Transformer-based Autoencoder with ID Constraint for Unsupervised Anomalous Sound Detection

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

[1*]College of Computer Science and Technology, Harbin Engineering University, Harbin, 150001, Heilongjiang, China.
[2]Bytedance, Shanghai, China.
[3]Centre for Audio, Acoustics and Vibration, University of Technology Sydney, Ultimo, NSW 2007, NSW, Australia.
[4]Centre for Vision, Speech and Signal Processing, University of Surrey, University of Surrey, GU2 7XH, U.K.

*Corresponding author(s). E-mail(s): j.guan@hrbeu.edu.cn;
Contributing authors: liuyoude@hrbeu.edu.cn; qiuqiangkong@gmail.com; xiaofeiyang128@gmail.com; qiaoxi.zhu@gmail.com; tianjiantong2022@hrbeu.edu.cn; w.wang@surrey.ac.uk;

Abstract

Unsupervised anomalous sound detection (ASD) aims to detect unknown anomalous sounds of devices when only normal sound data is available. The autoencoder (AE) and self-supervised learning based methods are two mainstream methods. However, the AE-based methods could be limited as the feature learned from normal sounds can also fit with anomalous sounds, reducing the ability of the model in detecting anomalies from sound. The self-supervised methods are not always stable and perform differently, even for machines of the same type. In addition, the anomalous sound may be short-lived, making it even harder to distinguish from normal sound. This paper proposes an ID constrained Transformer-based autoencoder (IDC-TransAE) architecture with weighted anomaly score computation for unsupervised ASD. Machine ID is employed to constrain the latent space of the Transformer-based autoencoder (TransAE) by introducing a simple ID classifier to learn the difference in the distribution for the same machine type

and enhance the ability of the model in distinguishing anomalous sound. Moreover, weighted anomaly score computation is introduced to highlight the anomaly scores of anomalous events that only appear for a short time. Experiments performed on DCASE 2020 Challenge Task2 development dataset demonstrate the effectiveness and superiority of our proposed method.

Keywords: Anomalous sound detection, autoencoder, ID classifier, weighted anomaly score computation

1 Introduction

Anomalous sound detection (ASD) aims to detect anomalies from acoustic signals. Since anomalous sounds can indicate system error or malicious activities, ASD has received much attention [1–5], which has been widely used in various applications, such as road surveillance [6, 7], animal disease detection [8], and industrial equipment predictive maintenance [9]. Recently, ASD has also been used to monitor the abnormality of industrial machinery equipment, such as anomaly detection for surface-mounted device machine [10, 11], and the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge Task2 from 2020 to 2023 [12–15], to reduce the loss caused by machine damage and the cost of manual inspection.

Supervised learning based methods usually train a binary classifier to detect the anomaly [7, 16]. However, it is hard to collect enough anomalous data for supervised learning, as actual anomalous sounds rarely occur in real scenarios. In addition, the high diversity of the anomalies can reduce the robustness of supervised methods. Therefore, unsupervised methods are often employed to detect unknown anomalous sounds without using anomalous sound samples.

In unsupervised ASD, a method is to employ the autoencoder (AE) to learn the distributions of sound signals and perform anomaly detection. Conventional AE-based approaches adopt autoencoder to reconstruct multiple frames of spectrogram to learn the distribution of normal sounds, and then the reconstruction error is used to obtain the anomaly score for anomaly detection [10, 12, 17–19]. However, the conventional AE-based methods do not work well for non-stationary ASD [20], as non-stationary normal sounds (e.g., sound signals of valves) can easily have larger reconstruction errors than abnormal sounds, thus deteriorating the detection performance. In [20], an interpolation deep neural network (IDNN) method is proposed, which masks the center frame of the input, and only uses the reconstruction error of the masked center frame to improve non-stationary sound reconstruction, without considering the edge frames. While the method in [21] adopts a similar strategy as IDNN, and applies the local area mask on the input and employs attentive neural process (ANP) [22] for the reconstruction of the masked input.

Instead of reconstructing spectrogram feature, the method in [23] mixes multiple features as the input, and adopts a fully connected U-Net for the mixed feature reconstruction. To utilize the intra-frame statistics of sound signal, a novel group masked autoencoder for distribution estimation (Group MADE) is proposed for unsupervised

ASD [24, 25], which estimates the density of an audio time series and achieves better performance. However, the distributions of normal audio clips from different machines are different even for the same sound class. This difference can be even greater than that between normal and anomalous sound, which makes it harder to distinguish normal and anomalous sounds for these purely AE-based methods, as the learned feature from these normal sounds may also fit with the anomalous sounds [26].

Machine identity (ID) has been used as the additional condition for encoding in the latent feature space of AE, in order to allow the decoder to provide different reconstructions for each machine [27, 28]. However, the encoder is unable to learn the difference in distributions for different machines, and as a result, the anomalous sound may be well reconstructed. For this reason, it could still be difficult to distinguish normal and anomalous sound. In addition, the above mentioned AE-based methods often use averaged anomaly score for detection, which does not take into account the short-lived condition in anomalous sound, resulting in low anomaly scores for anomalous events that appear only for a short time, which makes it even more challenging for the AE-based methods.

Therefore, instead of reconstructing normal sounds to learn the feature representation, the self-supervised methods are presented to learn the feature representation by utilizing the difference in distributions among different machines [29–36]. The study in [29] uses machine type and machine ID in addition to the machine condition (normal/abnormal) as training labels for self-supervised classification. The flow-based self-supervised method [37] adopts normalizing flow (NF) [38, 39] models, such as generative flow (Glow) [40] and masked autoregressive flow (MAF) [41], to obtain the likelihood estimation for anomaly detection. In this method, an auxiliary task is introduced to distinguish the sound data of that machine ID (i.e., target data) from the sound data of other machine IDs with the same machine type (i.e., outlier data). Moreover, although the self-supervised learning based methods can achieve better performance than the AE-based methods, they are not always stable and could perform differently even for the machines of the same type.

In this paper, we present an ID constrained Transformer-based autoencoder (IDC-TransAE) architecture with weighted anomaly score computation for unsupervised ASD. Our method includes two stages, namely, spectrogram reconstruction and anomaly detection. First, an IDC-TransAE is introduced to reconstruct the spectrogram of normal sounds, where Transformer [42] is employed to build the AE architecture, and a simple ID classifier is incorporated into the AE. Specifically, the Transformer captures the time-dependent information of the sound signal, and the classifier utilizes machine ID to constrain the latent space of AE, so that our proposed IDC-TransAE can learn different distributions of normal machines, even with the same type. In the proposed IDC-TransAE architecture, instead of using the positional encoding (PE) for Transformer to provide additional temporal information, a linear phase embedding (LPE) method is proposed to represent the temporal information of sound signal by using its phase information, which can further enhance the classification performance of the proposed IDC-TransAE. In addition, the center frame prediction (CFP) is also employed in our IDC-TransAE to improve the ASD ability for non-stationary signals (e.g., Valve). Then, the reconstruction error from the trained

IDC-TransAE can be used to calculate the anomaly score to detect the anomaly. Here, we introduce a weighted anomaly score computation method via global weighted ranking pooling (GWRP) [43], which can highlight the anomaly scores for the anomalous events that only appear for a short time. Finally, we obtain the final anomaly score with the combination of the classification anomaly score and weighted reconstruction anomaly score, to obtain more stable and consistent detection performance.

In summary, the innovations and contributions of this paper for unsupervised anomalous sound detection can be summarized as follows:

1. We analyze the generalization problem of AE for ASD and point out the main reason for this problem, and propose a solution, i.e., IDC-TransAE, to mitigate the generalization of AE and improve the detection performance. To the best of our knowledge, this is the first work to clearly point out the main reason for the generalization problem of AE for ASD.
2. We propose an ID constraint (IDC) classifier to learn different audio feature distributions from the same machine type, which can enhance the distinguishing ability for anomaly detection.
3. We design a linear phase embedding (LPE) to replace the traditional positional encoding (PE) to preserve the own temporal information of machine sounds by the phase of sounds.
4. In the anomaly score calculation, we introduce the global weighted ranking pooling (GWRP) to highlight the anomaly score of sounds with short-time non-stationary anomalies, which obtains a more stable and consistent detection performance.
5. Experimental results verify that the proposed IDC-TransAE method can mitigate the generalization problem of AE for ASD. Ablation studies and visualizations further verify the effectiveness of the design of ID constraint, LPE and GWRP for ASD. Our study employs the DCASE 2020 Challenge Task2 dataset to address AE's generalization problem in ASD, excluding DCASE 2022 and 2023 datasets tailored for domain-shift and first-shot scenarios beyond our paper's scope.

