] is to divide an image into a fixed number of small blocks, and then convert these small blocks into vectors. These vectors are passed as input to the Transformer encoder, which learns how to combine these vectors under a specific task. Like MobileNetV2, we train an SEC and an SEC model separately.

The detailed configuration of ViT is as follows. Use 16 × 16 patch to block the input features, and the block features are changed from 256 to 128 dimensions after linear transformation. The depth of the Transformer encoder is set to 1, and the multi-head attention mechanism is used, and the head is set to 8. The mlp dimension in the feedforward neural network is set to 4 times the feature dimension.

Table 3: Model Ensemble Configuration.

	Ensenmble
Ensemble1	**MV2_SEC + MV2_SED**
Ensemble2	**ViT_SEC + ViT_SED**
Ensemble3	**Ensemble1 + Ensemble2**
Ensemble4	**Ensemble1 * Ensemble2**

2.3. ANOMALY SCORE

We use the Mahalanobis distance as the anomaly score. The formula for calculating Mahalanobis distance is

$$\mathcal{A} = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \tag{3}$$

where $\boldsymbol{x}$ is the mean vector, $\boldsymbol{\mu}$ is the mean vector representation corresponding to the machine and $\boldsymbol{\Sigma}$ is the covariance matrix corresponding to the machine. For the source domain and the target domain, we use the data of different domains to calculate the average vector $\boldsymbol{\mu}$, and use all the training data of the machine to calculate the covariance matrix $\boldsymbol{\Sigma}$.

Table 4: The average results of AUC-S for different machine types.

	ToyCar	ToyTrain	bearing	fan	gearbox	slider	valve	Average
AE_MSE	70.10%	57.93%	65.92%	80.19%	60.31%	70.31%	55.35%	64.79%
AE_MAH	**74.53%**	55.98%	65.16%	**87.10%**	71.88%	84.02%	56.31%	68.84%
MV2_SEC	61.60%	65.60%	74.72%	60.40%	70.48%	93.80%	69.84%	69.59%
MV2_SED	61.08%	61.12%	73.92%	66.32%	81.56%	91.96%	68.04%	70.58%
ViT_SEC	53.60%	50.68%	56.60%	73.40%	86.56%	**99.24%**	64.12%	65.48%
ViT_SED	51.00%	51.64%	57.56%	73.92%	86.80%	98.92%	63.12%	65.21%
Ensemble1	62.20%	**65.96%**	**75.64%**	65.64%	78.00%	94.88%	**70.00%**	**71.91%**
Ensemble2	52.08%	51.72%	55.20%	74.76%	89.52%	99.20%	64.08%	65.49%
Ensemble3	57.32%	57.56%	67.68%	71.24%	**90.04%**	**99.24%**	66.44%	70.09%
Ensemble4	57.84%	57.80%	70.12%	71.44%	89.48%	99.16%	66.92%	70.67%

Table 5: The average results of AUC-T for different machine types.

	ToyCar	ToyTrain	bearing	fan	gearbox	slider	valve	Average
AE_MSE	46.89%	57.02%	55.75%	36.18%	60.69%	48.77%	50.69%	49.59%
AE_MAH	43.42%	42.45%	55.28%	45.98%	70.78%	73.29%	51.40%	52.37%
MV2_SEC	54.04%	**63.28%**	**69.20%**	51.00%	66.59%	88.84%	59.60%	62.83%
MV2_SED	54.40%	53.80%	59.00%	55.04%	71.92%	87.12%	**70.60%**	62.70%
ViT_SEC	49.44%	58.20%	61.64%	71.48%	**80.08%**	97.44%	57.52%	64.96%
ViT_SED	54.16%	57.44%	64.24%	74.72%	75.24%	**97.88%**	56.80%	66.08%
Ensemble1	54.88%	58.84%	65.72%	53.68%	70.48%	89.16%	66.68%	63.95%
Ensemble2	52.84%	57.64%	63.08%	**73.68%**	79.56%	97.76%	57.36%	66.08%
Ensemble3	**55.08%**	60.84%	65.32%	62.44%	80.04%	97.16%	60.68%	66.55%
Ensemble4	**55.08%**	60.56%	65.48%	66.80%	79.36%	96.72%	61.04%	**67.16%**

Table 6: The average results of pAUC for different machine types.

