HIERARCHICAL METADATA INFORMATION CONSTRAINED SELF-SUPERVISED LEARNING FOR ANOMALOUS SOUND DETECTION UNDER DOMAIN SHIFT

Haiyan Lan[1,2], *Qiaoxi Zhu*[3], *Jian Guan*[1,2*], *Yuming Wei*[1,2], *Wenwu Wang*[4]

[1]Group of Intelligent Signal Processing, College of Computer Science and Technology, Harbin Engineering University, China
[2]National Engineering Laboratory for Modelling and Emulation in E-government, Harbin Engineering University, China
[3]Centre for Audio, Acoustics and Vibration, University of Technology Sydney, Australia
[4]Centre for Vision Speech and Signal Processing, University of Surrey, UK

ABSTRACT

Self-supervised learning methods have achieved promising performance for anomalous sound detection (ASD) under domain shift by incorporating the metadata of domain shift types and machine sound attributes in feature learning. However, the relation between domain shifts and machine sound attributes has yet to be fully utilised despite their potential benefits for characterising domain shifts. This paper presents a hierarchical metadata information constrained self-supervised ASD method, where the hierarchical relation between domain shift types (section IDs) and attributes is constructed and used as constraints to improve feature representation. In addition, we propose an attribute-group-centre based method for calculating the anomaly score under the domain shift condition. Experiments show improved audio feature learning over the state-of-the-art methods in DCASE 2022 challenge Task 2.

Index Terms— Anomalous sound detection, domain shift, self-supervised learning, metadata

1. INTRODUCTION

Anomalous sound detection (ASD) is a task for automatically identifying the working condition of a machine as normal or abnormal based on the sound emitted from the machine. Due to the difficulty in collecting rare and diverse anomalous sounds, it is a challenging unsupervised learning task with only normal sounds available for model training [1]. Unsupervised ASD methods based on autoencoder (AE) [2–4] or self-supervised classification models [5,6] with metadata (e.g. machine IDs) incorporated achieved state-of-the-art performance on the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge 2020 Task 2 [1]. However, ASD often has limited performance in practice due to the domain shift problem [7]. That is, acoustic characteristics differ between the source domain (in training) and the target domain (in detection), with the change of attributes [7] (e.g., machine operating conditions or types of noise). Due to this problem, the anomalies in the target domain can be misidentified with the model trained using sounds from the source domain.

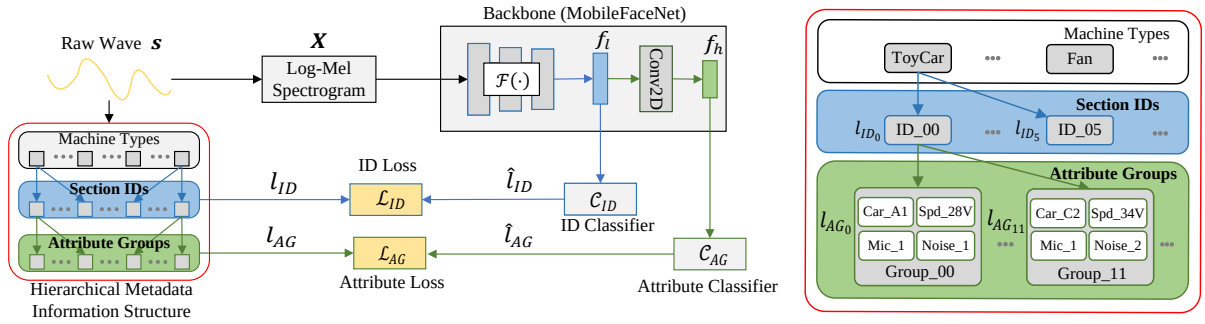
With a focus on the domain shift problem, DCASE 2022 challenge Task 2 launched a new task for unsupervised ASD [7–10]. Its dataset has hierarchical metadata of machine type, section ID and attributes, as we illustrated in Fig. 1(b). Each section ID refers to a subset of the data within a domain shift scenario under a machine type, and domain shifts result from the change of attributes, e.g., the machine's operation speed and the environmental noise level. The 1st ranked method in the challenge [8] adopts self-supervised classification with section IDs as labels for feature learning. On the other hand, our previous work achieved 3rd place [9] using attributes as labels, considering their effect on acoustic characteristics. However, only using either section IDs [8,11] or attributes [9] may not be sufficient to obtain features helpful for characterising domain shifts.