2 Preliminary

The AE-based methods are widely used for unsupervised ASD [10, 12, 17, 18] An AE model is trained with normal sounds to learn their feature distribution. It implicitly assumes that it can reconstruct normal sounds better than anomalous sounds, so that anomalous sounds often have larger reconstruction errors than normal sound. The reconstruction error is then used for deriving the anomaly scores for anomaly detection. Figure 1 shows the AE architecture for unsupervised ASD.

Regarding model training, multiple frames of a spectrogram are usually used as the input, and the same number of frames are generated as the output. Suppose $\boldsymbol{X} \in \mathbb{R}^{N \times M}$ is the log-Mel spectrogram of the sound signal, where N is the number of frames and M is the feature dimension of each frame of $\boldsymbol{X}$. The loss for the AE model training is

$$L_{\mathrm{AE}} = \|\boldsymbol{X} - D(E(\boldsymbol{X}))\|_2^2 , \tag{1}$$

where $E(\cdot)$ and $D(\cdot)$ are the encoder and the decoder of AE, respectively.

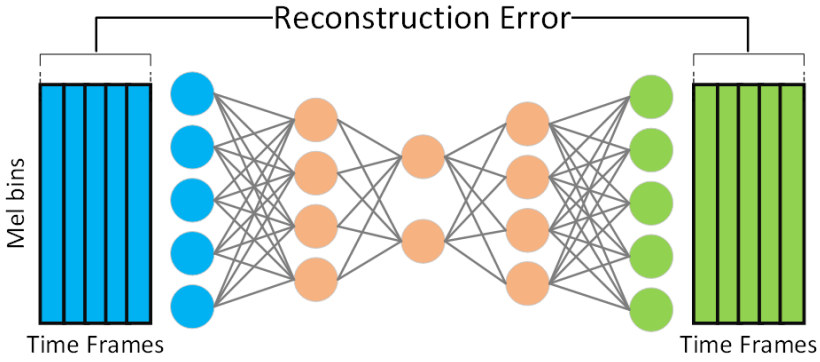


Fig. 1: Typical architecture of AE for unsupervised ASD uses the reconstruction error between the input and output as the anomaly score.

Then, the trained AE model can be used to detect the anomaly. $\boldsymbol{Y}$ is the test audio clip, split into I segments, $\{\boldsymbol{Y}_i\}_{i=1}^{I}$. Here $\boldsymbol{Y}_i \in \mathbb{R}^{N\times M}$ is the i-th segment and also the i-th input of the model. The reconstruction error e_i for $\boldsymbol{Y}_i$ is

$$e_i = \frac{1}{NM}\left\|(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}}_i)\right\|_F^2, \tag{2}$$

where $\overline{\boldsymbol{Y}}_i = D(E(\boldsymbol{Y}_i))$ is the corresponding output frames, and $\|\cdot\|_F$ denotes Frobenius norm. It results in a reconstruction error sequence $\boldsymbol{e} = \{e_i\}_{i=1}^{I}$ for $\boldsymbol{Y}$, and the mean reconstruction error of $\boldsymbol{e}$ can be used as the anomaly score

$$\mathcal{A}(\boldsymbol{e})_{mean} = \frac{1}{I}\sum_{i=1}^{I} e_i. \tag{3}$$

Here $\mathcal{A}(\boldsymbol{e})_{mean}$ represents the anomalous degree of the audio clip. The normal or anomaly of the clip is determined by $\mathcal{H}(\boldsymbol{e},\theta)$ [44]:

$$\mathcal{H}(\boldsymbol{e},\theta) = \begin{cases} 0\ (Normal) & \mathcal{A}(\boldsymbol{e})_{mean} \le \theta \\ 1\ (Anomaly) & \mathcal{A}(\boldsymbol{e})_{mean} > \theta \end{cases}, \tag{4}$$

where θ is a pre-defined threshold value to determine whether an audio clip is anomalous.

However, for normal non-stationary sounds, the AE-based methods tend to give large reconstruction errors for both normal and abnormal sounds, this is because the edge frames of non-stationary sound are hard to reconstruct. In [20], IDNN is proposed for non-stationary sound ASD, which removes the center frame of the multiple frames as the input, and predicts the removed frame as the output, as shown in Figure 2. The input multiple frames of IDNN can be expressed as $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{\frac{N+1}{2}-1}, \boldsymbol{x}_{\frac{N+1}{2}+1}, \cdots, \boldsymbol{x}_N]^T$, and T denotes transposition. The loss function

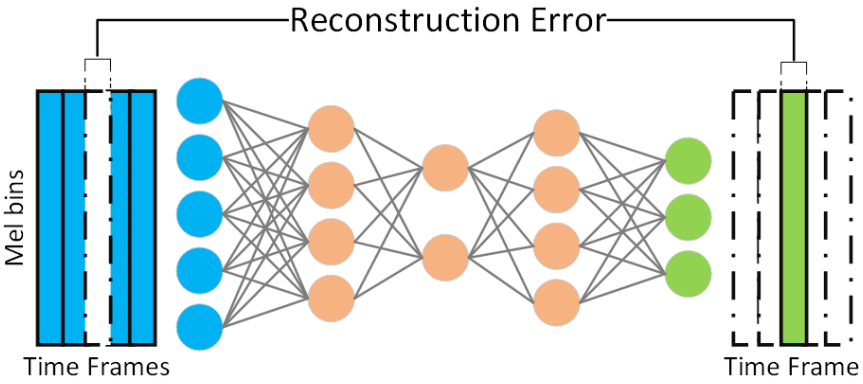


Fig. 2: The architecture of IDNN uses the reconstruction error of center frame as the anomaly score.

of IDNN is formulated as

$$L_{\text{IDNN}} = \left\| \boldsymbol{x}_{\frac{N+1}{2}} - D(E(\boldsymbol{X})) \right\|_2^2, \tag{5}$$

where $\boldsymbol{x}_{\frac{N+1}{2}}$ is the removed center frame of original input frames. Unlike conventional AE-based methods, the reconstruction error e_i of the i-th input is only calculated by the center frame.

However, the training procedure does not involve the anomalous sound, as a result, the AE-based methods could be limited in the scenario where the learned feature also fits with the anomalous sound [2]. In this case, the anomalous sound could be well reconstructed with a smaller reconstruction error than that of the normal sounds of different machines, even of the same type. For example, the anomalous sounds from one machine may be similar to the normal sounds of another machine, due to different usage of different machines. In this case, the AE trained with these different machines of the same machine type can reconstruct the anomalous sounds well, and thus it may not be able to detect these anomalous sounds.

In addition, for anomalous events that only appear for a short time in audio clips, the anomaly score calculated by mean reconstruction error is often too small, making it difficult to detect the anomaly.

3 Proposed Method

This section presents our IDC-TransAE with weighted anomaly score computation for unsupervised ASD. We introduce IDC-TransAE to reconstruct the spectrogram of normal sounds to learn their distributions, and apply GWRP for weighted anomaly score computation to perform anomaly detection.

3.1 ID Constraint Transformer Autoencoder

We utilize Transformer to exploit temporal information for better reconstruction of normal sounds, where only the encoder layer of Transformer is employed to build the encoder and decoder of our IDC-TransAE architecture. In addition, machine ID is adopted to constrain the latent space of the AE by introducing a simple ID classifier to learn different representations for different normal sounds. The framework of the proposed IDC-TransAE is illustrated in Figure 3.

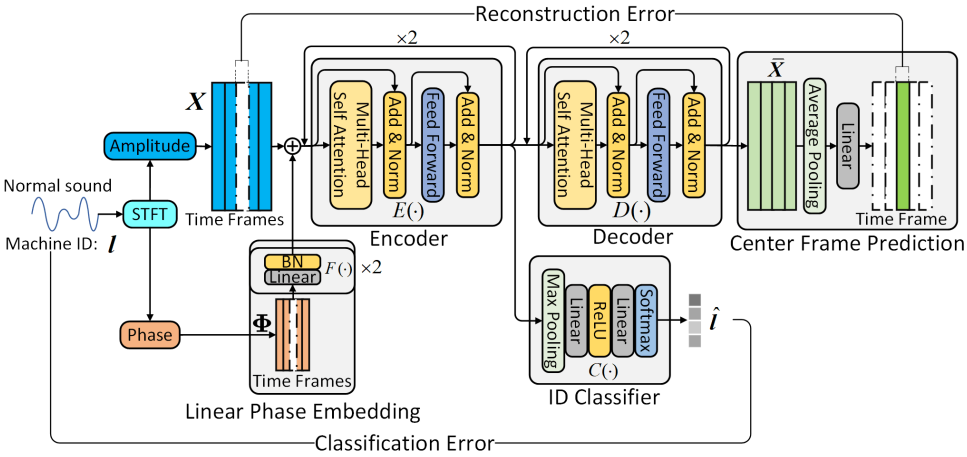


Fig. 3: The architecture of the proposed IDC-TransAE for normal sound reconstruction. $\boldsymbol{X}$ and $\boldsymbol{\Phi}$ are the inputs to the model, which are obtained from the sound signal by removing the center frame. The final predicted center frame is obtained by average pooling of the output of the decoder $\overline{\boldsymbol{X}}$ in frames and a linear layer, and $\hat{\boldsymbol{l}}$ is the predicted machine ID probability of sound signal, which is obtained by max pooling of output of encoder $\boldsymbol{z}$ in frames and two linear layers with softmax. IDC-TransAE is optimized by the combination of reconstruction error and classification error.

