Self-Supervised Augmented Diffusion Model for Anomalous Sound Detection

Jiawei Yin[1,2], Wenbin Zhang[2], Mingjun Zhang[2] and Yu Gao[2,*]
[1] Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education,
East China University of Science and Technology, Shanghai, 200237, China
[2] AI Research Center, Midea Group (Shanghai) Co.,Ltd., Shanghai 201702, China
E-mail: y30221023@mail.ecust.edu.cn, {zhangwb87, zhangmj139, gaoyu11}@midea.com

***Abstract*—Generative models have significantly enhanced the capability of unsupervised anomalous sound detection (ASD) with their strong data modeling capabilities. However, many existing ASD methods based on generative models focus solely on accurately reconstructing sound data itself, neglecting the use of metadata. This results in limited features learned by the models. Additionally, these methods suffer from issues such as low generation quality and mode collapse. To address these challenges, we propose a self-supervised augmented diffusion model (SSDM) to improve ASD performance. SSDM learns expressive embeddings through a self-supervised learning module with a dual-path time-frequency self-attention ASD framework and then uses a denoising diffusion module to learn the distribution of these embeddings as the basis for anomaly detection. The reconstruction loss, which is derived from the reconstruction of test data embeddings measured by the self-supervised learning module in the denoising diffusion module, is used as the anomaly score. Experiments on the DCASE Challenge 2023 Task 2 development dataset demonstrate the effectiveness of the proposed method.**

I. Introduction

Anomalous sound detection (ASD) is a task that involves distinguishing between normal and abnormal states of a machine by analyzing the sounds it emits [1]. However, due to the infrequency and diversity of anomalous sounds, it is challenging to collect data that covers all possible anomalies. Consequently, this task primarily involves learning the distribution of normal sounds to detect sounds that deviate from this distribution and classify them as anomalies. The main approach in existing methods is to use generative models to minimize the reconstruction error of normal sounds during the training phase, and then use the reconstruction error as an anomaly score during the inference phase. By effectively capturing the complex distribution patterns in the data, generative models have become an intuitive and widely applied method for ASD [2]–[4].

Despite the success of generative models, they still face several challenges. Firstly, generative models are designed to learn the distribution of normal data, so the optimization objectives of these methods are primarily focused on the audio data itself [5]. This often leads to insufficient utilization of metadata [6], such as machine type or operating conditions. This approach is suboptimal because operating conditions often contain features that reflect the differences in sounds of the same type of machine [7]. As a result, the audio features learned by generative models for the same machine type can become more complex, leading to unstable ASD performance. Secondly, Variational Autoencoders (VAEs) [8]–[10] and Generative Adversarial Networks (GANs) [3], [4], [11] are the most commonly used generative models for ASD. However, VAEs tend to generate samples of lower quality, while GANs are prone to mode collapse, which hinders the accurate detection of anomalous sounds.

Recently, several self-supervised classification methods have been proposed [7], [12]–[14], leveraging metadata such as machine type and operating conditions as classification conditions. These methods utilize advanced classification networks to learn discriminative high-level features from normal data. K-nearest neighbors (KNN) are then employed as anomaly detectors to measure the distance between test features and normal features to assess anomalies. This approach has achieved satisfactory performance in ASD tasks. However, it is not always stable [15]. This instability arises due to the difficulty of the classification task, as the features learned through auxiliary tasks may not be sufficiently fine-grained [16], leading to the anomaly detector's inability to effectively distinguish between normal and anomalous sounds.

To this end, we propose a self-supervised augmented diffusion model (SSDM) for ASD. First, we design a self-supervised learning module that leverages auxiliary classification tasks to learn expressive and discriminative latent features. Self-attention mechanisms are incorporated into the module to enhance the learning of features along the time and frequency axes. Then, we design a denoising diffusion module based on the diffusion model [17]. The denoising diffusion module learns the data distribution directly from these discriminative latent features rather than from the original data, enabling more effective anomaly detection. In the denoising diffusion module, during the training phase, the latent features are progressively corrupted through a forward diffusion process, and a reverse denoising process is employed to reconstruct and capture the distribution of normal data. During the testing phase, anomalous sounds are identified based on the reconstruction error of the input. We conducted experiments on the DCASE 2023 Challenge Task 2 development dataset, and the results demonstrate the effectiveness of our proposed method.

