# ANOMALOUS SOUND DETECTION USING AUDIO REPRESENTATION WITH MACHINE ID BASED CONTRASTIVE LEARNING PRETRAINING

*Jian Guan*[1]*, Feiyang Xiao*[1]*, Youde Liu*[2]*, Qiaoxi Zhu*[3]*, Wenwu Wang*[4]

[1]Group of Intelligent Signal Processing, College of Computer Science and Technology, Harbin Engineering University, Harbin, 150001, China
[2]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China
[3]Centre for Audio, Acoustics and Vibration, University of Technology Sydney, Ultimo, NSW, Australia
[4]Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK

## ABSTRACT

Existing contrastive learning methods for anomalous sound detection refine the audio representation of each audio sample by using the contrast between the samples' augmentations (e.g., with time or frequency masking). However, they might be biased by the augmented data, due to the lack of physical properties of machine sound, thereby limiting the detection performance. This paper uses contrastive learning to refine audio representations for each machine ID, rather than for each audio sample. The proposed two-stage method uses contrastive learning to pretrain the audio representation model by incorporating machine ID and a self-supervised ID classifier to fine-tune the learnt model, while enhancing the relation between audio features from the same ID. Experiments show that our method outperforms the state-of-the-art methods using contrastive learning or self-supervised classification in overall anomaly detection performance and stability on DCASE 2020 Challenge Task2 dataset.

***Index Terms***— Anomalous sound detection, metadata information, contrastive learning, self-supervised learning

## 1. INTRODUCTION

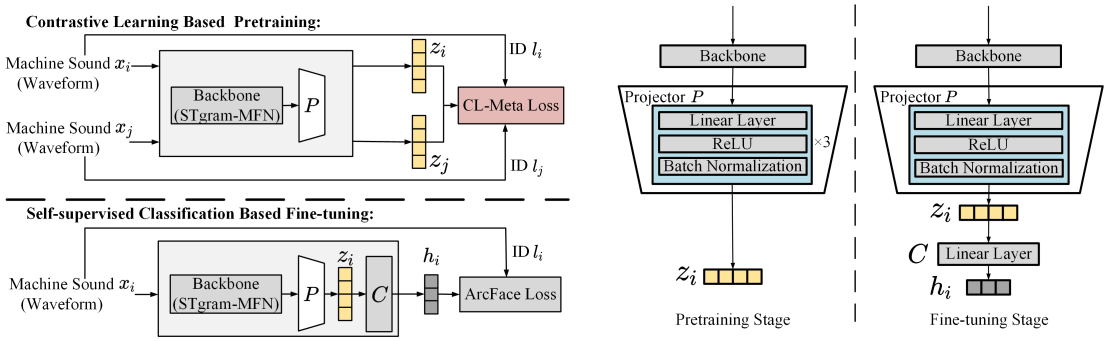
Anomalous sound detection (ASD) aims to detect the unknown anomalous sounds with only normal sounds available in training [1–7]. It has the potential in acoustic scene monitoring [3], quality assurance [8], and artificial intelligence-based factory automation [5]. Due to the unavailability of anomalous sounds, the audio feature acquisition and representation of normal sounds is key to distinguishing the normal and unknown anomalous sounds [9, 10].

To learn the audio representation, early methods employ auto-encoder for reconstructing the input audio feature (i.e., log-Mel spectrogram) and employ the reconstruction error as the anomaly score for anomalous sound detection [1,2]. However, these methods often ignore the use of metadata about audio files that can describe the states or properties of machines, e.g., machine ID. The operating sounds of machines with different IDs often have unique characteristics reflecting the difference between machines. In this case, the learnt audio representation may be degraded by the difference between machines of different IDs under the same machine type, which can limit the detection performance [3, 9].