3.1.1 Center Frame Prediction

For better reconstruction of the spectrogram of normal non-stationary sound, following IDNN, we introduce a center frame prediction (CFP) method by removing the center frame of input frames and predicting the removed frame. After removing center frame $\boldsymbol{x}_{\frac{N+1}{2}}$, the input frames can be expressed as

$$\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{\frac{N+1}{2}-1}, \boldsymbol{x}_{\frac{N+1}{2}+1}, \cdots, \boldsymbol{x}_N]^T, \tag{6}$$

where T denotes the matrix transposition operation. Unlike IDNN, the predicted center frame obtained by CFP is the average pooling of decoder output in frames and then processed by a linear layer, as shown in Figure 3.

3.1.2 Linear Phase Embedding

To represent the appropriate positional relationship of the sound signal, we propose a linear phase embedding (LPE) method for IDC-TransAE, to replace positional encoding (PE), often used in Transformer to provide additional position information via sinusoid function [42], which is, however, not strongly correlated with the sound signal. In contrast, LPE in the proposed method preserves the signal's temporal information by linearly embedding the phase angles of the signal to the same dimensions with the input $\boldsymbol{X}$. The phase angle is obtained via the short-time Fourier transform (STFT).

Assuming the phase angles corresponding to $\boldsymbol{X}$ are denoted as $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_{\frac{N+1}{2}-1}, \boldsymbol{\phi}_{\frac{N+1}{2}+1}, \cdots, \boldsymbol{\phi}_N]^T$, with center frame removed, and $F(\cdot)$ is the linear embedding function, including two linear layers with batch normalization. The output $\overline{\boldsymbol{X}}$ of decoder $D(\cdot)$ can be obtained as

$$\overline{\boldsymbol{X}} = D(E(\boldsymbol{X} + F(\boldsymbol{\Phi}))), \tag{7}$$

where $\overline{\boldsymbol{X}} = [\overline{\boldsymbol{x}}_1, \cdots, \overline{\boldsymbol{x}}_n, \cdots, \overline{\boldsymbol{x}}_{N-1}]^T$. Then, the average pooling of $\overline{\boldsymbol{X}}$ is used to predict the center frame, and the reconstruction loss for the center frame is

$$L_r = \left\| \boldsymbol{x}_{\frac{N+1}{2}} - (W_o \frac{1}{N-1} \sum_{n=1}^{N-1} \overline{\boldsymbol{x}}_n + b_o) \right\|_2^2, \tag{8}$$

where W_o and b_o are the learnable parameters of the last output linear layer. The LPE module helps preserve the temporal information of the signal to enhance the ability of the model for anomalous sound detection.

3.1.3 ID Classifier

We observed that the performance of the trained AE model on different machines with the same type could be quite different. The potential cause is the difference in distributions of normal machine sound when individual machines have different usages. However, the trained model only learns how to reconstruct the general distribution of different normal machines sounds.

To enable the model to learn different representations for different machine sounds even with the same type, we introduce an ID classifier $C(\cdot)$ with machine ID information to constrain the latent feature $\boldsymbol{z}$ of AE. The structure of the ID classifier $C(\cdot)$ is given in Figure 3, which consists of a max pooling layer, two linear layers with a ReLU [45] function and a softmax activation function.

Here, the latent feature $\boldsymbol{z}$ is the output of the encoder of AE, which is the input of the classifier, defined as $\boldsymbol{z} = E(\boldsymbol{X} + F(\boldsymbol{\Phi}))$. The output of the ID classifier $\hat{\boldsymbol{l}} = C(\boldsymbol{z}) \in \mathbb{R}^K$ is the probability indicating normal/anomalous sound corresponding to the machine ID, and K is the number of machines with the same type. Then, the classification error of $C(\cdot)$ can be obtained via a cross-entropy loss function [46]

$$L_c = CrossEntropy(\boldsymbol{l}, \hat{\boldsymbol{l}}), \tag{9}$$

where $\boldsymbol{l} \in \mathbb{R}^K$ is the one-hot vector of machine ID label of the sound signal.

Therefore, the proposed IDC-TransAE can be jointly trained by minimizing the center frame reconstruction error and the machine ID classification error with the joint loss function

$$L_{total} = (1-\alpha)L_r + \alpha L_c, \tag{10}$$

where $\alpha \in [0, 1)$ is a hyper-parameter. The magnitude of α denotes the extent to which the machine ID classifier restricts $\boldsymbol{z}$. By jointly training the AE with the ID classifier, we can improve anomaly detection performance.

3.2 Weighted Anomaly Score Computation

For anomaly detection, the formula $\mathcal{A}(\boldsymbol{e})_{mean}$ in Equation (3) usually underestimates the anomaly scores of anomalous audio clips when the anomalous events only appear for a short time. One solution is to use the maximal reconstruction error as the anomaly score i.e., max anomaly score $\mathcal{A}(\boldsymbol{e})_{max} = max(\boldsymbol{e})$, to highlight the anomalies of these audio clips. However, it is not robust to use the maximum value of $\boldsymbol{e}$ as the anomaly score of the whole audio clip, as it may overestimate the anomaly scores of some normal audio clips.

To improve the reliability of the calculated anomaly score, we employ the global weighted rank pooling (GWRP) method to obtain weighted anomaly score, where GWRP is a generalization of max and mean, which can highlight the anomaly score by setting different weights to reconstruction error sequence $\boldsymbol{e}$. For example, let $\hat{\boldsymbol{e}} = \{\hat{e}_1, ..., \hat{e}_I\}$ be sorted by descending order of $\boldsymbol{e}$, the GWRP anomaly score can be calculated as

$$\mathcal{A}(\hat{\boldsymbol{e}})_{gwrp} = \frac{1}{Z(r)} \sum_{i=1}^{I} r^{i-1} \hat{e}_i, \tag{11}$$

where $0 \leq r \leq 1$ is a hyper-parameter and $Z(r) = \sum_{i=1}^{I} r^{i-1}$ is a normalization term. When $r = 0$, $\mathcal{A}(\hat{\boldsymbol{e}})_{gwrp}$ degenerates to $\mathcal{A}(\boldsymbol{e})_{max}$, and when $r = 1$, $\mathcal{A}(\hat{\boldsymbol{e}})_{gwrp}$ becomes $\mathcal{A}(\boldsymbol{e})_{mean}$. It intends to assign larger weights to anomalous audio clips and lower weights to normal audio clips, to generate high anomaly scores for the anomalous events of short duration. In addition, the classification error is combined with the reconstruction error to calculate the anomaly score, to allow the anomaly score to increase if the ID classifier misclassifies the machine ID. Finally, the weighted anomaly score can be calculated as

$$\mathcal{A}(\hat{\boldsymbol{e}}, \boldsymbol{l}, \hat{\boldsymbol{l}}) = (1-\beta)\mathcal{A}(\hat{\boldsymbol{e}})_{gwrp} + \beta L_c, \tag{12}$$

where $\beta \in [0, 1]$ is a parameter weighting the impact of a false prediction by the ID classifier on the anomaly score. For clarity, the proposed IDC-TransAE with weighted anomaly score computation is denoted as IDC-TransAE-W in the following section.

4 Experiments and Results

4.1 Experimental Setup

4.1.1 Dataset

We evaluate our method on the DCASE 2020 Challenge Task2 [12] dataset, which comprises parts of MIMII [47] and ToyADMOS dataset [48] including the normal/anomalous operating sounds of six types of real/toy machines. The MIMII dataset includes four types of machines (i.e., Fan, Pump, Slider and Valve), with four different machines for each machine type. The ToyADMOS dataset consists of two types of machines (i.e., ToyCar and ToyConveyor), with four and three different machines for each type, respectively. Each recording is a single-channel audio of 10-sec long with a 16kHz sampling rate that includes both a target machine's operating sound and environmental noise. Following [12], the training set only includes normal sounds, with around 6000 items for each machine type, and the test set consists of both normal and anomalous sounds, including about 500 to 1000 items for normal and anomaly in each machine type.

