# Machine Anomalous Sound Detection Based on Self-Supervised Classification

Shuxian Wang†, Jun Du†* and Yajian Wang†
† National Engineering Research Center of Speech and Language Information Processing
University of Science and Technology of China, Hefei, Anhui, China
E-mail: sxwang21@mail.ustc.edu.cn, jundu@ustc.edu.cn, yajian@mail.ustc.edu.cn

***Abstract*—The Machine Anomalous Sound Detection task aims to design a system to detect unknown anomalous sounds given only the sounds of machines working normally. The sounds emitted by different types of machines often have different characteristics, and the environments in which the machines work (such as temperature, noise, etc.) are constantly changing, which also affects the acoustic characteristics of the machine sound, so this is a challenging task. To this end, we propose a method for anomalous sound detection based on self-supervised classification. First, we obtain an effective feature representation of the sound by extracting frequency domain and time domain features from the raw wave and extracting pre-trained features based on the pre-trained model. Then, we design an auxiliary loss based on the attribute information of the audio, which helps the model to distinguish different operating conditions of the machine. Finally, we extract latent representations from the trained model, and calculate the anomaly score of the machine based on the distance metric. Experimental results on the DCASE 2022 Challenge Task 2 dataset demonstrate the effectiveness of our method. Moreover, we analyze the complementarity between different feature representations, which proves that the feature representations used in our method are effective.**

## I. Introduction

Anomalous Sound Detection (ASD) is the task of judging whether a machine is working normally or abnormally based on the sound it makes while working. The detection of machine anomalous sounds is of great significance to the development of the industry, so more and more researches have been carried out on the ASD task in recent years. ASD is a challenging task. On the one hand, in real-world conditions, it is often easier for us to obtain the sound of the machine working normally, while the anomalies are rare and highly diverse. Therefore, we need to design the ASD system to detect unknown anomalous sounds using only normal sounds. On the other hand, the acoustic characteristics of machine sounds are affected by changing operating conditions (such as environmental noise, temperature, etc.), which makes it difficult for the ASD system to accurately detect whether the machine is working abnormally.

Recently, many methods have been used in the ASD task. Among them, one main type is based on generative models, such as autoencoder (AE) [1], [2], [3], [4], [5], [6], [7], interpolation deep neural network (IDNN) [8], flow [9], WaveNet [10]. Another mainstream method is to use the information of machine type and section ID to implement ASD based on classification [11], [12], [13], [14], [15], [16], [17], [18]. Different section IDs of a machine represent different machine operating conditions that have changed, such as operating voltage, factory noise, etc. However, methods for detecting anomalies based on reconstruction errors (such as AE) often suffer from performance degradation due to the limitation that only normal samples can be used during training, if the characteristics of some training samples are similar to the abnormal samples during testing. Classification-based methods are also likely to perform poorly because some of the training samples (normal) have similar characteristics to the test samples (normal or abnormal). At the same time, if we only use the section ID information of the machine for classification and ignore the different working conditions of the machine, the ASD system often cannot detect anomalous sounds well when encountering new working conditions. Therefore, the discriminativeness of the acoustic features used and the ability of the model to distinguish and adapt to different working conditions are the keys to the success of the ASD system.

Log-Mel spectrograms are often used as input features for ASD systems [19], [20], [21], [22]. However, log-Mel spectrograms may lose useful high-frequency information, so time-domain features are extracted from the raw wave and concatenated with frequency-domain features to obtain STgram features [23]. However, compared to log-Mel spectrograms, the performance of ASD systems may be affected by the noise contained in the temporal information that Tgram features may bring in [23].