As a solution, the self-supervised classification methods [3, 4] employ machine IDs accompanying the machine sound as labels to improve the learning of audio features from different IDs, which may offer better performance [3]. However, these methods do not effectively enhance the inter-class relation between audio features with the same machine ID. As a result, the learnt features may not be sufficiently fine-grained for anomalous sound detection. These methods may perform differently even for machines of the same type, leading to instability for anomaly detection [4, 9]. In our recent work, we have further improved the performance and stability of the self-supervised classification method by introducing the spectral-temporal feature, i.e., STgram [9]. However, the inter-class relation between the learnt features from the same ID is rarely considered. Thus, the normal and anomalous sounds from the same ID cannot be well distinguished. The learnt feature still has the potential to be further improved. Recent studies in image representation learning indicate that contrastive learning may perform better for feature learning than self-supervised classification methods [11, 12], and could help enhance the inter-class relation between samples [12].

Following the success of contrastive learning in image representation, e.g., SimCLR [11], contrastive learning has also been introduced for audio representation in recent ASD studies [13, 14], where each audio signal is represented with audio embeddings using data augmentation (e.g., time masking, frequency masking). More specifically, the embeddings from the same audio signal are moved closer together, while the embeddings from different machine sounds are moved

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




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

(b) Different model structures in two stages.

**Fig. 1**. Framework of the proposed CLP-SCF method, where the training procedure includes two phases: contrastive learning based pretraining and self-supervised classification based fine-tuning. Different model structures are adopted in these two stages. In the pretraining stage, the model consists of a backbone module (i.e., STgram-MFN [9]) for audio feature extraction and a projector module $P$ for audio feature embedding. In the fine-tuning stage, a self-supervised classifier $C$ is used in addition to the backbone and projector modules for fine-tuning the model.

away from each other via contrastive processing, when learning the audio feature representation of the normal sounds. However, in practice, the embeddings of these audio signals may still be far away from each other. In other words, these methods can only learn general representation of normal machine sounds, which may not be fine-grained enough to distinguish the anomalous sounds from the normal sounds.

In this paper, we introduce contrastive learning to exploit the latent relation between the machine sound and its corresponding metadata (machine ID), and present a two-stage training method for representation learning of audio features in anomalous sound detection by combining contrastive learning based pretraining and self-supervised classification based fine-tuning (CLP-SCF). In our method, a backbone module (i.e., STgram-MFN [9]) is employed for audio feature extraction, and a multi-layer perception (MLP) is adopted to map the extracted audio feature to the audio embedding for contrastive learning. In the pretraining stage, a metadata-based contrastive learning (CL-Meta) loss is introduced for fine-grained feature learning, which not only increases the intra-class difference between audio features from different IDs by pushing their audio embeddings away from each other, but also enhances the inter-class relation between audio features from the same ID by clustering their audio embeddings together. In the fine-tuning stage, a self-supervised ID classifier is adopted to fine-tune our model by distinguishing the audio features of different IDs to further enhance the distinguishing ability of the learnt audio representation. Experiments conducted on DCASE 2020 dataset [1] show that the proposed method outperforms the state-of-the-art methods in both detection performance and stability.

## 2. PROPOSED METHOD

This section presents the proposed two-stage CLP-SCF method in detail. The overall framework and the training procedure are shown in Fig. 1. In our method, a novel metadata-based contrastive learning (CL-Meta) loss is introduced for audio feature pretraining, and a self-supervised classifier is then adopted for fine-tuning the model to learn improved audio representation. Our model includes a backbone module (i.e., STgram-MFN [9]) for audio feature extraction and an MLP projector module to obtain audio embeddings, with different structures in two stages as shown in Fig. 1(b).

### 2.1. Contrastive Learning Based Pretraining

In the pretraining stage, a novel metadata-based contrastive learning (CL-Meta) loss is introduced for audio feature pretraining. As the learnt feature captures the relation between audio signals and their corresponding machine IDs, it offers a better ability to identify the sound from different IDs and enhance the relation between sounds from the same ID.