4.1.2 Performance Metrics

Following [12, 20, 29, 37, 49], we employ area under the receiver operating characteristic (ROC) curve (AUC) and the partial-AUC (pAUC) as the performance metrics, where the pAUC is calculated as the AUC over a low false-positive-rate (FPR) range $[0, p]$ and $p = 0.1$ as in [12]. Higher AUC indicates better model performance. pAUC reflects the reliability of the ASD system based on practical requirements. It is important to increase pAUC to avoid the ASD system predicting false alerts frequently [12]. In addition, the minimum AUC (mAUC) is adopted to represent the worst detection performance achieved among individual machines of same machine type, following [37].

4.1.3 Implementation Details

Table 1: Implementation details for all machine types.

	Fan	Pump	Slider	Valve	ToyCar	ToyConveyor
n_FFT				1024		
n_Mels				128		
hop length				512		
frames				5		
α				0.3		
r	1.00	1.00	0.96	0.92	1.00	1.00
β	0.84	0.82	0.80	0.72	0.62	0.98

The implementation details of IDC-TransAE can be seen in Table 1. We use the log-Mel spectrogram and phase angle of the sound signal as the input of our IDC-TransAE. The frame size is 1024 with an overlapping 50%, i.e., the number of FFT bins (n_FFT) is 1024, and the hop length is 512. The number of Mel filter banks (n_Mels) is set as 128. The number of frames (i.e., N) is 5. The dimension of phase angles is 513, which is embedded to a 128-dimensional vector by the linear function

$F(\cdot)$. Here, $F(\cdot)$ consists of two linear layers with batch normalization. The encoder and decoder of IDC-TransAE include two layers, respectively. The classifier includes a max pooling layer, two linear layers with a ReLU and a softmax activation function. The hyper-parameter α of the joint loss function is empirically set as 0.3.

Adam optimizer [50] is used to optimize our model with a learning rate of 0.0001. For each machine type, our model is trained 300 epochs, and the batch size is set as 2000. In the joint training stage, we found that the classification loss converges much faster than the reconstruction loss, so we adopt a training strategy to avoid the overfitting of the classifier, by training the classifier every 10 epochs (i.e., using L_{total} loss) and the remaining epochs for autoencoder (i.e., using L_r loss). In weighted anomaly score computation, r and β are empirically selected, and the values of r and β are provided in Table 1.

4.2 Experimental Results and Performance Analysis

4.2.1 Comparison with Other Methods

To demonstrate the performance of our method for unsupervised ASD, we compare our approach with the AE baseline of DCASE 2020 Challenge Task2 [12] and mainstream models, including AE-based methods (i.e., IDNN [20], ANP-Boot [49], Group MADE [24] and IDCAE [27]) and self-supervised based methods (i.e., MobileNetV2 [29] and Glow_Aff [37]), where IDCAE, MobileNetV2 and Glow_Aff employ the ID information for anomalous sound detection.

Table 2: Performance comparison in terms of AUC (%) and pAUC (%) for different types of machines.

	Fan		Pump		Slider		Valve		ToyCar		ToyConveyor		Average	
	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
w/o ID information														
AE baseline [12]	65.91	51.93	70.20	61.69	83.42	65.72	67.78	51.67	78.77	67.58	72.53	**60.43**	73.10	59.84
IDNN [20]	65.94	52.48	74.26	62.20	84.34	65.48	83.70	62.02	77.42	62.64	69.36	58.58	75.67	60.57
ANP-Boot [21]	64.80	53.00	65.50	59.00	**94.90**	83.10	85.20	72.00	72.90	68.10	67.10	54.20	75.07	64.90
Group MADE [24]	68.00	53.10	74.10	66.20	94.40	**83.70**	95.60	85.50	79.50	68.40	**74.70**	60.30	81.05	69.53
TransAE-mean	73.91	54.14	77.31	68.96	91.51	74.66	96.09	84.65	80.62	72.65	74.32	59.80	82.29	69.14
TransAE-W	**73.91**	**54.14**	**77.31**	**68.96**	94.52	82.33	**99.68**	**98.31**	**80.62**	**72.65**	74.32	59.80	**83.39**	**72.70**
w/ ID information														
MobileNetV2 [29]	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 [37]	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
IDCAE [27]	77.45	70.32	77.29	70.33	80.04	68.25	78.26	55.80	78.07	74.22	70.29	59.46	76.90	66.40
IDC-TransAE-mean	80.44	70.21	83.41	79.24	92.17	77.10	94.04	78.94	93.17	87.43	75.69	62.96	86.49	75.98
IDC-TransAE-W	**80.44**	70.21	**83.41**	**79.24**	**96.20**	**86.38**	**99.60**	**98.29**	**93.40**	**87.43**	**75.69**	**62.96**	**88.12**	**80.94**

Table 2 shows the comparison results in terms of AUC and pAUC. Here, IDC-TransAE-W and IDC-TransAE-mean represent IDC-TransAE with weighted anomaly score computation and mean anomaly score computation, respectively. In addition, the proposed methods without using ID information are evaluated, i.e., TransAE-W and TransAE-mean.

As shown in Table 2, the methods with ID information (denoted as w/ ID) give better detection performance than the methods without ID information (denoted

as w/o ID), except IDCAE. The proposed IDC-TransAE-W performs the best in terms of average AUC and pAUC. Amongst the methods without using ID information, our TransAE-W also achieves the best overall performance. Especially, both TransAE-W and IDC-TransAE-W can substantially improve the performance on the non-stationary sound signal of Valve (i.e., with 13.66% and 19.35% pAUC improvements compared to TransAE-mean and IDC-TransAE-mean, respectively), which demonstrates the effectiveness of the weighted anomaly score computation for anomalous events appearing for a short time. In addition, the significantly improved average pAUC (i.e., 80.94%) shows that the proposed IDC-TransAE-W is more reliable than other methods.

Note that, r in the weighted anomaly score computation can be adjusted according to the time length of the anomalous event, for example, when $r = 1$, $A(\hat{\boldsymbol{e}})_{gwrp} = A(\boldsymbol{e})_{mean}$. This means it is more applicable than mean anomaly score. The influence of r will be discussed in Section 4.4.

4.2.2 Detection Stability

To demonstrate the effectiveness of our method for more stable detection, another experiment is conducted to show the worst detection performance on individual machines of the same type, where the self-supervised based methods (i.e., MobileNetV2 and Glow_Aff) and the typical AE-based method (i.e., IDNN) are employed for comparison. The results in terms of mAUC are given in Table 3.

Table 3: Performance comparison in terms mAUC (%) among the individual machines of the same type.

	MobileNetV2[29]	Glow_Aff[37]	IDNN[20]	**IDC-TransAE-W**
Fan	50.40	49.60	**56.56**	50.55
Pump	52.90	**65.70**	61.86	57.27
Slider	82.80	87.80	74.22	**88.64**
Valve	67.90	77.70	66.83	**99.24**
ToyCar	55.70	80.10	64.41	**81.35**
ToyConveyor	48.70	61.00	**62.89**	62.31
Average	59.73	70.32	64.46	**73.23**

As can be seen from Table 2 and Table 3, the self-supervised methods, i.e., MoblieNetV2 and Glow_Aff, can achieve significant improvements in average AUC and pAUC, as compared to the AE-based method IDNN. However, they perform dramatically different even for the machines of the same type, as observed from Table 3 and Table 2, e.g., MobileNetV2 has much smaller mAUC than AUC on Fan, Pump, ToyCar and ToyConveyor. The results demonstrate the instability of the self-supervised methods.