In order to obtain a more discriminative feature representation of machine sounds and help the model to distinguish and adapt to different machine conditions, we develop an effective ASD system. First, to utilize the complementarity between different features, we extract the pre-trained features from the raw wave based on the pre-trained model, and concatenate them with the frequency domain and time domain features. Second, in addition to the section ID, we also use the attribute information (including different operating conditions of the machine) to design auxiliary loss, so that the ASD system can adapt to the new environments to achieve good anomaly detection performance. Finally, in order to better measure the difference between abnormal samples and normal samples, we extract the latent embeddings of training and test data through the trained self-supervised classifier to measure the cosine

*corresponding author

978-616-590-477-3 ©2022 APSIPA APSIPA ASC 2022

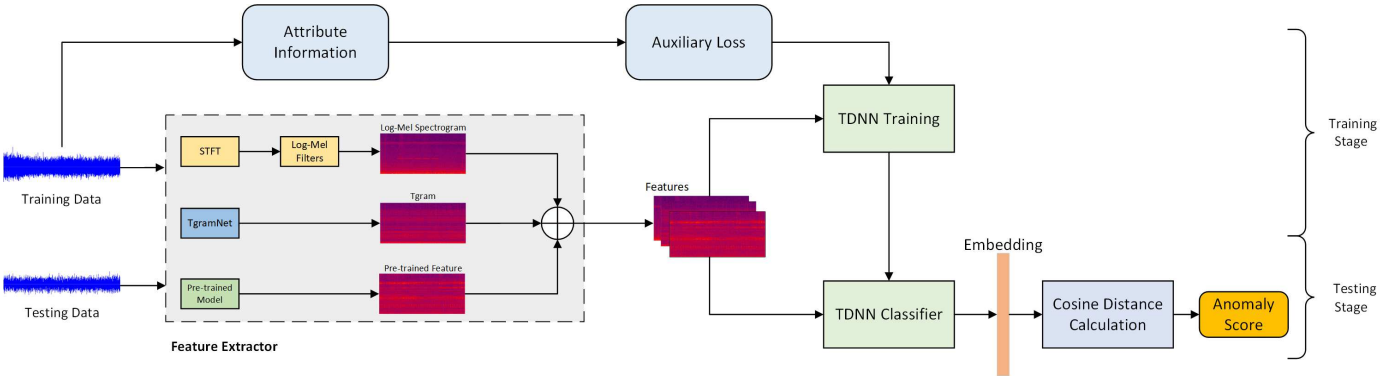



Fig. 1. The framework of the proposed anomalous sound detection system.

distance between test samples and training data to calculate the anomaly score.

## II. Method

The overall framework of our proposed anomalous sound detection method is shown in Fig. 1. For the training data, we first obtain an effective representation for each audio through a feature extractor, and then a self-supervised classifier is trained to learn the representation. Among them, the features extracted by the feature extractor include log-Mel spectrograms, Tgram features and pre-trained features. Specifically, the log-Mel spectrograms are generated by short-time Fourier transform (STFT) and Mel-filter banks, and Tgram features are extracted from the raw wave based on TgramNet [23]. Finally, in order to improve the representation ability of the feature and the generalization of the model, we extract pre-trained features based on the pre-trained model. During training, the attribute information of each audio is used to calculate the auxiliary loss, which helps us better train the self-supervised classifier. For the test data, firstly, the effective representation is obtained based on the feature extractor, and then the trained self-supervised classifier is used to obtain the embeddings, so that the cosine distances between the embeddings of the test sample and the training data are calculated. There are only normal sounds in the training data, while there are normal and anomalous sounds in the test data. For anomalous samples, they are obviously farther away from the normal samples, so their anomaly score is higher.

### A. Feature Extraction

In order to obtain a more powerful representation, the raw wave needs to be processed by three parts of the feature extractor, and finally the three features are concatenated together as a representation of the input audio.

The input $a$ is a single-channel audio signal. First, the spectrogram is obtained based on the short-time Fourier transform, and then the log-Mel spectrogram (denoted as $F_L$) is obtained by passing through the Mel-filter banks and taking the logarithm as follows:

$$F_{\mathrm{L}} = \text{log-Mel}(||\mathrm{STFT}(a)||^2), \tag{1}$$

where log-Mel represents Mel-filter banks and logarithmic operation, and the $F_{\mathrm{L}}$ dimension is $F \times T$ ($F$ is the number of Mel-frequency bins, and $T$ is the number of frames in the time domain).