Supposing $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_N]$ is a set of input audio signals that includes $N$ machine sounds. We select the $i$-th machine sound $\boldsymbol{x}_i$ $(1 \leq i \leq N)$ as the anchor, and build the contrast with the remaining $(N-1)$ audio signals. The machine ID label of $\boldsymbol{x}_i$, defined as $l_i$, and its audio embedding $\boldsymbol{z}_i \in \mathbb{R}^D$ can be extracted via the backbone and the projector modules, as shown in Fig. 1. For the remaining $(N-1)$ audio signals $\boldsymbol{x}_j$ $(1 \leq j \leq N, j \neq i)$, we can obtain their corresponding ID label $l_j$ and audio embedding $\boldsymbol{z}_j$ in the same way. For contrastive learning, we can use the cosine similarity score defined in terms of the audio embeddings $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ as

$$s_{i,j} = \frac{\boldsymbol{z}_i^\top * \boldsymbol{z}_j}{\|\boldsymbol{z}_i\|_2 \|\boldsymbol{z}_j\|_2} \tag{1}$$

where $*$ denotes matrix multiplication, $\top$ represents transposition operation, and $\|\cdot\|_2$ is the $l_2$-norm function.

To capture the relation of the audio embeddings from the same ID and distinguish audio embeddings from different IDs, the cosine similarity score of audio embeddings from

the same ID is expected to be maximized, whereas the cosine similarity score of the embeddings from different IDs to be minimized. Therefore, following [12], our CL-Meta loss for audio feature learning can be defined as

$$L_{\text{CL-Meta}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|K(i)|}\sum_{k\in K(i)}\log\frac{\exp(s_{i,k}/\tau)}{\sum_{j\neq i}^{N}\exp(s_{i,j}/\tau)} \quad (2)$$

where $\tau$ is the temperature scalar to scale the cosine similarity scores, which is used to enlarge the distance between audio embeddings from different IDs. $K(i) = \{k|1 \leq k \leq N, \text{and } k \neq i, l_i = l_k\}$ denotes the set of indexes that have the same ID as audio index $i$. $k$ is an index from $K(i)$, and $|K(i)|$ represents the number of indexes in $K(i)$.

With the contrastive learning loss Eq. (2), we can obtain a more effective audio representation to enhance the relation between sound from the same ID and the difference between different machine sounds. We can then use the learnt model parameters for the initialization of the model in the fine-tuning stage.

### 2.2. Self-supervised Classification Based Fine-tuning

With the pretrained audio representation, we then fine-tune our model by a self-supervised ID classifier with ArcFace loss [15] to further enhance the distinguishing ability of the learnt audio representation.

Note that, a different model structure is applied in this stage as illustrated in Fig. 1 (a) and (b), where a simple classifier is introduced after the projector to learn the latent feature $\boldsymbol{h}_i$ from the audio embedding $\boldsymbol{z}_i$ for ID prediction. Then, following [9], we employ the self-supervised classification loss, i.e., ArcFace loss [15] for the model fine-tuning, which can further improve the ability to distinguish the audio features from different IDs. The ArcFace loss is calculated as

$$L_{\text{ArcFace}} = \text{ArcFace}(\boldsymbol{h}_i, l_i). \quad (3)$$

For the anomalous sound detection, we use the proposed CLP-SCF method to predict the ID of an estimated machine sound, and calculate the negative log probability of the estimated machine sound and its corresponding ID as the anomaly score for anomalous sound detection. That is, a normal sound is less likely to be predicted as a non-corresponding ID. Therefore, in the inference stage, if the predicted ID differs from the actual ID, it will be considered an anomalous sound.

## 3. EXPERIMENTS AND RESULTS

### 3.1. Experimental Setup

***Dataset*** Following [9], our CLP-SCF method is evaluated on the DCASE 2020 Challenge Task2 development and additional datasets, which include four machine types (i.e., Fan, Pump, Slider and Valve) from the MIMII dataset [16] and two machine types (i.e., ToyCar and ToyConveyor) from the ToyADMOS dataset [17]. Each machine type has seven different machines, except for ToyConveyor, which only has six different machines. Therefore, we have audio signals from 41 different machines (41 ID labels), where each audio signal is around 10 seconds. The training data of the development dataset and the additional dataset are combined as the training set, and our model is trained for all machine IDs. The normal and anomalous sound from the test data of the development dataset is adopted for model evaluation.

Note that in our experiments, the DCASE 2022 Challenge Task2 dataset [5] was not used since it is designed to investigate domain shift, where the distribution of audio features of machine sounds may change from the known source domain to the unknown target domain. This is out of the scope of our work here, as we address the audio representation of sounds under known status. Therefore, we use DCASE 2020 dataset instead in our experiments.