Existing methods [12, 13] used both section IDs and attributes in a parallel way. Taking attributes and section IDs in parallel assumes that the attributes and section IDs are independent or that the same attribute works equally under different domain shifts. However, the same attribute can impact the machine sound differently under different domain shift types (section IDs). Thus, the relation between attributes and domain shift types has yet to be fully utilised despite their potential benefits for characterising domain shifts.

This paper is the first study exploiting the implicit hierarchical relation between domain shift type and attribute for more effective feature learning for anomalous sound detection under domain shift. We propose a hierarchical metadata information constrained (HMIC) self-supervised method us-

*Corresponding Author

This work was partly supported by the Natural Science Foundation of Heilongjiang Province under Grant No. LH2022F010 and No. YQ2020F010, and a Newton Institutional Links Award from the British Council with Grant No. 623805725.



(a) Training procedure of the proposed model

(b) Hierarchical metadata information structure

Fig. 1. Framework of the proposed HMIC method, where the hierarchical relation of the metadata information is exploited by the introduced hierarchical metadata information structure, and a backbone network (i.e., MobileFaceNet [14]) is used for the extraction of low-level feature $\boldsymbol{f}_l$ and high-level feature $\boldsymbol{f}_h$, which are constrained by section ID label l_{ID}, and attribute group label l_{AG}, respectively. Here, $\mathcal{F}(\cdot)$ denotes the feature extractor.

ing domain shift types and attributes in a hierarchical way. Specifically, we set the attribute groups (AGs) under each section ID (domain shift type) to cluster the data with the same attributes' values as an attribute group, as shown in Fig. 1 (b). Then, we use the hierarchical relation as the constraint in self-supervised learning to obtain finer audio feature representation, with the section IDs characterising the type of domain shift for low-level feature learning and the attribute groups exploiting acoustic characteristics of each domain shift for high-level feature learning. Moreover, we propose an attribute group centre (AGC) based method to calculate anomaly scores. AGC represents the average of the learnt audio features from each attribute group. We calculate the anomaly score using the minimal Mahalanobis distance between the test sound's audio feature and AGCs to better adapt to the variance of domain shifts. Experiments conducted on the DCASE 2022 challenge Task 2 dataset demonstrate the proposed method's improved self-supervised audio feature learning compared to the state-of-the-art methods.

2. PROPOSED METHOD

This section introduces our proposed HMIC, as illustrated in Fig. 1, which consists of a backbone (i.e., MobileFaceNet [14]) for feature extraction, and two classifiers (ID classifier $\mathcal{C}_{ID}$ and attribute classifier $\mathcal{C}_{AG}$) to predict section ID and attribute group label, respectively. It uses a hierarchical metadata information structure to exploit the implicit relation between section ID and attributes for finer feature learning. In addition, we introduce an AGC-based method for calculating the anomaly scores in the detection stage.

2.1. Hierarchical Metadata Information Structure

Addressing domain shift in ASD, metadata information (i.e., section IDs and attributes) related to domain shift is utilised, with their hierarchical relation being further exploited through a hierarchical metadata information structure in Fig. 1 (b). To emphasise audio clips under each section ID may have certain attributes with different values, we cluster data with the same attributes' values as an attribute group under this section ID. Therefore, each section ID has several AGs, and each AG under the same section ID has the same attribute types but different values. So, we constructed a metadata information tree for each machine type, with section IDs as nodes and AGs as leaves, as in Fig. 1 (b).

Taking the machine type ToyCar in DCASE 2022 challenge Task 2 [7] 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 to construct the hierarchical relation between section ID and attributes. Therefore, we can employ this hierarchical relation between section IDs and attributes as the constraint to learn finer audio features to mitigate the domain shift issue in ASD.

2.2. Hierarchical Constrained Classification

With the hierarchical information discussed above, we can employ section IDs and AGs as the self-supervision labels to constrain the learning of the low-level and high-level audio features, respectively, as shown in Fig. 1 (a). Here, low and high levels indicate the output coarse and fine-grained features from our model's low and high-level layers with the hierarchical constraint, respectively.

The log-Mel spectrogram $\boldsymbol{X}$ of the input audio signal $\boldsymbol{s}$ is the input of our model. We obtain the low-level feature $\boldsymbol{f}_l$ and high-level feature $\boldsymbol{f}_h$ from the backbone network

