THUEE SYSTEM FOR DCASE 2025 ANOMALOUS SOUND DETECTION CHALLENGE

Technical Report

Anbai Jiang[1], *Wenrui Liang*[1], *Shi Feng*[1], *Yihong Qiu*[2], *Yixiang Zhao*[1], *Junjie Li*[1], *Pingyi Fan*[1], *Wei-Qiang Zhang*[1], *Cheng Lu*[2], *Xie Chen*[3], *Yanmin Qian*[3], *Jia Liu*[1,4]

[1] Tsinghua University, Beijing, China
[2] North China Electric Power University, Beijing, China
[3] Shanghai Jiao Tong University, Shanghai, China
[4] Huakong AI Plus Company Limited, Beijing, China
Email: {jab22, lwr24}@mails.tsinghua.edu.cn

ABSTRACT

This technical report presents the THUEE system for the DCASE 2025 anomalous sound detection (ASD) challenge. Motivated by the success of self-supervised learning (SSL) and generative modeling in various modalities and tasks, we build the system by first adapting multiple SSL pre-trained models for ASD. We find that fine-tuning the model with all six DCASE ASD datasets significantly boosts the ASD performance. To address granularity mismatches in machine attributes, we adopt an adaptive prototype modeling scheme. Furthermore, we leverage powerful diffusion-based audio generation models to synthesize samples under minor working conditions, augmenting the imbalanced training set to mitigate domain gaps between source and target distributions. Finally, we conduct mega ensembling of dozens of single models by Bayesian optimization, achieving substantial performance gains. The best ensemble system reaches 74.29% on the DCASE23 dataset, 70.17% on the DCASE24 dataset and 69.35% on the DCASE25 development set.

Index Terms— Anomalous Sound Detection, Self-Supervised Learning, Generative Models, Ensembling

1. INTRODUCTION

Anomalous Sound Detection (ASD) aims to detect anomalous sound when only normal sound is provided for training. That is, the training set contains only normal sounds while the test set contains both normal and anomalous sounds. As an annual ASD challenge, the DCASE 2025 ASD challenge [1, 2, 3, 4] features 15 machine types with eight brand-new machine types. Compared with previous ASD challenges, the DCASE 2025 ASD challenge allows the use of previous DCASE ASD challenge datasets, which addresses one of the crucial difficulites when developing ASD systems. As known, the top-2 systems of the DCASE 2024 challenge [5, 6] and recent state-of-the-art (SOTA) models [7, 8, 9] conformably adopt self-supervised learning (SSL) pre-trained models as the basis and fine-tune them on the ASD dataset to adapt these models from the general domain of audio to the specific domain of machinery sound. However, due to the limited size of the ASD dataset, the fine-tuning task, typically attribute classification, is far too easy for SSL models pre-trained on large datasets, and thus these models can not be well-adapted for the machinery data. Now that all six ASD datasets can be utilized for training, the attribute classification task is much more complex and these models can be fine-tuned for longer rounds before over-fitting, which significantly boosts the performance.

Built upon our previous works [5, 6, 7, 10, 9], the THUEE system is developed by three steps. First of all, we more thoroughly exploit the transfer learning capabilities of SSL models by fine-tuning them on all six ASD datasets. We employ three top-performing SSL models, i.e. BEATs [11], EAT [12] and a self pre-trained SSL model, and adopt an implicit prototype modeling scheme [9] to handle the mixed label granularity. Secondly, we leverage generative models to generate samples of rare working conditions as augmentation for the imbalanced training set. We first train TangoFlux [13] and a diffusion-based model on the ASD dataset respectively, then fine-tune BEATs on the combined dataset of both the real and generated data. Finally, we conduct mega ensembling by linearly combining the anomaly scores of dozens of models, where the coefficients are optimized through Bayesian optimization.

The proposed systems are evaluated on previous ASD datasets to have a fair comparison with SOTA systems. Our best ensemble system achieves remarkable results of 74.29% on the DCASE23 dataset, 70.17% on the DCASE24 dataset and 69.35% on the DCASE25 development set, which are comparable and even superior than previous SOTA systems.