The main contributions of this paper can be summarized as

*Corresponding author.

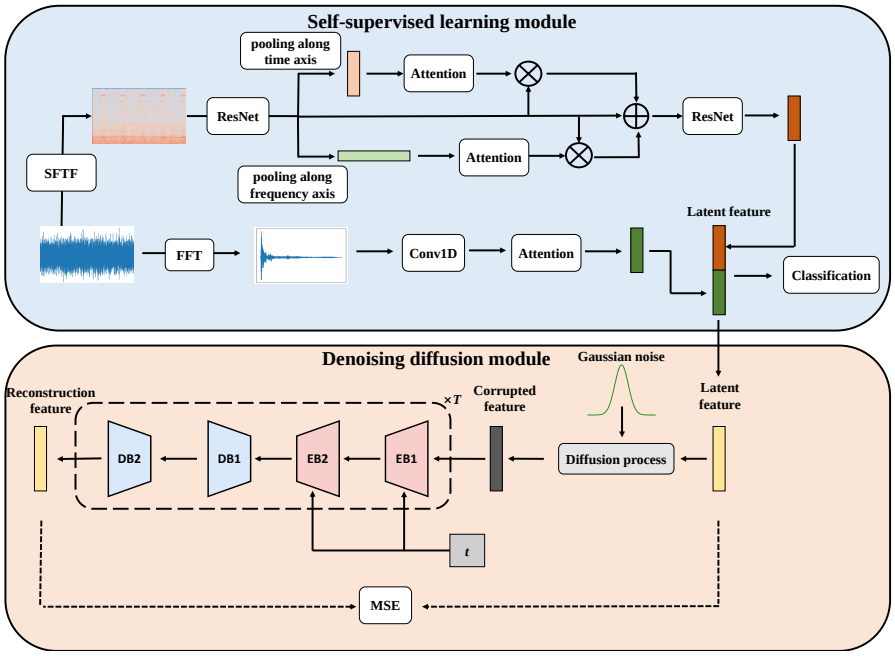


Fig. 1: The overview of our proposed SSDM method

follows:

1) To effectively utilize the meta-information of sound data, we design a self-supervised learning module with a dual-path time-frequency framework to extract discriminative feature embeddings. We incorporate time-frequency self-attention mechanisms to enhance the focus on features along the time and frequency axes.
2) To learn the fine-grained distribution of the embeddings, we design a denoising diffusion module that reconstructs the data through a diffusion and denoising process, serving as the basis for anomaly sound detection. The reconstruction error of the embeddings from the test data is used as the anomaly score to identify anomalous sounds.
3) Experiments on the DCASE 2023 development set show that our proposed SSDM method outperforms baseline and state-of-the-art methods.

II. METHOD

Figure 1 provides an overview of our proposed SSDM method. First, in the self-supervised learning module, we use a dual-path ASD framework to accomplish an auxiliary classification task, introducing a self-attention module to enhance focus on the temporal and frequency aspects of the input. The self-supervised learning module outputs latent feature embeddings. These embeddings are then used in a denoising diffusion module to learn the distribution. The latent features are first progressively corrupted by a forward diffusion process and then reconstructed through a reverse denoising process to capture the distribution of normal data. Finally, anomalies are identified by using the reconstruction error as the anomaly score to detect anomalous sounds.

A. Self-supervised learning module

The primary purpose of generative models is to learn the distribution of input data, but they cannot fully utilize the metadata within the data to learn high-level features with discriminative semantics. To address this, we introduce a self-supervised learning module designed to extract data representations suitable for the ASD task. We employ a dual-path ASD framework [6] as the backbone of our self-supervised learning module to separately extract spectrogram and full magnitude frequency spectrum features. The spectrum path is processed by 1D convolutional layers, while the spectrogram path is handled by ResNet [18] layers. The dual-path framework has been proven to effectively extract expressive features from sounds compared to using a single path alone. We integrate self-attention modules along specific dimensions into both branches. The self-attention module used in this work is an improved version of the SE block [19], enhancing information along specific dimensions, as defined below:

$$y = x + \sigma\left(avgpool_i(x) \cdot W^{\mathrm{T}} + b\right) \cdot x \tag{1}$$

where $avgpool_i$ is the average pooling operation along a specific i dimension and σ is the sigmoid function operation. W and b are learning parameters, and x represents the input features. In the spectrogram path, we employ self-attention modules along the time and frequency dimensions. In the spectrum path, we use self-attention modules along the frequency dimensions. The encoded representations from both paths are combined to form the latent feature z, which is used for training an auxiliary classification task through the classification head.