(MobileFaceNet [14]) via a feature extractor $\mathcal{F}(\cdot)$ and a 2-dimensional convolutional layer (Conv2D), respectively,

$$\boldsymbol{f}_l = \mathcal{F}(\boldsymbol{X}) \tag{1}$$

$$\boldsymbol{f}_h = \text{Conv2D}(\boldsymbol{f}_l) \tag{2}$$

To utilise hierarchical relation to learn features relevant to domain shift, section ID and AG are employed as self-supervision labels, l_{ID} and l_{AG}, to constrain the learning of $\boldsymbol{f}_l$ and $\boldsymbol{f}_h$, respectively. First, two simple linear classifiers (ID classifier $\mathcal{C}_{ID}$ and attribute classifier $\mathcal{C}_{AG}$) are adopted for section ID label and AG label prediction, that $\hat{l}_{ID} = \mathcal{C}_{ID}(\boldsymbol{f}_l)$ and $\hat{l}_{AG} = \mathcal{C}_{AG}(\boldsymbol{f}_h)$, respectively. Then, an ID loss $\mathcal{L}_{ID}$ and an attribute loss $\mathcal{L}_{AG}$ are introduced to constrain the process of learning the low-level and high-level features, using e.g. the cross-entropy (*CE*) loss,

$$\mathcal{L}_{ID} = CE(l_{ID}, \hat{l}_{ID}) \tag{3}$$

$$\mathcal{L}_{AG} = CE(l_{AG}, \hat{l}_{AG}) \tag{4}$$

Finally, the total loss $\mathcal{L}_{total}$ for model training is

$$\mathcal{L}_{total} = \lambda \mathcal{L}_{ID} + (1 - \lambda)\mathcal{L}_{AG} \tag{5}$$

The weight λ is empirically tuned for each machine type.

2.3. Attribute Group Centre-based Anomaly Detection

Anomaly score calculation is the key to evaluating the test sound in the anomaly detection stage. We introduce the attribute group centre (AGC) to calculate the anomaly score. Each attribute group's AGC is the average of the learnt audio features from that attribute group. Then, the feature of the test sound is compared with all the AGCs to allow better anomaly detection in the presence of domain shift.

Assume N training audio clips under the m-th attribute group with the label $l_{AG_{m-1}}$, $m = 1, 2, \ldots, M$ and M is the number of attribute groups under the corresponding section ID. The m-th attribute group centre $\boldsymbol{c}_m$ is

$$\boldsymbol{c}_m = \frac{1}{N} \sum_{n=1}^{N} \hat{\boldsymbol{f}}_{h_n} \tag{6}$$

where $\hat{\boldsymbol{f}}_{h_n}$ denotes the high-level audio feature derived from the model for the n-th training audio clip, $n = 1, 2, \ldots, N$.

Then, the Mahalanobis distance [15] is used to measure similarity between the audio feature representation $\bar{\boldsymbol{f}}$ of the sound under test and each AGC $\boldsymbol{c}_m$, $m = 1, 2, \ldots, M$, and the minimal Mahalanobis distance is taken as the anomaly score $\mathcal{A}$

$$\mathcal{A} = \min_{m \in [1,M]} \sqrt{(\bar{\boldsymbol{f}} - \boldsymbol{c}_m)^T \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{f}} - \boldsymbol{c}_m)} \tag{7}$$

where $\boldsymbol{\Sigma}^{-1}$ is the inverse of the covariance matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}$ is obtained from the feature of all the audio clips under the m-th attribute group of the same section.

Our proposed ASD method with the AGC-based anomaly score calculation is named **HMIC-AGC**. The proposed method is later compared with **HMIC-DC**, which calculates the Mahalanobis distance between $\bar{\boldsymbol{f}}$ and the domain centre (DC), i.e., the average feature of each domain, instead of each attribute group, to derive the anomaly score. Though DC is widely used for anomaly score calculation, such as the 1st ranked method in DCASE 2022 challenge [8], DC uses average features from multiple domains, while AGC considers acoustic characteristics of each specific domain to obtain more accurate feature representation under domain shifts.

3. EXPERIMENTAL RESULTS

3.1. Experimental Setup

Dataset The training data for model training is from the development and the additional datasets of the DCASE 2022 challenge Task 2 [7], which includes five different machine types (bearing, fan, gearbox, slider, and valve) [16] and two types of toys (i.e., ToyCar and ToyTrain) [17]. Each machine type contains six section IDs, each with 990 and 10 audio clips from the source and target domain, respectively. We evaluate the performance on the evaluation dataset of the DCASE 2022 challenge Task 2. Note that the evaluation dataset's domain information (source and target) is unknown to verify the generalization ability of the ASD systems.

Evaluation Metrics The evaluation metrics include the area under the receiver operating characteristic curve (AUC), partial AUC (pAUC), and the harmonic mean of AUC and pAUC scores over all the machine types, sections, and domains [7].