***Implementation Details*** In the pretraining stage, we randomly select 6 machine sounds from each ID to construct the set of input audio signals with the batch size of 246 ($41 \times 6$), which is used to build the contrast for the signals from all ID labels. Adam optimizer [18] with a learning rate of 0.0005 is used for model optimization, and the model is pretrained with 100 epochs. The temperature score $\tau$ in Eq. (2) is empirically selected as 0.05 following [12].

In the fine-tuning stage, the batch size is converted to 128, the learning rate is set as 0.0001, and our model is fine-tuned with 300 epochs. For the self-supervised classification, the margin and scale hyper-parameters of the ArcFace loss are set as 1.0 and 30, respectively. Note that, the cosine annealing strategy is adopted as the learning rate decay schedule in both stages [19].

***Performance Metrics*** Following [1–4, 9], we employ the area under the receiver operating characteristic curve (AUC) and the partial-AUC (pAUC) for performance evaluation. Here, pAUC denotes the AUC value over a low false-positive rate range $[0, p]$, where $p$ is set as 0.1 following [1, 9]. Meanwhile, minimum AUC (mAUC) is also adopted for detection stability evaluation, which reflects the worst detection performance of the machines from the same machine type [4, 9].

### 3.2. Performance Comparison

To show the performance of the proposed CLP-SCF, we compare our method with the state-of-the-art methods on DCASE 2020 Task2 dataset, including IDNN [2], MobileNetV2 [3], Glow_Aff [4], STgram-MFN (ArcFace) [9] and AADCL [13]. Here, IDNN is the AE-based method without machine information, and MobileNetV2, Glow_Aff, and STgram-MFN (ArcFace) are the state-of-the-art self-supervised classification methods that also adopt machine ID for anomaly detection. The AADCL is the method using contrastive learning

**Table 1**. Performance comparison in terms of AUC (%) and pAUC (%) on the test data of the development dataset.

| 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** |

**Table 2**. Performance comparison in terms of mAUC (%).

| 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** |

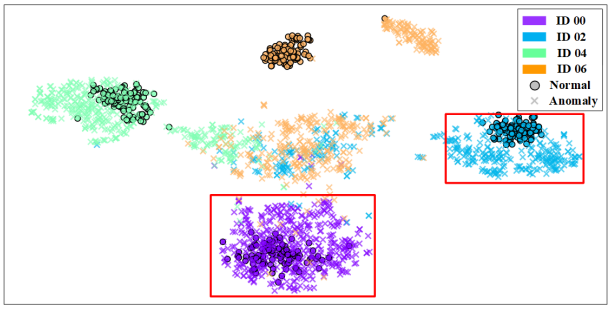
to learn audio representation via data augmentation, without exploring the relation between machine sound and its corresponding metadata. The results are shown in Table 1.

Our proposed method can achieve the best performance in terms of average AUC and average pAUC on all machine types, which provides 1.39% and 2.14% improvements in terms of average AUC and average pAUC, respectively, over the second best method, i.e., STgram-MFN (ArcFace), which is the backbone of our method. Except for the pAUC performance on ToyConveyor, the proposed method offers better detection performance in terms of both AUC and pAUC for all machine types, than the state-of-the-art self-supervised classification methods that also adopt machine IDs as self-supervision labels, and the contrastive learning based ASD method, i.e., AADCL.

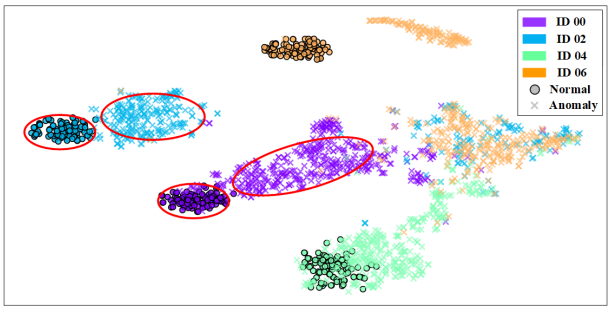
### 3.3. Detection Stability

Table 2 presents the mAUC performance of our CLP-SCF, as compared with that of STgram-MFN (ArcFace) [9]. The study in [9] significantly improved the detection stability and performance via its proposed spectral-temporal fusion feature (STgram). From Table 2, we can see that our proposed method can further improve detection stability with significant improvement in mAUC for all machine types, especially for the machine types of Fan, Pump, and ToyCar. The results from Tables 1 and 2 verify the effectiveness of the proposed method for improving detection performance and stability.