	ToyCar	ToyTrain	bearing	fan	gearbox	slider	valve	Average
AE_MSE	**52.47%**	48.57%	50.42%	59.04%	53.22%	56.37%	51.18%	52.84%
AE_MAH	49.18%	48.13%	51.37%	**59.33%**	54.34%	54.72%	51.08%	52.36%
MV2_SEC	50.94%	50.58%	61.42%	50.52%	55.73%	68.68%	53.05%	55.19%
MV2_SED	47.58%	49.47%	57.52%	52.95%	55.95%	64.95%	52.73%	53.95%
ViT_SEC	51.26%	**50.63%**	50.26%	50.57%	**58.15%**	91.10%	53.68%	**55.67%**
ViT_SED	51.21%	49.16%	50.58%	51.26%	53.42%	90.68%	52.78%	54.75%
Ensemble1	48.47%	49.74%	**61.68%**	51.73%	56.89%	68.11%	53.36%	55.00%
Ensemble2	51.71%	49.95%	51.10%	50.31%	55.26%	**91.47%**	53.47%	55.32%
Ensemble3	51.10%	48.89%	53.05%	51.57%	53.31%	89.53%	53.95%	55.24%
Ensemble4	50.89%	49.00%	54.15%	51.36%	53.47%	87.95%	**54.26%**	55.34%

3. EXPERIMENTS

3.1. DATASETS

3.1.1. MobileNetV2

The pretrained dataset used in this work is Audioset [9]. And in the fine-tuning stage, this work uses the DCASE 2023 Challenge Task2 dataset [10, 11], which includes recordings of 14 classes and a sampling rate of 16 kHz.

3.1.2. ViT

The model only use the DCASE 2023 Challenge Task2 dataset, which includes recordings of 14 classes and a sampling rate of 16 kHz.

3.2. SETUP

We load the audio data using the default sample rate and apply a short time Fourier transform (STFT) with a window size of 512 and a hop length of 160. The STFT spectrogram convert into a Mel spectrogram with a 64-band Mel filter. We used Adam as the optimizer and the learning rate of the model was set to 0.001. Training runs with a batch size of 32, and see Table 2 for detailed configurations.

4. RESULT

To show the performance, we evaluate the detection performance of the area under the receiver operating characteristic curve (AUC) and the partial AUC (pAUC) with $p = 0.1$. Table 3 shows the configurations of the 4 ensemble models we submitted. We simply add or multiply the anomaly scores output from different models to obtain the final anomaly score.

AE_MSE and **AE_MAH** indicate that the official model uses MSE and Mahalanobis distance as anomaly scores.

MV2_SEC and **MV2_SED** respectively represent the SEC model and SED model trained by MobileNetV2.

ViT_SEC and **ViT_SED** respectively represent the SEC model and SED model trained by ViT.

The experimental results are shown in Table 4, 5, and 6. The average represents the harmonic mean. We also used t-SNE [12] to visualize the embeddings of 14 machines, as shown in Figures 1, 2, 3, and 4.

5. CONCLUSION

We presented our submission systems for DCASE2023 Challenge Task 2 in this technical report, using two proposed methods. Experimental results show that our proposed systems outperformed the baseline systems.

6. REFERENCES

[1] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, November 2020, pp. 81–85. [Online]. Available: http://dcase.community/documents/workshop2020/proceedings/DCASE2020Workshop_Koizumi_3.pdf

[2] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, 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," *arXiv preprint arXiv:2206.05876*, 2022.

[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, Tech. Rep., 2022.

[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 arXiv e-prints: 2303.00455*, 2023.

[5] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Description and discussion on dcase 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *In arXiv e-prints: 2305.07828*, 2023.

[6] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[8] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[9] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[10] 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.

[11] 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.

[12] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.