The log-Mel spectrogram may lose critical high frequency information because the Mel-filter banks are used. In previous work, [23] has demonstrated that extracting temporal feature information from the raw wave based on TgramNet [23] is highly complementary to log-Mel spectrogram. Therefore, we obtain the Tgram feature $F_{\mathrm{T}}$ based on TgramNet as follows:

$$F_{\mathrm{T}} = \mathrm{TgramNet}(a), \tag{2}$$

where TgramNet is a CNN-based network [23], and the dimension of $F_{\mathrm{T}}$ is $F \times T$, which is the same as that of the log-Mel spectrogram $F_{\mathrm{L}}$.

We use the pre-trained model to increase the generalization of the model. First, general training is performed on a large number of labeled or unlabeled data, and then fine-tuning is conducted with a limited amount of target data to improve the performance of downstream tasks.

We employ the wav2vec 2.0 [24] pre-trained model, which uses a multilayer convolutional neural network to process the raw wave of speech audio. We fine-tune it on the dataset used in our experiments, and then based on this fine-tuned model, latent representations $F_{\mathrm{P}}$ of the training and test data are obtained as follows:

$$F_{\mathrm{P}} = \mathrm{PTM}(a), \tag{3}$$

where PTM represents the pre-trained model, and the dimension of $F_{\mathrm{P}}$ is $F \times T$.

Finally, we concatenate the above three features to obtain an effective representation $F_{\mathrm{LTP}}$ of the input audio as follows:

$$F_{\mathrm{LTP}} = \mathrm{Concatenate}(F_{\mathrm{L}}, F_{\mathrm{T}}, F_{\mathrm{P}}). \tag{4}$$

where the dimension of $F_{\mathrm{LTP}}$ is $3 \times F \times T$.

### B. Auxiliary Loss

The datasets we use are from the development dataset and additional training dataset of the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge 2022 Task

2 [25]. They are recorded in different environments or states (e.g., different noises, different machine speeds, etc.), and the file name of each audio gives its attribute label information. Taking Slide rail in the dataset as an example, we can know its attribute information such as velocity, acceleration and factory noise. In the real world, the operating environment and state of each machine are constantly changing, so if the attribute information such as different environments or states can be distinguished, it will help the model to adapt to new environmental noise, machine working state and potential other new conditions, so that the model can detect anomalous sounds more accurately in different environments or states.

Therefore, when training the self-supervised classifier, in addition to the cross entropy (CE) loss and batch hard triplet loss [26] used to classify the section ID of the machine, an auxiliary loss is also added to classify the different attribute information of the machine. The loss function we use is as follows:

$$Loss = L_{\mathrm{CE}} + L_{\mathrm{Triplet}} + L_{\mathrm{Auxiliary}}. \tag{5}$$

### C. Distance Metric

In order to get the outlier of each test sample, firstly, we extract the embeddings of the training and test data based on the trained self-supervised classifier, and then measure the cosine distance between the test sample and the training data (all normal samples) to obtain the anomaly score. Obviously, for the anomalous sound in the test data, it is farther from normal samples, and the anomaly score is higher, so as to achieve the purpose of detecting anomalous sound. Compared to directly obtaining the posterior probability from the softmax layer of the self-supervised classifier to calculate the anomaly score, the distance metric approach can characterize the difference between anomalous and normal samples based on a more effective latent representation obtained from the model.

In the dataset we use, the test data has three types of section IDs, namely Section 00, Section 01 and Section 02. The first way to calculate the anomaly score is to obtain the posterior probability $p_\theta$ that each test sample predicted by the model belongs to its correct section ID from the output of the softmax layer, so as to calculate the anomaly score as follows:

$$A_\theta(X_j) = \log \frac{1 - p_\theta}{p_\theta}, \tag{6}$$

where $X_j$ is the audio feature of the $j$-th test sample, $p_\theta$ is the softmax output of the model for the correct section ID, and $A_\theta$ is the anomaly score for each audio.

To calculate the anomaly score based on the distance metric, firstly, the embeddings of the training and test data are obtained from the trained model, and then the distances between the test sample and all the training data belonging to the same section ID are calculated, and they are sorted from smallest to largest. Thus, the anomaly score can be calculated as follows:

$$A_\theta(X_j) = 1 - \max_{k=1,\ldots,K} (cos(E_{i,j}^{\mathrm{Test}}, E_{i,k}^{\mathrm{Train}})). \tag{7}$$

where $X_j$ is the audio feature of the $j$-th test sample, $E_{i,j}^{\mathrm{Test}}$ is the embedding of the the $j$-th test sample whose section ID is $i$, and $E_{i,k}^{\mathrm{Train}}$ is the embedding of the the $k$-th training data whose section ID is also $i$. There are a total of $K$ training data with section ID $i$.