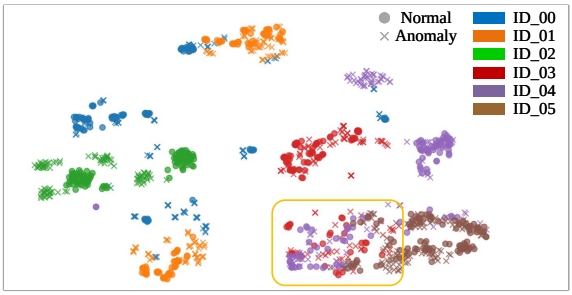
Implementation The log-Mel spectrogram of the audio clips is used as the input feature for our model, where the frame size is set as 1024 with an overlap of 50%, and the number of Mel filter banks is 128. The dimension of the input log-Mel spectrogram is 128 × 313. Our model is trained with 120 epochs, using Adam optimizer [19] with an initial learning rate of 0.0001, and the cosine annealing strategy is then applied for changing the learning rate.

3.2. Experimental Results

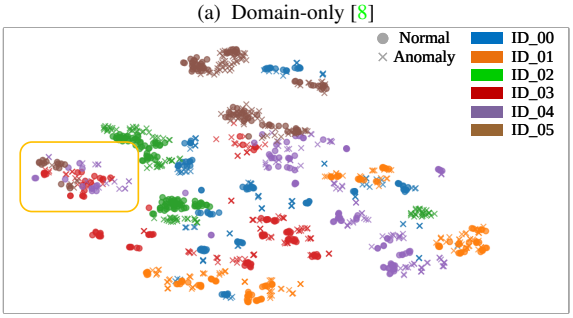
Performance Comparison Our proposed HMIC-AGC method hierarchically uses the domain shift type and attribute for self-supervised ASD with AGC for anomaly score calculation. In the experiment, it was compared with baseline methods (AE [18] and MobileNetV2 [18]), the self-supervised methods only using the attribute (Attribute-only, as the 3rd ranked method in DCASE 2022 Task 2 [9]) or section ID (Domain-only, as the 1st ranked method [8]), and HMIC-DC using DC for anomaly score calculation. For a fair comparison, all methods are performed without model pretraining, and all adopt log-Mel spectrogram as the input without the high-pass filter in [8] or temporal information fusion in [9].

Table 1. Performance comparison in terms of AUC (%) and pAUC (%) on the evaluation dataset of DCASE 2022 challenge Task 2. **Total**: harmonic mean (%) of AUC and pAUC scores over all the machine types, sections, and domains.

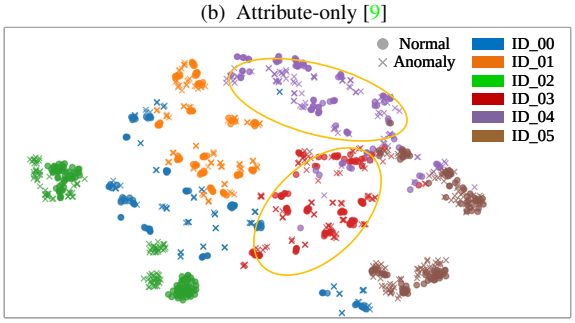
Methods	ToyCar		ToyTrain		Bearing		Fan		Gearbox		Slider		Valve		Total	
	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
AE [18]	61.18	60.21	43.14	49.36	59.93	53.95	41.16	50.12	61.92	51.95	58.95	54.16	54.26	51.30	53.01	52.80
MobileNetV2 [18]	42.79	53.44	51.22	50.98	58.23	52.16	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)

Fig. 2. The t-SNE visualisation of the learnt audio features using different self-supervised methods for machine type Bearing. Different colours represent different section IDs. "•" and "×" respectively represent normal and anomalous sounds.

As can be seen from Table 1, both HMIC-DC and HMIC-AGC can significantly improve the detection performance for all machine types except Fan, as compared with the baseline methods, i.e., AE [18] and MobileNetV2 [18]. In addition, the proposed methods achieve the best overall performance compared to Domain-only or Attribute-only methods from the 1st and 3rd ranked submissions in DCASE 2022, respectively. Although the pAUC performance on Slider and Fan of HMIC-AGC is slightly lower than the Domain-only method, it significantly improves the pAUC performance on ToyCar and Gearbox, with 10.04% and 7.99% improvement, respectively. Specifically, both HMIC-DC and HMIC-AGC can improve the total harmonic mean performance, and HMIC-AGC achieves the best total harmonic mean performance. The results demonstrate the effectiveness of the hierarchical metadata information constraint and the AGC-based anomaly scores calculation, which show the superior generalisation ability of our proposed methods for ASD under domain shift conditions. In addition, the performance on the development set has the same trend as that on the evaluation set, though not presented in this paper.