The rest of the report is organized as follows. Section 2 depicts the fine-tuning details of the SSL models. Section 3 depicts the training process of the generative models. Section 4 introduces the mega ensembling procedure and section 5 demonstrates the ASD results.

2. SSL MODELS ADAPTATION

2.1. SSL Models

Three SSL models are employed in the THUEE system, namely BEATs [11], EAT [12] and a self pre-trained SSL model. For BEATs, we adopt the official iteration 3 checkpoint. For EAT, we adopt the official base30 checkpoint. In contrary with BEATs and EAT, our self pre-trained SSL model substitutes log-mel spectrogram with short time Fourier transform (STFT) to retain linear frequency scale and preserve essential high frequency components

This work was supported by the National Key Research and Development Program of China (Grant NO.2021YFA1000500(4)) and the National Natural Science Foundation of China under Grant No. 62276153.

Table 1: Performances of ensemble systems on the DCASE23, DCASE24 and DCASE25 dataset

System	Generation		Model	DCASE23			DCASE24			DCASE25
	TangoFlux	Small Diff	Count	dev	eval	hmean	dev	eval	hmean	dev
23 Challenge Best [14]			-	68.11	66.97	67.54	-	-	-	-
24 Challenge Best [5]			-	-	-	-	67.82	66.24	67.02	-
THUEE System 1			15	70.03	78.15	73.87	70.25	69.07	69.66	68.96
THUEE System 2	✓	✓	74	70.29	**78.77**	**74.29**	70.52	**69.82**	**70.17**	**69.35**
THUEE System 3		✓	34	**70.49**	78.22	74.15	**70.78**	69.35	70.06	68.93
THUEE System 4	✓	✓	38	70.34	78.42	74.16	70.50	69.38	69.94	69.17

that are strongly associated with malfunctions, which would be diluted if mel-scale was adopted. After converting the waveform to STFT spectrogram, the model splits the spectrogram with a fixed band width and conduct masked-and-predict SSL training on these sub-bands in a teacher student framework, where the network is ViT-base [15]. The final utterance-level embedding is the concatenation of the [CLS] embeddings of all sub-bands. The model is trained on a combined dataset of Audioset [16], Freesound[1], MTG-Jamendo [17] and Music4all [18] with a total volume of 17k hours. Our self pre-trained model will be formally introduced in an upcoming research paper.

2.2. Fine-tuning

The above SSL models are fine-tuned on all six DCASE ASD datasets [19, 20, 21, 2, 3], i.e. DCASE20, DCASE21, DCASE22, DCASE23, DCASE24 and DCASE25. It is noted that we do not utilize the supplementary data of the DCASE25 dataset since adding them deprecates the performance. The classification label is selected as the combination of dataset year, machine type, section and attributes (if available), where each unique combination is select as a new class, resulting in 1114 classes. It is noted that these ASD datasets have shared machine types, whereas some datasets do not provide attribute information (all of DCASE20 and some machine types of DCASE24 and DCASE25), leading to mixed label granularity. As a solution, we adopt the adaptive prototype learning scheme proposed in our previous work [9] where we insert 16 learnable sub-centers into the embedding space for each class that does not contain attribute information. During training, the network embedding is mapped to the nearest sub-center and the corresponding sub-center rather than the original embedding is utilized for loss computation. In this way, these sub-centers act as implicitly learned attribute centers, and thus label granularity can be conformably aligned to the attribute level.

The rest of the fine-tuning details are mostly identical with AnoPatch [7]. For BEATs, we append an attentive statistical pooling layer to the ViT backbone. For EAT and our self pre-trained model, we extract the embedding corresponding to the [CLS] token. All parameters are updated during fine-tuning since we find that full fine-tuning is more effective than LoRA fine-tuning [8] when the data size scales up. All models are trained for 20k steps with warmup learning schedulers.