To illustrate the effect of the learnt audio feature representation, Fig. 2 shows the t-distributed stochastic neighbour embedding (t-SNE) cluster visualization of the latent features of these two methods, where we can see that our method shows better distinguishing ability. For example, compared to STgram-MFN (ArcFace), the proposed method significantly reduces the overlapping between the normal and anomalous latent features of "ID 00" and "ID 02" in Fig. 2. The result further demonstrates the effectiveness of the proposed method.




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




(b) Proposed CLP-SCF

**Fig. 2**. The t-SNE visualization of STgram-MFN (ArcFace) [9] and the proposed CLP-SCF for the machine type Fan. (a) denotes the latent feature distribution obtained using the STgram-MFN (ArcFace) method. (b) denotes the latent feature distribution obtained using the proposed CLP-SCF method. The symbol "•" and "×" denote normal and anomalous sound classes, respectively. The normal and anomalous latent feature distributions for "ID 00" and "ID 02" are highlighted by the red contours.

## 4. CONCLUSION

In this paper, we have studied the relation between the metadata and the machine sound in audio representation for anomalous sound detection. We have presented a two-stage method to improve the quality of the audio representation, which consists of model pretraining using the metadata-based contrastive learning in the first stage, and model fine-tuning using the self-supervised ID classification in the second stage. Experiments show that the proposed method achieves better detection performance than the state-of-the-art methods.

## 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 DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proc. of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 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 *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 271–275.

[3] R. Giri, S. V. Tenneti, F. Cheng, K. Helwani, U. Isik, and A. Krishnaswamy, "Self-supervised classification for detecting anomalous sounds," in *Proc. of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2020, pp. 46–50.

[4] K. Dohi, T. Endo, H. Purohit, R. Tanabe, and Y. Kawaguchi, "Flow-based self-supervised density estimation for anomalous sound detection," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 336–340.

[5] 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," *In arXiv e-prints: 2206.05876*, 2022.

[6] F. Xiao, Y. Liu, Y. Wei, J. Guan, Q. Zhu, T. Zheng, and J. Han, "The DCASE2022 challenge task 2 system: Anomalous sound detection with self-supervised attribute classification and GMM-based clustering," DCASE2022 Challenge, Tech. Rep., July 2022.

[7] Y. Wei, J. Guan, H. Lan, and W. Wang, "Anomalous sound detection system with self-challenge and metric evaluation for DCASE2022 challenge task 2," DCASE2022 Challenge, Tech. Rep., July 2022.

[8] B. Chen, L. Bondi, and S. Das, "Learning to adapt to domain shifts with few-shot samples in anomalous sound detection," *arXiv preprint arXiv:2204.01905*, 2022.

[9] Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous sound detection using spectral-temporal information fusion," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 816–820.

[10] 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 *Proc. of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2020, pp. 51–55.

[11] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. of International Conference on Lachine Learning (ICML)*. PMLR, 2020, pp. 1597–1607.

[12] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, vol. 33. Curran Associates, Inc., 2020, pp. 18 661–18 673.

[13] H. Hojjati and N. Armanfard, "Self-supervised acoustic anomaly detection via contrastive learning," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3253–3257.

[14] X. Cai and H. Dinkel, "A contrastive semi-supervised learning framework for anomaly sound detection," in *Proc. of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2021, pp. 31–34.

[15] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699.

[16] H. Purohit, R. Tanabe, T. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in *Proc. of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2019, pp. 209–213.

[17] 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. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 308–312.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *Proc. of International Conference on Learning Representations (ICLR)*. OpenReview.net, 2017.