TABLE I: Comparison of different methods with harmonic mean of AUCs and pAUCs.

Method	ToyCar	ToyTrain	Bearing	Fan	Gearbox	Slider	Valve	All hmean
AE (Mahala). [20]	52.83	48.23	60.02	59.90	64.60	71.75	52.83	57.68
MobileNetV2. [21]	55.18	59.03	68.00	53.61	63.63	82.24	60.06	62.00
FeatEx. [12]	52.35	55.45	64.30	63.80	75.39	85.56	88.13	66.88
Han et al. [22]	63.47	57.35	57.10	62.76	67.52	79.11	67.79	64.31
Self-implement Baseline	53.04	53.13	66.00	63.42	72.36	84.23	63.05	63.55
SSDM (Ours)	56.26	57.75	65.84	67.77	72.09	89.55	78.55	68.09

B. Denoising diffusion module

To effectively capture the distribution of the latent features z of normal data output from the self-supervised learning module, we designed a denoising diffusion module based on the denoising diffusion probabilistic model (DDPM) [17]. DDPM has demonstrated its excellent performance in terms of sample data generation quality and training stability [23]–[25]. The denoising diffusion module consists of a forward diffusion process $q(\,z_t|\,z_{t-1})$ and the reverse denoising process $p_\theta\,(z_{t-1}\,|z_t\,)$. In the diffusion process with T steps, random Gaussian noise progressively corrupts the original data distribution, which follows a Markov process:

$$q(\,z_t|\,z_{t-1}) = \mathcal{N}\left(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t\mathbf{I}\right) \tag{2}$$

where $t = 1, \ldots, T$. The amount of noise added at each step is defined by the variance schedule β_t. This allows the distribution of z_t at any given time to be computed solely based on the original input z_0 and the noise addition step t. It can be defined as:

$$z_t = \sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}e \tag{3}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$. In the denoising process, the model learns to optimize the parameters θ. The reverse process is defined as follows:

$$p_\theta\,(z_{t-1}\,|z_t\,) = \mathcal{N}\,(z_{t-1}; \mu_\theta\,(z_t, t)\,, \beta_t\mathbf{I}) \tag{4}$$

where $\mu_\theta\,(z_t, t)$ is mean of noise predicted by the network.

The training of DDPM can be regarded as an autoencoder. In our work, we use two encoder blocks (EBs) and two decoder blocks (DBs), each composed of dense layers, to train the DDPM. The input and output sizes are consistent and the output is predicted noise ε_t. The goal is to make the predicted noise consistent with the real noise e, so MSE loss can be used.

$$Loss = \arg\min\left[\|\varepsilon_t - e\|^2\right] \tag{5}$$

Based on the predicted noise ε_t , the denoising process can be defined as follows:

$$z'_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_t\right) + \sqrt{\beta_t}e \tag{6}$$

In ASD, the anomaly score is derived from the reconstruction error, with data points showing high deviations being more likely to be classified as anomalies. The anomaly score, denoted as $error$, is defined as follows:

$$error = \|z_0 - z'_0\|_2^2 \tag{7}$$

III. EXPERIMENTS AND RESULTS

A. Dataset

We use the development set from Task 2 of the DCASE 2023 challenge to validate the effectiveness of our proposed method. This dataset includes seven types of machines: Bearing, Fan, Gearbox, Slider, ToyCar, ToyTrain, and Valve. The dataset is divided into training and test sets. The training set contains 1000 normal sound samples for each machine type, while the test set includes 100 normal and 100 abnormal sound samples. Additionally, the training set provides attributes for each sample, which can be used during training. We utilize these attributes as classification labels in the self-supervised learning module.

We evaluate performance using the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) and the partial AUC (pAUC). In our paper, we report the harmonic mean of all AUCs and pAUCs for each machine type, as well as the harmonic mean for the entire dataset.