## III. Experiments and Analysis

### A. Dataset and Evaluation Metrics

We conduct experiments using the DCASE Challenge 2022 Task 2 [25] development and additional training datasets, which are generated from the ToyADMOS2 [27] and MIMII DG [28] datasets, with a total of seven classes of machines, namely ToyCar and ToyTrain from ToyADMOS2 and Fan, Gearbox, Bearing, Slide rail (Slider), and Valve from MIMII DG. Each recording is a single-channel and 10-second long audio. Each type of machine in the development dataset has three section IDs (i.e., Section 00, Section 01 and Section 02), and each type of machine in the additional training dataset also has three section IDs (i.e., Section 03, Section 04 and Section 05). The section is the unit in which performance metrics are calculated. In our experiments, the training data (only normal sound) from the development dataset and the additional training dataset are used as training set, and the test data (normal and anomaly sound) in the development dataset are used for evaluation. When different audios are recorded, there are differences in properties such as factory noise, running speed, acceleration, etc. The data of each type of machine is divided into the source domain and the target domain. From the source domain to the target domain, there are differences in these attributes. The training set gives these attribute information. The specific number of attribute classes is expressed as machine type (number of attributes) as follows: ToyCar (22), ToyTrain (23), Fan (12), Gearbox (44), Bearing (34), Slider (37), Valve (15).

The machine anomalous sound detection task is evaluated using the area under the receiver operating characteristic (ROC) curve (AUC) and the partial-AUC (pAUC) [25]. The pAUC is calculated as the AUC over a low false-positive-rate (FPR) range [0, $p$] ($p = 0.1$) [25]. The final anomaly score $\Omega$ is given by the harmonic mean of the AUC and pAUC scores over all the machine types, sections, and domains as follows:

$$\begin{aligned} \Omega = h[\ &\mathrm{AUC}_{\mathrm{m,n,d}},\ \mathrm{pAUC} \mid m \in M, \\ &n \in S(m),\ d \in \{\mathrm{source}, \mathrm{target}\}\ ], \end{aligned} \tag{8}$$

where $h[\cdot]$ represents the harmonic mean (over all machine types, sections, and domains), $M$ represents the set of machine types, and $S(m)$ represents the set of sections for machine type $m$.

Similarly, the anomaly score for each machine is given by the harmonic mean of the AUC and pAUC scores over all the sections and domains as follows:

$$\begin{aligned} \Omega = h[\ &\mathrm{AUC}_{\mathrm{n,d}},\ \mathrm{pAUC} \mid n \in S(m), \\ &d \in \{\mathrm{source}, \mathrm{target}\}\ ]. \end{aligned} \tag{9}$$

TABLE I
EXPERIMENTAL CONFIGURATIONS FOR ABLATION EXPERIMENTS

| Experiment No. | Features | Loss | Anomaly Score Calculation Method |
|---|---|---|---|
| 1(Baseline) | LogMel | CE Loss+Batch Hard Triplet Loss | Classification Confidence (Eq. (6)) |
| 2 | LogMel+Tgram | CE Loss+Batch Hard Triplet Loss | Classification Confidence |
| 3 | LogMel+Pre-trained Features | CE Loss+Batch Hard Triplet Loss | Classification Confidence |
| 4 | LogMel+Tgram+Pre-trained Features | CE Loss+Batch Hard Triplet Loss | Classification Confidence |
| 5 | LogMel+Tgram+Pre-trained Features | CE Loss+Batch Hard Triplet Loss+Auxiliary Loss | Classification Confidence |
| 6 | LogMel+Tgram+Pre-trained Features | CE Loss+Batch Hard Triplet Loss+Auxiliary Loss | Distance Metric (Eq. (7)) |

### *B. Ablation Experiments*

On the basis of the baseline system, we conduct ablation experiments from three aspects: features, loss, and anomaly score calculation methods. The experimental numbers and their corresponding experimental configurations are shown in Table I. It is worth noting that in the following experiments, except for the features, loss, and anomaly score calculation methods, other experimental configurations are the same as the baseline system.