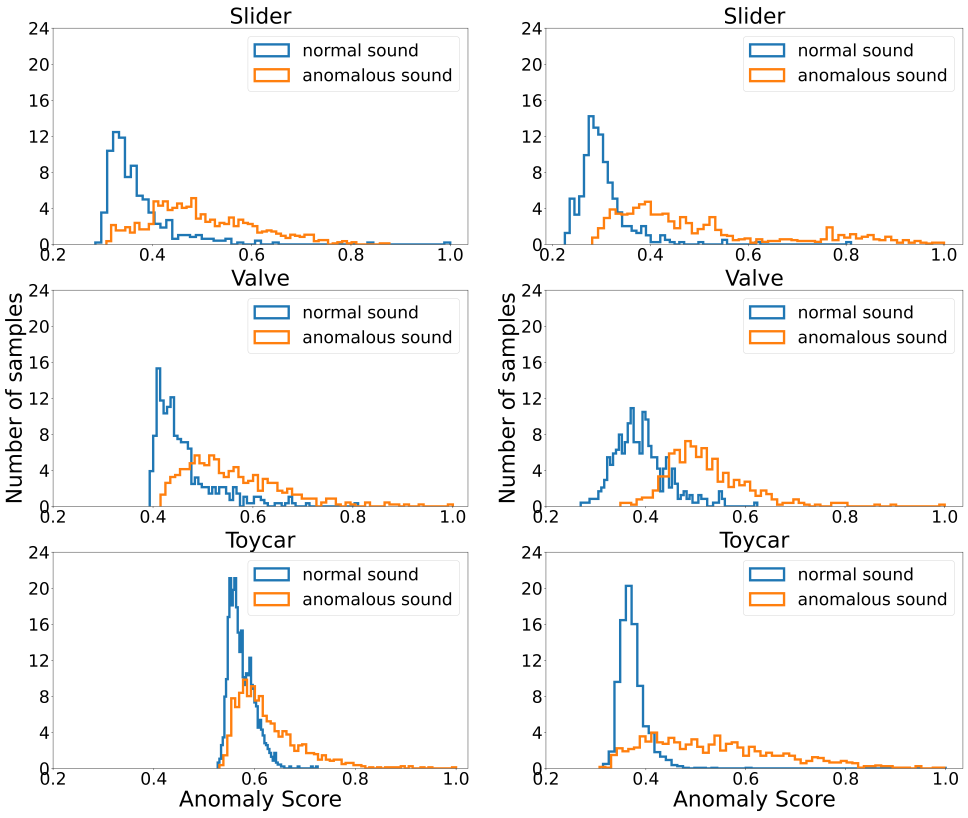
Especially, the average mAUC (i.e., 59.73%) of MobileNetV2 is lower than that of IDNN (i.e., 64.46%), which indicates that the self-supervised classification method (i.e., MobileNetV2) indeed easily fails on some individual machines and lacks performance consistency. In contrast, the AE-based method IDNN can provide a relatively stable detection performance. Although the flow-based self-supervised method

(Glow_Aff) can improve detection stability to some extent compared to the AE-based method, our proposed method can achieve the best average mAUC performance and obtain more stable performance for some machine types, i.e., Slider, Valve, and ToyCar.

Although Glow_Aff has a higher mAUC on Pump than our proposed method, the model needs to be trained for each individual machine which could be limited in real-world applications. In contrast, our proposed method only needs to train one model for each machine type.

4.2.3 Generalization to Anomaly

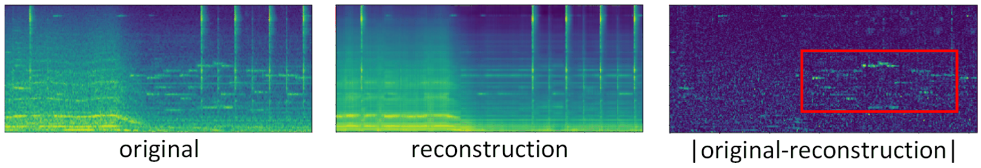
To demonstrate the proposed IDC-TransAE can mitigate the generalization of AE for anomalous sound and improve its detection performance, experiments are conducted to compare it with the typical AE-based method (i.e., IDNN).



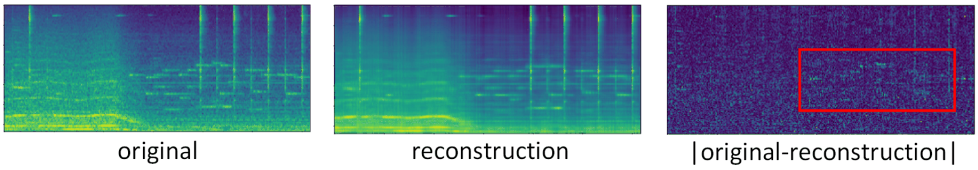
(a) IDNN (b) IDC-TransAE-mean

Fig. 4: The histograms of anomaly scores distribution on Slider, Valve and ToyCar using IDNN and the proposed IDC-TransAE-mean, where blue and orange indicate the anomaly score distribution of normal and anomalous sound, respectively.

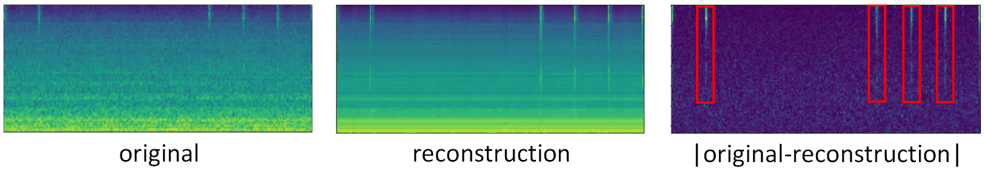
First, we show the histograms of anomaly score distribution on Slider, Valve and ToyCar using IDNN and our proposed IDC-TransAE. For a fair comparison, our method (i.e., IDC-TransAE-mean) also adopts mean anomaly score computation as IDNN, and the results are provided in Figure 4. Here, the anomaly score is on the horizontal axis of the histogram, which is normalized to facilitate comparison. The vertical axis represents the number of audio samples corresponding to the anomaly score distribution on the histogram.



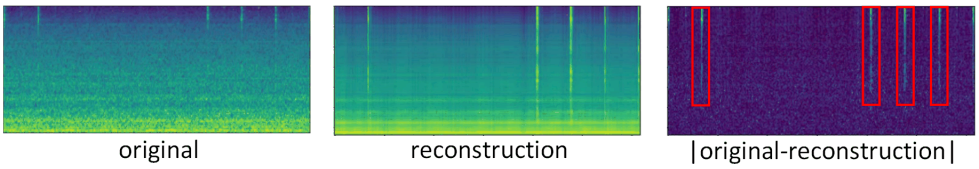
(a) IDNN for normal sound



(b) IDC-TransAE for normal sound



(c) IDNN for anomalous sound



(d) IDC-TransAE for anomalous sound

Fig. 5: The log-Mel spectrogram reconstruction analysis of IDNN and IDC-TransAE on normal and anomalous Valve's sound, the "original", "reconstruction" and "|original-reconstruction|" represent original spectrogram, reconstructed spectrogram and the absolute value of their difference, respectively.

From Figure 4, we can see that, for IDNN, the anomaly score distribution of the anomalous sound tends to be similar to that of the normal sound, especially on ToyCar, as shown in Figure 4(a). It shows that most anomalous sound have a small anomaly score similar to normal sound. This indicates that the AE-based method (i.e., IDNN)

is able to generalize the representation for anomalous sound, which reduces its ability to distinguish between normal and abnormal sound. In contrast, our proposed method can give higher anomaly scores for the anomalous sound, and provide better detection ability than the AE-based method, as shown in the histograms in Figure 4(b), which demonstrates the effectiveness of our proposed IDC-TransAE architecture.

To further demonstrate that our proposed IDC-TransAE can mitigate its generalization for the anomaly, we perform another experiment for non-stationary anomalous sound detection (i.e., sound of Valve) as compared with IDNN, where the log-Mel spectrogram reconstruction of normal and anomalous sound is illustrated in Figure 5. From left to right, Figure 5 shows the original log-Mel spectrograms, the reconstructed log-Mel spectrograms, and the absolute values of their difference.

Comparing the red box areas illustrated in Figure 5(a) and Figure 5(b), the proposed IDC-TransAE can provide better normal sound reconstruction, as it can achieve smaller reconstruction error for normal sound than that of the typical AE-based method (i.e., IDNN). This can be clearly observed in the comparison of the absolute value difference of original log-Mel spectrogram and reconstrcuted log-Mel spectrogram, as the red box indicated areas in Figure 5(a) and Figure 5(b). Whereas for the anomalous sound reconstruction, our proposed method can give larger reconstruction error than the typical AE-based method, which means that our method has a better ability to highlight the anomalies when reconstructing the anomalous sound. This can be observed from the comparison between the red box areas in Figure 5(c) and Figure 5(d), where the absolute value difference shown in Figure 5(d) is much more clear than that in Figure 5(c). The results further demonstrate that our proposed IDC-TransAE can solve the generalization problem of the AE-based method and has a better ability in anomaly detection.

Note that the log-Mel spectrogram of the anomalous sound also shows that the anomalies may appear for a short time in the sound, as illustrated in Figure 5. In this case, the mean anomaly score computation method will give low anomaly scores for the anomalous events that only appear for a short time.