B. Implementation details

In the self-supervised learning module, we use the mixup strategy to augment the original waveforms. The sub-cluster AdaCos [6] is employed as the classification loss function, and we use all possible combinations of machine types and machine attributes as classification labels. The training consists of 50 epochs with a batch size of 128. The optimizer is Adam with an initial learning rate of 0.01, which decreases to 90% of its value every 5 epochs. In the denoising diffusion module, the network hidden layer size is set to 512. During training, the number of diffusion steps is set to 500, and during testing, it is set to 200. The training consists of 500 epochs with a batch size of 128. The optimizer is Adam with a fixed learning rate of 0.001.

C. Performance comparison and ablation studies

We compare our proposed method with the baseline system [20], [21] and other SOTA methods [12], [22], as shown in Table 1. Our method achieves the highest overall harmonic mean, with improvements of 10.41% and 6.09% compared to the baseline model, and a 1.21% increase compared to the SOTA self-supervised method FeatEx. The self-implement baseline in the table is a combination of a dual-path ASD framework and KNN as the anomaly detector. The results demonstrate that the proposed SSDM enhances overall ASD performance.

To validate the roles of the self-supervised learning module and the denoising diffusion module in SSDM, we conducted

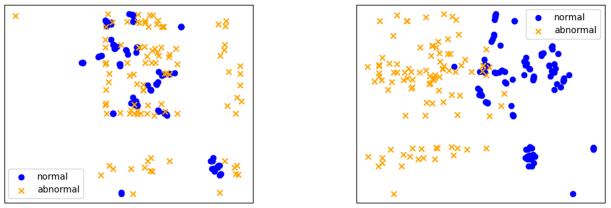
ablation experiments. Table 2 shows the harmonic mean of the performance across all categories for different modules in SSDM. The results indicate that SSDM is highly effective when both modules are included. The self-supervised learning module significantly improves overall ASD performance, while the denoising diffusion module complements it to achieve even better ASD performance.

TABLE II: Ablation studies on the modules of SSDM with harmonic mean of AUCs and pAUCs.

Module	SS+KNN	DM	SSDM
Self-supervised learning module	✓	×	✓
Denoising diffusion module	×	✓	✓
All hmean	65.87	56.11	68.09

D. Visualization analysis

To demonstrate the effectiveness of incorporating time-frequency self-attention in our self-supervised learning module, we use t-SNE to visualize the embeddings of the slider. The slider has been shown in previous studies to have highly distinctive time-frequency characteristics [11], which can help distinguish normal sounds from abnormal ones. As shown in Figure 2, after adding time-frequency self-attention, the overlap between abnormal and normal embeddings is significantly reduced, which will aid in subsequent anomaly detection module.



(a) without time-frequency self-attention (b) with time-frequency self-attention

Fig. 2: Illustration of t-SNE of time-frequency self-attention on slider.

To better illustrate the performance of each module in our proposed SSDM, we visualized the data using kernel density distribution plots. After training with normal data, we input the test data into the trained model and detect anomalies by evaluating the reconstruction error. As shown in Figure 3, different kernel density plots correspond to the probability density distributions of the test data within various modules of SSDM. Figure 3(a) shows the original data distribution, where the distribution of anomalous data largely overlaps with that of normal data. After processing by the self-supervised learning module, as shown in Figure 3(b), the anomalous data becomes roughly separated from the normal data, with the difference in distribution becoming more pronounced. Subsequently, the embedded features are reconstructed using the denoising diffusion module, where the reconstructed anomalous data follows

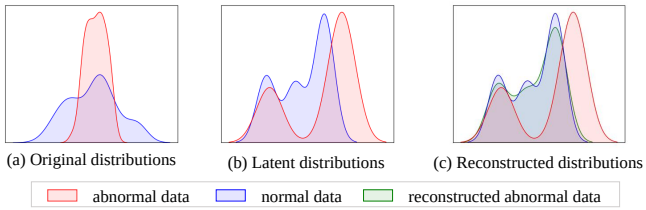


Fig. 3: Kernel density distribution plots of test data in different modules. (a). Original distribution of the test data. (b). Distribution of the latent features after processing by the self-supervised learning module. (c). Distribution of the latent features after reconstruction by the denoising diffusion model.

the distribution pattern of normal data, resulting in significantly higher reconstruction errors for anomalies. As shown in Figure 3(c), accurate anomaly detection can be achieved based on the reconstruction error using SSDM.