*1) Baseline System:* The log-Mel spectrogram is used as the input to the baseline system. To generate log-Mel spectrogram, the short-time Fourier transform (STFT) with 1024 FFT points is applied, utilizing a window size of 1024 samples, a hop length of 512 samples and a dimensional Mel basis of 128. We choose the time-delay neural network (TDNN) as the baseline classifier. To train this model, Adam optimizer [29] is used with the learning rate of 0.0002 for 100 epochs, and CE loss and batch hard triplet loss are adopted. In addition, the calculation method of the anomaly score of the test data is shown in Eq. (6). The performance of the baseline system on the test data is shown in Table II.

TABLE II
MACHINE ANOMALY SCORES FOR BASELINE SYSTEM (%)

| Experiment No. | 1(Baseline) |
|---|---|
| ToyCar | 50.08 |
| ToyTrain | 48.96 |
| Bearing | 59.82 |
| Fan | 57.83 |
| Gearbox | 63.56 |
| Slider | 77.18 |
| Valve | 69.92 |
| Total | 59.63 |

*2) Pre-trained Features:* We use TgramNet to extract Tgram features and concatenate them with log-Mel spectrograms to obtain STgram [23] features. Meanwhile, the wav2vec2.0 pre-trained features are concatenated with log-Mel spectrograms. Finally, the two features are connected to log-Mel spectrograms simultaneously. The experimental results are shown in Table III.

As can be seen from Table III, after the Tgram features and pre-trained features are concatenated respectively, the total anomaly score is improved from 59.63% to 61.05% and 60.38%, respectively. When these two features are concatenated at the same time, the total anomaly score improves from 59.63% to 63.26%, a total improvement of 3.63%. And in the anomaly scores of each type of machines, except for Bearing, the other six types of machines perform better than the baseline. However, when these two features are concatenated separately, the anomaly scores on only some machines are improved relative to the baseline, which illustrates that the feature representations obtained by concatenating these two features at the same time are highly discriminative.

TABLE III
MACHINE ANOMALY SCORES UNDER DIFFERENT FEATURES (%)

| Experiment No. | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| ToyCar | 50.08 | 48.10 | 49.88 | **50.46** |
| ToyTrain | 48.96 | 50.14 | 50.20 | **51.32** |
| Bearing | 59.82 | 57.34 | 55.96 | 59.16 |
| Fan | 57.83 | 74.54 | 55.55 | **70.43** |
| Gearbox | 63.56 | 62.09 | 69.29 | **65.07** |
| Slider | 77.18 | 71.10 | 73.43 | **78.76** |
| Valve | 69.92 | 77.49 | 83.01 | **82.01** |
| Total | 59.63 | 61.05 | 60.38 | **63.26** |

*3) Auxiliary Loss:* Through the above ablation experiments, we demonstrate the effectiveness of concatenating Tgram features and pre-trained features with the log-Mel spectrogram simultaneously. Based on this, the auxiliary loss is designed to train the model based on the attribute information. The loss function is shown in Eq. (5). Auxiliary loss can be used to distinguish different environments and states of audio recording, so as to help the model better adapt to different environments and states, and then more accurately detect anomalous sounds. The experimental results are shown in Table IV.

TABLE IV
MACHINE ANOMALY SCORES UNDER DIFFERENT LOSS FUNCTIONS (%)

| Experiment No. | 4 | 5 |
|---|---|---|
| ToyCar | 50.46 | **53.92** |
| ToyTrain | 51.32 | 51.20 |
| Bearing | 59.16 | 57.85 |
| Fan | 70.43 | **78.34** |
| Gearbox | 65.07 | **74.39** |
| Slider | 78.76 | 78.10 |
| Valve | 82.01 | 79.01 |
| Total | 63.26 | **65.40** |

Observing Table IV, we can find that after the auxiliary loss is added, although it is slightly decreased on some machines, the performance is significantly improved on three machines (i.e. ToyCar, Fan, Gearbox), and the anomaly score is increased by 3.46%, 7.91%, and 9.32% respectively, and the final total

anomaly score is also increased from 63.26% to 65.40%, which shows the effectiveness of the auxiliary loss.