4.3 Ablation Studies

To show the effectiveness of different parts of our proposed IDC-TransAE-W, ablation studies are conducted, where AUC and pAUC are used as performance metric. The results are given in Table 4. Here, TransAE/PE-W denotes the proposed model without using machine ID constraint (IDC) module and CFP module, and adopts PE instead of LPE, with weighted anomaly score computation for anomaly detection. TransAE/PE/CFP-W denotes the TransAE/PE-W using CFP, and TransAE/LPE/CFP-W denotes replacing PE with LPE in Transformer/PE/CFP-W.

As shown in Table 4, TransAE/PE/CFP-W can significantly improve the detection performance for the non-stationary sound signal of Valve, with 12.61% AUC and 30.40% pAUC improvements as compared with TransAE/PE-W. To show the effectiveness of LPE, we compare the performance of TransAE/PE/CFP-W and TransAE/LPE/CFP-W. The result shows that TransAE/LPE/CFP-W can improve the detection performance on Fan, Slider, Valve, ToyConveyor, and achieve better

Table 4: Validation of different modules of IDC-TransAE.

	TransAE/PE-W		TransAE/PE/CFP-W		TransAE/LPE/CFP-W		IDC-TransAE-W	
	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
Fan	70.06	52.73	72.46	52.95	73.91	54.14	**80.44**	**70.21**
Pump	79.76	70.79	77.66	71.30	77.31	68.96	**83.41**	**79.24**
Valve	92.88	81.05	94.12	82.16	94.52	82.33	**96.20**	**86.38**
Slider	85.10	62.91	97.71	93.31	**99.68**	**98.31**	99.60	98.29
ToyCar	82.95	72.47	81.27	72.69	80.62	72.65	**93.40**	**87.43**
ToyConveyor	**76.81**	**63.88**	74.35	58.51	74.32	59.80	75.69	62.96
Average	81.26	67.31	82.93	71.82	83.39	72.70	**88.12**	**80.94**

average AUC and pAUC performance. It indicates that LPE can better represent the temporal information of the sound signal by using its phase information. By introducing the ID classifier, the proposed IDC-TransAE-W with the IDC module can achieve the best overall detection performance, giving more than 10% improvement in pAUC on Fan, Pump, and ToyCar as compared with TransAE/LPE/CFP-W. Besides, we can see that the proposed IDC module contributes the most to the performance improvement in Table 4, which further verifies the effectiveness of the proposed IDC module to enhance the ability of the model in distinguishing anomalous sound.

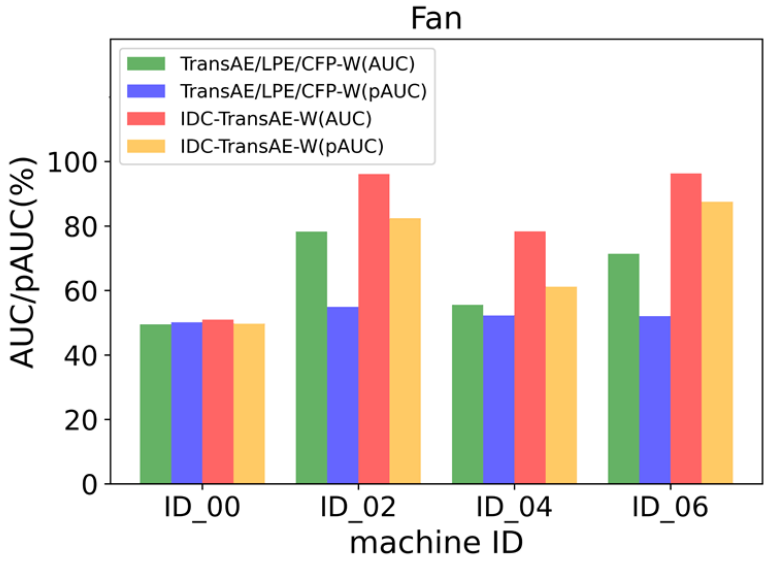


Fig. 6: Performance illustration for 4 different machines with the same type, i.e., Fan.

To further demonstrate the effectiveness of the IDC module, we compare the performance of IDC-TransAE-W and TransAE/LPE/CFP-W in terms of AUC and pAUC on four different machines of the machine type Fan. The result is illustrated in

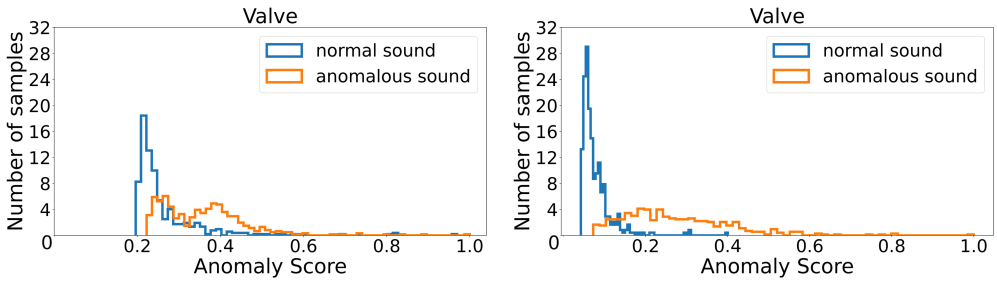
Figure 6. From Figure 6, we can see that IDC-TransAE-W can significantly improve the performance on ID_02, ID_04 and ID_06, as compared with TransAE/LPE/CFP-W. This means the IDC method can better distinguish the anomalous sound for different machines with the same type. The results in Table 4 and Figure 6 verify the effectiveness of different modules of our proposed method. To further illustrate the effectiveness of each module, we give the visualization analysis for each module in the following Section 4.4.

4.4 Visualization Analysis

In this section, visualization analysis is provided for better understanding the experimental results in the ablation studies. Specifically, the effectiveness of CFP, LPE module and IDC module in our proposed IDC-TransAE method are further evaluated. Besides, the influence of the parameter in the GWRP operation of anomaly score calculation is also explored in this section.

4.4.1 Effectiveness of CFP

To show how CFP operation affects the anomaly detection for non-stationary sound signals, we compare the histograms of anomaly score distribution between TransAE/PE-W and TransAE/PE/CFP-W on Valve. The result is given in Figure 7. Same as Figure 4, the anomaly score is also normalized to facilitate comparison. By comparing Figure 7(a) and Figure 7(b), we can see that the distribution of normal sound samples is on a smaller range of anomaly scores when adopting the CFP module (i.e., TransAE/PE/CFP-W), as illustrated in Figure 7(b). It verifies that CFP operation can improve the reconstruction of non-stationary signals as described in [20]. Therefore, it can improve the performance of our proposed method for anomaly detection of non-stationary sound signals.

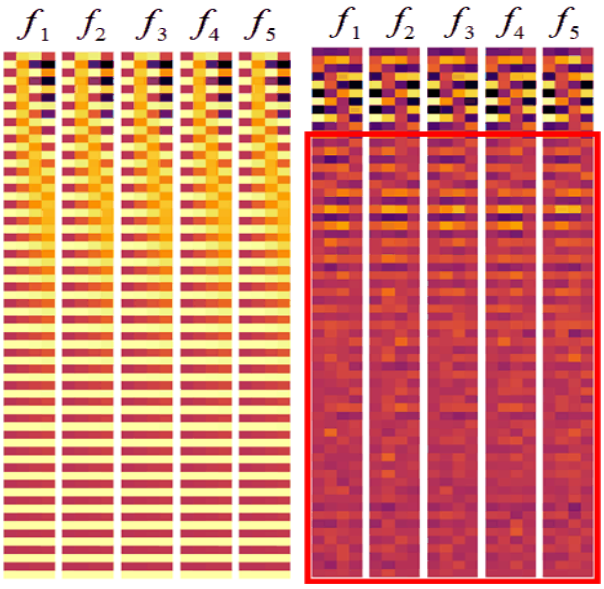


(a) TransAE/PE-W (b) TransAE/PE/CFP-W

Fig. 7: The comparison between TransAE/PE-W and TransAE/PE/CFP-W on histograms of anomaly scores distribution of Valve.

4.4.2 Visualization of Linear Phase Embedding

To show why the LPE module can enhance the ability of the model for anomalous sound detection, we visualize the encoding result of five consecutive input sound signals of TransAE/LPE/CFP, and compare it with the encoding result of positional encoding for TransAE/PE/CFP, as illustrated in Figure 8. Here, f_1 to f_5 are the encoding visualizations corresponding to the five consecutive input sound signals, respectively, where each input includes four frames.