IV. Conclusions

In this paper, we introduce a self-supervised augmented diffusion model (SSDM) for anomaly sound detection. SSDM utilizes a self-supervised learning module with dual-path time-frequency self-attention framework to leverage metadata from sounds and learn expressive and discriminative features. The denoising diffusion module learns the distribution of these discriminative latent features and uses the reconstruction error to detect anomalous sounds. Experimental results on the development dataset of DCASE 2023 Task 2 demonstrate the effectiveness of our proposed method.

References

[1] T. Nishida, N. Harada, D. Niizumi, *et al.*, "Description and discussion on dcase 2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2406.07250*, 2024.

[2] H. Yun, H. Kim, Y. H. Jeong, and M. B. Jun, "Autoencoder-based anomaly detection of industrial robot arm using stethoscope based internal sound sensor," *Journal of Intelligent Manufacturing*, vol. 34, no. 3, pp. 1427–1444, 2023.

[3] A. Jiang, W.-Q. Zhang, Y. Deng, P. Fan, and J. Liu, "Unsupervised anomaly detection and localization of machine audio: A gan-based approach," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[4] S. Liu, J. Li, W. Ke, and H. Yin, "Multi-attention enhanced discriminator for gan-based anomalous sound detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 6715–6719.

[5] S. Li, J. Yu, Y. Lu, G. Yang, X. Du, and S. Liu, "Self-supervised enhanced denoising diffusion for anomaly detection," *Information Sciences*, vol. 669, p. 120 612, 2024.

[6] K. Wilkinghoff, "Sub-cluster adacos: Learning representations for anomalous sound detection," in *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2021, pp. 1–8.

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

[8] M.-H. Nguyen, D.-Q. Nguyen, D.-Q. Nguyen, C.-N. Pham, D. Bui, and H.-D. Han, "Deep convolutional variational autoencoder for anomalous sound detection," in *2020 IEEE Eighth International Conference on Communications and Electronics (ICCE)*, IEEE, 2021, pp. 313–318.

[9] H. Purohit, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Hierarchical conditional variational autoencoder based acoustic anomaly detection," in *2022 30th European Signal Processing Conference (EUSIPCO)*, IEEE, 2022, pp. 274–278.

[10] V. Zavrtanik, M. Marolt, M. Kristan, and D. Skočaj, "Anomalous sound detection by feature-level anomaly simulation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 1466–1470.

[11] X. Huang, F. Guo, and L. Chen, "A res-ganomaly method for machine sound anomaly detection," *IEEE Access*, 2024.

[12] K. Wilkinghoff, "Self-supervised learning for anomalous sound detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 276–280.

[13] S. Choi and J.-W. Choi, "Noisy-arcmix: Additive noisy angular margin loss combined with mixup for anomalous sound detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 516–520.

[14] H. Chen, Y. Song, L.-R. Dai, I. McLoughlin, and L. Liu, "Self-supervised representation learning for unsupervised anomalous sound detection under domain shift," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 471–475.

[15] 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)*, IEEE, 2022, pp. 816–820.

[16] J. Guan, F. Xiao, Y. Liu, Q. Zhu, and W. Wang, "Anomalous sound detection using audio representation with machine id based contrastive learning pretraining," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[20] K. Dohi, K. Imoto, N. Harada, *et al.*, "Description and discussion on dcase 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2305.07828*, 2023.

[21] N. Harada, D. Niizumi, Y. Ohishi, D. Takeuchi, and M. Yasuda, "First-shot anomaly sound detection for machine condition monitoring: A domain generalization baseline," in *2023 31st European Signal Processing Conference (EUSIPCO)*, IEEE, 2023, pp. 191–195.

[22] B. Han, Z. Lv, A. Jiang, *et al.*, "Exploring large scale pre-trained models for robust machine anomalous sound detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 1326–1330.

[23] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin, "Diffusion models for medical anomaly detection," in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2022, pp. 35–45.

[24] J. Wyatt, A. Leach, S. M. Schmon, and C. G. Willcocks, "Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 650–656.

[25] X. Zhang, N. Li, J. Li, T. Dai, Y. Jiang, and S.-T. Xia, "Unsupervised surface anomaly detection with diffusion probabilistic model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6782–6791.