*4) Distance Metric:* In the above experiments, we used the posterior probability to calculate the anomaly score (as shown in Eq. (6)). On the basis of the previous experiments, we extract the embeddings of training and test data based on the trained self-supervised classifier, and then use the distance metric to calculate the anomaly score of the machine (as shown in Eq. (7)). The experimental results are shown in Table V. It can be found that the performance of the ToyCar has been significantly improved after using the distance metric, and the anomaly score of ToyCar has increased from 53.92% to 69.70%. At the same time, there are also certain improvements in ToyTrain, Bearing, and Gearbox. On the total anomaly score, it improved by 2.89%, which demonstrates that the distance metric can more effectively measure the difference between anomalous and normal samples based on the latent representation obtained from the model.

TABLE V
MACHINE ANOMALY SCORES UNDER DIFFERENT ANOMALY SCORE CALCULATION METHODS (%)

| Experiment No. | 5 | 6 |
|---|---|---|
| ToyCar | 53.92 | **69.70** |
| ToyTrain | 51.20 | **52.04** |
| Bearing | 57.85 | **58.70** |
| Fan | 78.34 | 76.95 |
| Gearbox | 74.39 | **76.83** |
| Slider | 78.10 | 77.35 |
| Valve | 79.01 | 77.09 |
| Total | 65.40 | **68.29** |

## C. Performance Comparison

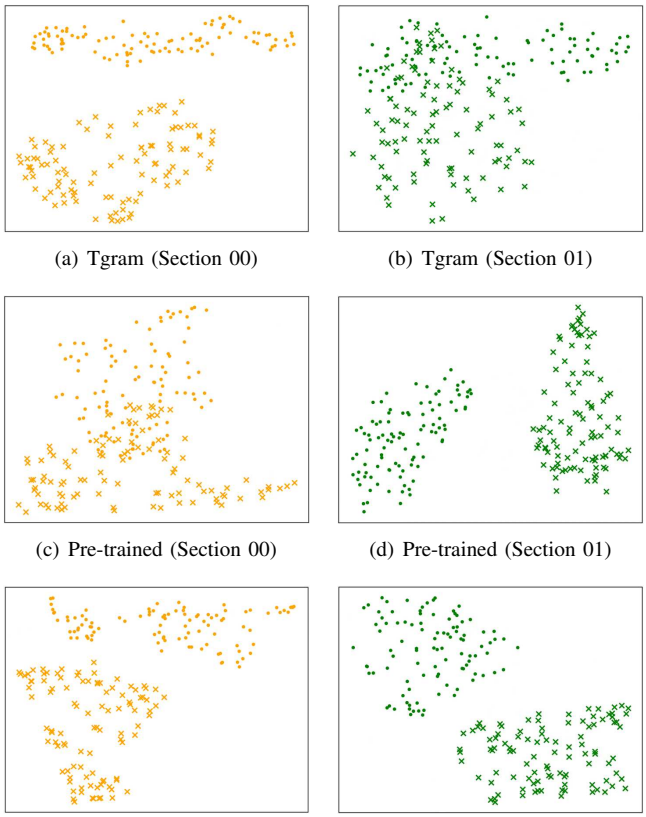
Table VI shows the performance comparison of our proposed method with other competing methods (AE, MobileNetV2) [25]. AE is based on reconstruction error and MobileNetV2 is based on machine section ID classification to detect anomalous sounds. It can be seen that our method outperforms other methods on six types of machines. Moreover, compared with the best performance of other methods, our method has a significant improvement on ToyCar, Fan, Gearbox, Slider, and Valve by 15.30%, 18.45%, 13.76%, 19.35%, and 14.96%, respectively. Furthermore, our method also improves by 11.71% over the best results of other methods on the total anomaly score, which clearly shows that our method has good performance for anomalous sound detection of different machines.