(a) PE (b) LPE

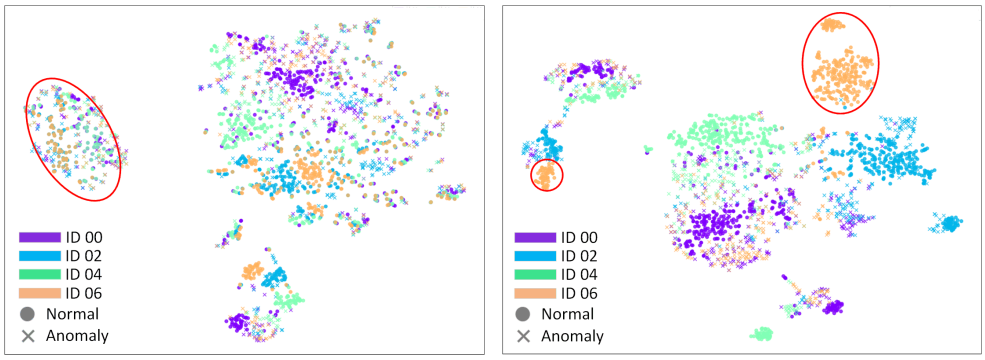
Fig. 8: Encoding visualization of five consecutive input sound signals using Positional Encoding (PE) and Linear Phase Embedding (LPE), respectively.

From Figure 8(a), we can see that the positional encoding visualization of each input is the same because the positional encoding operation adopts the same cosine representation for signal encoding. In contrast, by linearly encoding the phase information of the signal, our proposed LPE can preserve the signal's own temporal information and give different encoding representations for each different input signal, as indicated in the red box in Figure 8(b). Therefore, our proposed method can learn better latent features with unique characteristics from each signal, and enhance the ability of the model for anomalous sound detection.

4.4.3 Validation of IDC Module

We show the t-distributed stochastic neighbor embedding (t-SNE) cluster visualization of the latent features to validate the IDC module further. The experiment is conducted

on the test dataset of the machine type ToyCar, where the proposed method IDC-TransAE without using ID information (i.e., TransAE/LPE/CFP) is employed for comparison. The result is illustrated in Figure 9.



(a) TransAE/LPE/CFP (b) IDC-TransAE

Fig. 9: The t-SNE visualization of latent feature on the test dataset for the machine type ToyCar using TransAE/LPE/CFP and IDC-TransAE. Different color represents different machine ID. The "•" and "×" denote normal and anomalous samples, respectively.

As observed from Figure 9(a), the latent features of normal and anomalous sound samples from different machines overlap with each other when using the method without IDC module (i.e., TransAE/LPE/CFP). In addition, the latent features of the normal sound samples of one machine may be close to that of the anomalous samples from other machines, rather than the normal samples from the same machine, as illustrated in Figure 9(a). It results in the latent features of some anomalous samples from one machine on the manifold of the normal samples from another machine. Thereby these anomalous sounds will be well reconstructed, making it hard to distinguish the anomalies and reducing the detection performance. By introducing the IDC module to constrain the latent feature, the proposed method can reduce the generalization of AE for anomalous sound and further improve its distinguishing ability that the normal and anomalous latent features are well separated, as illustrated in Figure 9(b).

4.4.4 Influence of Parameter r for Anomaly Detection

As mentioned in Section 3.2, we introduce the weighted anomaly score computation to highlight the anomalous events that only appear for a short time. The parameter r in Equation (11) will decide the way for anomaly score computation, i.e., weighted anomaly score computation will degenerate to max anomaly score computation when $r = 0$, and it will become mean anomaly score computation when $r = 1$. Therefore, we also carry out another experiment to show the impact of parameter r on the performance of our proposed IDC-TransAE for anomaly detection. Here, different values of r from $0 \leq r \leq 1$ with an interval of 0.05 are selected to evaluate the

performance of our proposed method in terms AUC and pAUC on all six machine types. The result is shown in Figure 10.

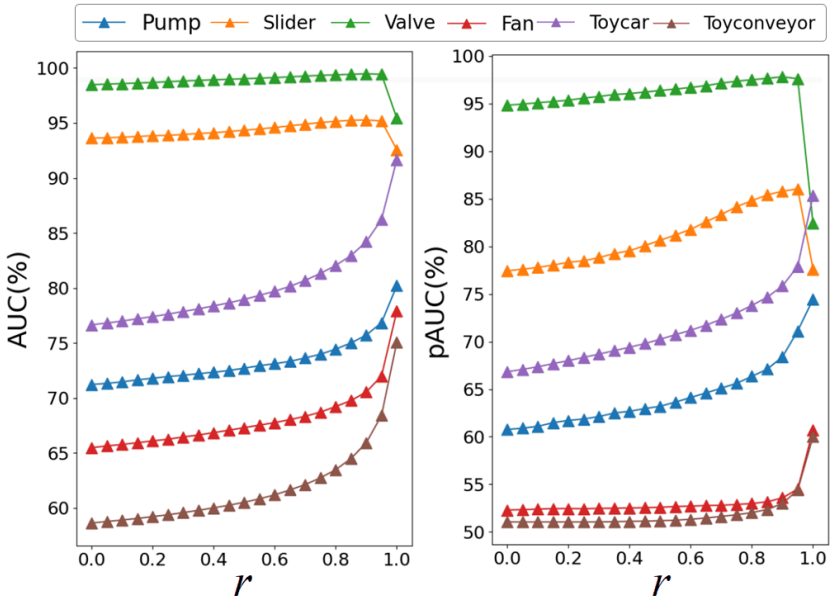


Fig. 10: Performance of our proposed IDC-TransAE in terms of AUC and pAUC under different r values for all six machine types.

From Figure 10, we can see that the mean score computation (i.e., $r = 1$) can achieve the best performance for the machine types of Fan, Pump, ToyCar and ToyConveyor. However, it obtains the worst performance for the machine types of Slider and Valve, where the anomalous sound often occurs in a short time. Though using max anomaly score computation ($r = 0$) can achieve better performance than adopting mean anomaly score computation on Slider and Valve, the weighted anomaly score computation method can provide the best performance for the machine type of Slider and Valve. Especially, the weighted anomaly score computation method can significantly improve the pAUC performance over the mean and max anomaly score computation on Slider and Valve. The result verifies the effectiveness of weighted anomaly score computation for the anomalous sound that appears over short time. In addition, the values of r can be adjusted according to different machine types, which makes it more applicable than mean anomaly score and max anomaly score computation.

5 Conclusions

In this paper, we have presented an IDC-TransAE architecture with weighted anomaly score computation for unsupervised ASD, where an ID classifier was introduced to

mitigate the generalization of AE for anomalous sound and enhance the distinguishing ability for different machines with the same type. In addition, center frame prediction was utilized to improve the reconstruction of the non-stationary sound signal, and a linear phase embedding strategy was applied to preserve the signal's temporal information and further improve its distinguishing ability for anomalous sound detection. Moreover, a weighted anomaly score computation method was introduced to highlight the anomaly scores for anomalous events that only appear for a short time. The experiments demonstrate the effectiveness and superiority of our proposed method, as compared with the baseline methods.

6 Declarations

Availability of data and materials
The datasets for experimental evaluation in this study are from the DCASE 2020 challenge Task 2, which are available on the internet.

Competing interests
W. Wang is an editorial board member of EURASIP Journal on Audio Speech and Music Processing and also a guest editor of the special issue "AI for Computational Audition: Sound and Music Processing", other authors declare that they have no competing interests.

Funding
This work was partly supported by the Natural Science Foundation of Heilongjiang Province under Grant No. YQ2020F010, and a GHfund with Grant No. 202302026860.

Authors' contributions
J. Guan: Conceptualization, Methodology, Writing-Original Manuscript, Funding Acquisition, and Supervision; Y. Liu: Methodology, Experimental Validation, and Data Analysis; Q. Kong: Conceptualization, Methodology, and Writing - Revision & Review; F. Xiao: Experimental Validation and Data Analysis; Q. Zhu: Writing - Revision & Review; J. Tian: Experimental Validation and Data Analysis; W. Wang: Writing - Revision & Review.

Acknowledgements
The authors would like to thank the Associate Editor and the anonymous reviewers for reviewing the manuscript.