Visualisation Analysis To further verify the effectiveness of the proposed HMIC for improved feature learning, the test data (with all 6 section IDs) from both development and evaluation datasets are evaluated. The t-distributed stochastic neighbour embedding (t-SNE) [20] cluster visualisation of the learnt features using section ID only, attribute only, or the proposed method are illustrated in Fig. 2. It can be seen that audio features are misclassified in the presence of overlapping between sections ID_03 and ID_04, when only using section ID or attribute (metadata without hierarchical relation) for self-supervised classification. In contrast, they can be distinguished with the proposed method from different sections as the areas marked with orange in Fig. 2 (c). The results demonstrate the effect of HMIC for more distinguishable feature learning under domain shift.

4. CONCLUSION

We have presented a self-supervised method for anomalous sound detection under domain shift, where a hierarchical metadata information structure is constructed and used as the constraint in self-supervised learning for improved feature learning. In addition, an attribute group centre based anomaly scores calculation method is introduced, which further enhances the domain generalisation ability by considering the attributes of domain shift. Experimental results demonstrate the effectiveness of the proposed method, with substantial improvements in the audio feature learning over those that only use section ID or attributes, as in the state-of-the-art methods in DCASE 2022 challenge Task 2.

5. 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 DCASE 2020 challenge Task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proceedings of DCASE Workshop*, 2020, pp. 81–85.

[2] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 271–275.

[3] R. Giri, F. Cheng, K. Helwani, S. V. Tenneti, U. Isik, and A. Krishnaswamy, "Group masked autoencoder based density estimator for audio anomaly detection." in *Proceedings of DCASE Workshop*, 2020, pp. 51–55.

[4] S. Kapka, "ID-conditioned auto-encoder for unsupervised anomaly detection," in *Proceedings of DCASE Workshop*, 2020, pp. 71–75.

[5] R. Giri, F. Cheng, K. Helwani, S. V. Tenneti, U. Isik, and A. Krishnaswamy, "Self-supervised classification for detecting anomalous sounds," in *Proceedings of DCASE Workshop*, 2020, pp. 46–50.

[6] K. Morita, T. Yano, and K. Tran, "Anomalous sound detection using CNN-based features by self supervised learning," DCASE 2021 Challenge, Tech. Rep., July 2021.

[7] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Description and discussion on DCASE 2022 challenge Task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," in *Proceedings of DCASE Workshop*, Nancy, France, November 2022.

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

[9] F. Xiao, Y. Liu, Y. Wei, J. Guan, Q. Zhu, T. Zheng, and J. Han, "The DCASE 2022 challenge Task 2 system: Anomalous sound detection with self-supervised attribute classification and GMM-based clustering," DCASE 2022 Challenge, Tech. Rep., July 2022.

[10] Y. Wei, J. Guan, H. Lan, and W. Wang, "Anomalous sound detection system with self-challenge and metric evaluation for DCASE 2022 challenge Task 2," DCASE 2022 Challenge, Tech. Rep., July 2022.

[11] I. Kuroyanagi, T. Hayashi, K. Takeda, and T. Toda, "Two-stage anomalous sound detection systems using domain generalization and specialization techniques," DCASE 2022 Challenge, Tech. Rep., July 2022.

[12] Y. Deng, J. Liu, and W.-Q. Zhang, "AITHU system for unsupervised anomalous detection of machine working status via sounding," DCASE 2022 Challenge, Tech. Rep., July 2022.

[13] K. Wilkinghoff, "An outlier exposed anomalous sound detection system for domain generalization in machine condition monitoring," DCASE 2022 Challenge, Tech. Rep., July 2022.

[14] S. Chen, Y. Liu, X. Gao, and Z. Han, "MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices," in *Proceedings of Chinese Conference on Biometric Recognition (CCBR)*. Springer, 2018, pp. 428–438.

[15] Y. Sakamoto and N. Miyamoto, "Anomaly calculation for each components of sound data and its integration for DCASE 2020 challenge Task2," DCASE 2020 Challenge, Tech. Rep., July 2020.

[16] 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 DCASE Workshop*, Nancy, France, November 2022, pp. 1–5.

[17] 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 DCASE Workshop*, 2021, pp. 1–5.

[18] K. Dohi, K. Imoto, Y. Koizumi, and N. Daisuke, "Description and discussion on DCASE 2022 challenge Task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," DCASE 2022 Challenge, Tech. Rep., July 2022.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2015.

[20] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of Machine Learning Research (JMLR)*, vol. 9, no. 11, p. 2579–2605, 2008.