## D. Results Analysis

As can be seen from Table III, compared to concatenating one of the Tgram features and the pre-trained features separately, when these two features are concatenated with the log-Mel spectrogram at the same time, the anomaly score of most types of machine and the total anomaly score will be better, which reflects the complementarity of these two features. Taking Slider as an example, the t-distributed Stochastic Neighbor Embedding (t-SNE) cluster visualization of latent features when Tgram features and pre-trained features are concatenated separately and when they are concatenated simultaneously is shown in Fig. 2. As can be seen from Fig. 2(a) and Fig. 2(b), the normal and abnormal samples of Section 00 can be well distinguished, but some of the normal and abnormal samples of Section 01 are overlapping. In Fig. 2(c) and Fig. 2(d), some normal and abnormal samples in Section 00 are overlapping, but the normal and abnormal samples in Section 01 are easier

TABLE VI
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON MACHINE ANOMALY SCORES (%)

| Algorithm | MobileNetV2 | AE | Our Method |
|---|---|---|---|
| ToyCar | 54.40 | 51.30 | **69.70** |
| ToyTrain | 51.56 | 39.77 | **52.04** |
| Bearing | 60.64 | 54.91 | 58.70 |
| Fan | 57.53 | 58.50 | **76.95** |
| Gearbox | 60.17 | 63.07 | **76.83** |
| Slider | 51.69 | 58.00 | **77.35** |
| Valve | 62.13 | 50.60 | **77.09** |
| Total | 56.58 | 52.71 | **68.29** |



(a) Tgram (Section 00) (b) Tgram (Section 01)

(c) Pre-trained (Section 00) (d) Pre-trained (Section 01)

(e) Tgram + Pre-trained (Section 00) (f) Tgram + Pre-trained (Section 01)

Fig. 2. t-SNE visualization of latent embeddings for Slider's test data. (a) and (b), (c) and (d), (e) and (f) respectively show the visualization results of Section 00 and Section 01 in the test data when Tgram features, pre-trained features, and both are concatenated with log-Mel spectrograms, respectively. Normal samples are marked with "•", and anomalous samples are marked with "×". Orange and green represent the data of Section 00 and Section 01 respectively.

to distinguish. However, from Fig. 2(e) and Fig. 2(f), it can be found that when these two features are concatenated at the same time, both normal and abnormal samples of Section 00 and Section 01 are well distinguished. Therefore, these two features are complementary, and the feature representation obtained when they are concatenated with log-Mel spectrograms at the same time is more discriminative, so the performance of the ASD system in detecting anomalies will be better.

## IV. Conclusions

In this paper, we have proposed an effective self-supervised method for machine anomalous sound detection. First, we obtain a more discriminative feature representation of machine sounds by combining frequency-domain features, time-domain features and pre-trained features extracted from the raw wave. Then, we use the attribute information of machine sounds to design auxiliary loss when training the model to help the ASD system to distinguish and adapt to different environments and working conditions, so that it can detect anomalous sounds more accurately under different working conditions. Finally, we use the distance metric to calculate the machine's anomaly score based on the latent embeddings extracted from the trained self-supervised classifier. The experimental results demonstrate the effectiveness of the proposed method. Our future work includes developing a more efficient model structure for machine anomalous sound detection.

## Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant No. 62171427.

## 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 (DCASE),* 2020, pp. 81–85.

[2] A. Ribeiro, L. M. Matos, P. J. Pereira, E. C. Nunes, A. L. Ferreira, P. Cortez, and A. Pilastri. "Deep dense and convolutional autoencoders for unsupervised anomaly detection in machine condition sounds," *arXiv preprint arXiv:2006.10417*, 2020.

[3] Y. Chen, Y. Song and T. Cheng, "Anomalous Sounds Detection Using A New Type of Autoencoder based on Residual Connection," Tech. Rep., DCASE2020 Challenge, 2020.

[4] N. K. Chaudhary, J. Mathew, and S. Sivadas, "Machine Condition Monitoring from Acoustic Signatures using Auto-Encoders," Tech. Rep., DCASE2020 Challenge, 2020.

[5] P. Daniluk, M. Gozdziewski, S. Kapka, and M. Kosmider, "Ensemble of auto-encoder based systems for anomaly detection," Tech. Rep., DCASE2020 Challenge, 2020.