References

[1] Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys (CSUR) **41**(3), 1–58 (2009)

[2] Koizumi, Y., Saito, S., Uematsu, H., Kawachi, Y., Harada, N.: Unsupervised detection of anomalous sound based on deep learning and the neyman-pearson

lemma. IEEE/ACM Transactions on Audio, Speech, and Language Processing **27**(1), 212–224 (2018)

[3] Chalapathy, R., Chawla, S.: Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407 (2019)

[4] Nunes, E.C.: Anomalous sound detection with machine learning: A systematic review. arXiv preprint arXiv:2102.07820 (2021)

[5] Guan, J., Liu, Y., Zhu, Q., Zheng, T., Han, J., Wang, W.: Time-weighted frequency domain audio representation with GMM estimator for anomalous sound detection. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023). IEEE

[6] Foggia, P., Petkov, N., Saggese, A., Strisciuglio, N., Vento, M.: Audio surveillance of roads: A system for detecting anomalous sounds. IEEE Transactions on Intelligent Transportation Systems **17**(1), 279–288 (2015)

[7] Li, Y., Li, X., Zhang, Y., Liu, M., Wang, W.: Anomalous sound detection using deep audio representation and a BLSTM network for audio surveillance of roads. IEEE Access **6**, 58043–58055 (2018)

[8] Chung, Y., Oh, S., Lee, J., Park, D., Chang, H.-H., Kim, S.: Automatic detection and recognition of pig wasting diseases using sound data in audio surveillance systems. Sensors **13**(10), 12929–12942 (2013)

[9] Henze, D., Gorishti, K., Bruegge, B., Simen, J.-P.: AudioForesight: A process model for audio predictive maintenance in industrial environments. In: Proceedings of International Conference On Machine Learning And Applications (ICMLA), pp. 352–357 (2019). IEEE

[10] Oh, D.Y., Yun, I.D.: Residual error based anomaly detection using auto-encoder in SMD machine sound. Sensors **18**(5), 1308–1321 (2018)

[11] Park, Y., Yun, I.D.: Fast adaptive RNN encoder–decoder for anomaly detection in SMD assembly machine. Sensors **18**(10), 3573–3583 (2018)

[12] Koizumi, Y., Kawaguchi, Y., Imoto, K., Nakamura, T., Nikaido, Y., Tanabe, R., Purohit, H., Suefusa, K., Endo, T., Yasuda, M., Harada, N.: Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring. In: Proceedings of Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop, Tokyo, Japan, pp. 81–85 (2020)

[13] Kawaguchi, Y., Imoto, K., Koizumi, Y., Harada, N., Niizumi, D., Dohi, K., Tanabe, R., Purohit, H., Endo, T.: Description and discussion on DCASE2021

challenge task 2: Unsupervised anomalous detection for machine condition monitoring under domain shifted conditions. In: Proceedings of Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop, Barcelona, Spain, pp. 186–190 (2021)

[14] Dohi, K., Imoto, K., Harada, N., Niizumi, D., Koizumi, Y., Nishida, T., Purohit, H., Tanabe, R., Endo, T., Yamamoto, M., Kawaguchi, Y.: Description and discussion on DCASE2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques. In: Proceedings of Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop, Nancy, France (2022)

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

[16] Zabihi, M., Rad, A.B., Kiranyaz, S., Gabbouj, M., Katsaggelos, A.K.: Heart sound anomaly and quality detection using ensemble of neural networks without segmentation. In: Proceedings of Computing in Cardiology Conference (CinC), pp. 613–616 (2016)

[17] Tagawa, T., Tadokoro, Y., Yairi, T.: Structured denoising autoencoder for fault detection and analysis. In: Proceedings of Asian Conference on Machine Learning (ACML), pp. 96–111 (2015)

[18] Marchi, E., Vesperini, F., Eyben, F., Squartini, S., Schuller, B.: A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM neural networks. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1996–2000 (2015). IEEE

[19] Marchi, E., Vesperini, F., Weninger, F., Eyben, F., Squartini, S., Schuller, B.: Non-linear prediction with LSTM recurrent neural networks for acoustic novelty detection. In: Proceedings of International Joint Conference on Neural Networks (IJCNN), pp. 1–7 (2015). IEEE

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

[21] Wichern, G., Chakrabarty, A., Wang, Z.-Q., Le Roux, J.: Anomalous sound detection using attentive neural processes. In: Proceedings of Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 186–190 (2021). IEEE

[22] Kim, H., Mnih, A., Schwarz, J., Garnelo, M., Eslami, A., Rosenbaum, D., Vinyals,

O., Teh, Y.W.: Attentive neural processes. arXiv preprint arXiv:1901.05761 (2019)

[23] Van Truong, H., Hieu, N.C., Giao, P.N., Phong, N.X.: Unsupervised detection of anomalous sound for machine condition monitoring using fully connected U-Net. Journal of ICT Research & Applications **15**(1), 41–55 (2021)

[24] Giri, R., Cheng, F., Helwani, K., Tenneti, S.V., Isik, U., Krishnaswamy, A.: Group masked autoencoder based density estimator for audio anomaly detection. In: Proceedings of Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop, pp. 51–55 (2020)

[25] Giri, R., Tenneti, S.V., Helwani, K., Cheng, F., Isik, U., Krishnaswamy, A.: Unsupervised anomalous sound detection using self-supervised classification and group masked autoencoder for density estimation. Technical report, DCASE2020 Challenge (2020)

[26] Zavrtanik, V., Kristan, M., Skočaj, D.: DRAEM - A discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceedings of International Conference on Computer Vision (ICCV), pp. 8330–8339 (2021)

[27] Kapka, S.: ID-conditioned auto-encoder for unsupervised anomaly detection. arXiv preprint arXiv:2007.05314 (2020)

[28] Kuroyanagi, I., Hayashi, T., Adachi, Y., Yoshimura, T., Takeda, K., Toda, T.: An ensemble approach to anomalous sound detection based on Conformer-based autoencoder and binary classifier incorporated with metric learning. In: Proceedings of Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop, Barcelona, Spain, pp. 110–114 (2021)

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

[30] Wilkinghoff, K.: Combining multiple distributions based on sub-cluster adacos for anomalous sound detection under domain shifted conditions. In: Proceedings of Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop, Barcelona, Spain, pp. 55–59 (2021)

[31] Venkatesh, S., Wichern, G., Subramanian, A., Le Roux, J.: Improved domain generalization via disentangled multi-task learning in unsupervised anomalous sound detection. In: Proceedings of Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop, Nancy, France (2022)

[32] Liu, Y., Guan, J., Zhu, Q., Wang, W.: Anomalous sound detection using spectral-temporal information fusion. In: Proceedings of International Conference on

Acoustics, Speech and Signal Processing (ICASSP), pp. 816–820 (2022). IEEE

[33] Guan, J., Xiao, F., Liu, Y., Zhu, Q., Wang, W.: Anomalous sound detection using audio representation with machine ID based contrastive learning pretraining. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023). IEEE

[34] Hejing, Z., Jian, G., Qiaoxi, Z., Feiyang, X., Youde, L.: Anomalous sound detection using self-attention-based frequency pattern analysis of machine sounds. In: Proceedings of INTERSPEECH, pp. 336–340 (2023)

[35] Xiao, F., Liu, Y., Wei, Y., Guan, J., Zhu, Q., Zheng, T., Han, J.: The DCASE2022 challenge task 2 system: Anomalous sound detection with self-supervised attribute classification and GMM-based clustering. Technical report, DCASE2022 Challenge (2022)

[36] Wei, Y., Guan, J., Lan, H., Wang, W.: Anomalous sound detection system with self-challenge and metric evaluation for DCASE2022 challenge task 2. Technical report, DCASE2022 Challenge (2022)

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

[38] Tabak, E.G., Turner, C.V.: A family of nonparametric density estimation algorithms. Communications on Pure and Applied Mathematics **66**(2), 145–164 (2013)

[39] Dinh, L., Krueger, D., Bengio, Y.: Nice: Non-linear independent components estimation. arXiv preprint arXiv:1410.8516 (2014)

[40] Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. In: Proceedings of Advances in Neural Information Processing Systems (NIPS) (2018)

[41] Papamakarios, G., Pavlakou, T., Murray, I.: Masked autoregressive flow for density estimation. In: Proceedings of Advances in Neural Information Processing Systems (NIPS) (2017)

[42] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems (NIPS) (2017)

[43] Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: Proceedings of European Conference on Computer Vision (ECCV), pp. 695–711 (2016). Springer

[44] Koizumi, Y., Saito, S., Uematsu, H., Harada, N.: Optimizing acoustic feature extractor for anomalous sound detection based on Neyman-Pearson lemma. In: Proceedings of European Signal Processing Conference (EUSIPCO), pp. 698–702 (2017). IEEE

[45] Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 315–323 (2011). PMLR

[46] Murphy, K.P.: Machine Learning: A Probabilistic Perspective. MIT press (2012)

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

[48] Koizumi, Y., Saito, S., Uematsu, H., Harada, N., Imoto, K.: ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection. In: Proceedings of Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 313–317 (2019). IEEE

[49] Perez-Castanos, S., Naranjo-Alcazar, J., Zuccarello, P., Cobos, M.: Anomalous sound detection using unsupervised and semi-supervised autoencoders and gammatone audio representation. arXiv preprint arXiv:2006.15321 (2020)

[50] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)