[6] S. Grollmisch, D. Johnson, J. Abeßer, and H. Lukashevich, "IAEO3-Combining OpenL3 Embeddings and Interpolation Autoencoder for Anomalous Sound Detection," Tech. Rep., DCASE2020 Challenge, 2020.

[7] C. Zhang, Y. Yao, Y. Zhou, G. Fu, S. Li, G. Tang, and X. Shao, "Unsupervised detection of anomalous sounds based on dictionary learning and autoencoder," Tech. Rep., DCASE2020 Challenge, 2020.

[8] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous Sound Detection Based on Interpolation Deep Neural Network," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2020, pp. 271-275.

[9] K. Dohi, T. Endo, H. Purohit, R. Tanabe, and Y. Kawaguchi, "Flow-Based Self-Supervised Density Estimation for Anomalous Sound Detection," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2021, pp. 336-340.

[10] E. Rushe and B. M. Namee, "Anomaly Detection in Raw Audio Using Deep Autoregressive Networks," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2019, pp. 3597-3601.

[11] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions," *arXiv preprint arXiv: 2106.04492,* 2021.

[12] P. Primus, "Reframing unsupervised machine condition monitoring as a supervised classification task with outlierexposed classifiers," Tech. Rep., DCASE2020 Challenge, 2020.

[13] T. Inoue, P. Vinayavekhin, S. Morikuni, S. Wang, T. H. Trong, D. Wood, M. Tatsubori, and R. Tachibana, "Detection of anomalous sounds for machine condition monitoring using classification confidence," Tech. Rep., DCASE2020 Challenge, 2020.

[14] Q. Zhou, "ArcFace based Sound MobileNets for DCASE 2020 Task 2," Tech. Rep., DCASE2020 Challenge, 2020.

[15] J. A. Lopez, H. Lu, P. Lopez-Meyer, L. Nachman, G. Stemmer, and J. Huang, "A Speaker Recognition Approach To Anomaly Detection," Tech. Rep., DCASE2020 Challenge, 2020.

[16] J. Bai, Z. Wang, M. Wang and J. Chen, "DPTRANS: Dual-Path Transformer for Machine Condition Monitoring," Tech. Rep., DCASE2021 Challenge, 2021.

[17] Y. Deng, J. Liu, J. Ma, X. Chen, C. Lu, R. Xu, and W. Q. Zhang, "AITHU system for unsupervised anomalous sound detection," Tech. Rep., DCASE2021 Challenge, 2021.

[18] Y. Sakamoto and N. Miyamoto, "Combine mahalanobis distance, interpolation auto encoder and classification approach for anomaly detection," Tech. Rep., DCASE2021 Challenge, 2021.

[19] S. Kapka, "ID-conditioned auto-encoder for unsupervised anomaly detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE),* 2020, pp. 71–75.

[20] I. Kuroyanagi, T. Hayashi, Y. Adachi, T. Yoshimura, K. Takeda, and T. Toda, "Anomalous sound detection with ensemble of autoencoder and binary classification approaches," Tech. Rep., DCASE2021 Challenge, 2021.

[21] J. Tozicka, D. Karel, and L. Michal, "Unsupervised Anomalous Sound Detection by Siamese Network and Auto-Encoder," Tech. Rep., DCASE2021 Challenge, 2021.

[22] K. Wilkinghoff, "Utilizing Sub-Cluster AdaCos for Anomalous Sound Detection under Domain Shifted Conditions," Tech. Rep., DCASE2021 Challenge, 2021.

[23] Y. Liu, J. Guan, Q. Zhu and W. Wang, "Anomalous Sound Detection Using Spectral-Temporal Information Fusion," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2022, pp. 816-820.

[24] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems* 33 (2020): 12449-12460.

[25] 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.

[26] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737,* 2017.

[27] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "Toyadmos2: Another dataset of miniaturemachine operating sounds for anomalous sound detection under domain shift conditions," in *Proc. 6th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE),* 2021, pp. 1–5.

[28] 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," *arXiv preprint arXiv:2205.13879,* 2022.

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