2.3. Anomaly Detection

The anomaly detection procedure is mainly identical to [9]. We use k-nearest neighbor (KNN) detector (k=1) for anomaly detection, where embeddings of normal samples are first extracted to form two memory banks, one for the source domain and the other for the target domain. The anomaly score of a query embedding is the minimum distance to the nearest neighbor in two memory banks. For EAT, we employ SMOTE [22] to oversample target embeddings. Anomaly detection is conducted for each section respectively.

[1] https://freesound.org/

3. GENERATIVE MODELS AS AUGMENTATION

Recent works [23, 24, 25] have demonstrated that generating rare samples by diffusion-based models is effective for ASD performance. To further exploit the ASD capabilities of SSL models, we leverage two different generative models to augment the ASD training set.

3.1. Fine-tuning Pre-trained TangoFlux

To augment the training set with realistic synthetic data, we employ a fine-tuned version of the TangoFlux [13] model to generate machine audio conditioned on machine type and attribute labels. We first construct text-audio pairs by generating captions for each audio clip in all six DCASE ASD datasets. These captions are created using predefined textual templates that reflect both the machine type and operational condition. The resulting text-audio pairs are then used to fine-tune a pretrained TangoFlux model, enabling it to generate machine sounds guided by textual prompts corresponding to target machine types and attributes.

To improve the fidelity of the generated audio, we adopt a reference-based generation strategy. For each sample, we select a reference audio from the original dataset and encode it into the latent space using the model's VAE encoder [26]. The encoded reference is then processed through a forward diffusion and reverse denoising process, conditioned on both the text prompt and the reference embedding. This approach helps preserve realistic acoustic features related to the target machine sound.

To ensure the quality and semantic correctness of the generated data, we apply a sample screening step. A BEATs model is fine-tuned on the original dataset to perform attribute classification and is used to evaluate each generated audio clip. Only samples whose predicted attribute label matches the intended condition are retained. This filtering process ensures that only high-quality, label-consistent audio is added to the final training set.

3.2. Training Diffusion Models from Scratch

In addition to direct fine-tuning SSL models, we also trained a diffusion model [27] from scratch to augment the original dataset. This model directly generates log-mel spectrograms of machine audio,

Table 2: Detailed results of ensemble systems on the DCASE25 development set

Machine	Metric	System 1	System 2	System 3	System 4
bearing	AUC_s	65.42	68.30	66.16	67.10
	AUC_t	67.82	69.56	67.72	67.96
	pAUC	57.16	60.58	58.74	59.42
	hmean	63.12	65.90	63.96	64.59
fan	AUC_s	61.66	61.24	61.12	61.60
	AUC_t	63.46	62.58	64.00	63.64
	pAUC	55.63	55.00	56.00	56.05
	hmean	60.06	59.42	60.19	60.26
gearbox	AUC_s	84.20	81.88	82.38	82.92
	AUC_t	83.30	82.02	82.58	83.10
	pAUC	68.42	68.32	67.16	69.00
	hmean	77.93	76.84	76.65	77.75
slider	AUC_s	91.72	92.56	92.52	92.44
	AUC_t	76.90	78.56	76.54	76.96
	pAUC	58.84	60.26	58.11	58.53
	hmean	73.35	74.76	73.02	73.35
ToyCar	AUC_s	69.42	63.78	69.44	70.02
	AUC_t	63.56	63.78	63.40	63.28
	pAUC	52.95	53.00	52.32	52.53
	hmean	61.19	61.44	60.87	61.07
ToyTrain	AUC_s	76.52	76.50	77.26	76.78
	AUC_t	70.04	68.92	69.94	69.06
	pAUC	57.89	55.95	56.26	56.11
	hmean	67.24	66.00	66.64	66.18
valve	AUC_s	87.68	89.74	89.44	89.26
	AUC_t	93.42	94.34	93.88	93.84
	pAUC	83.95	86.37	86.63	86.53
	hmean	88.18	90.03	89.89	89.78
hmean	AUC_s	75.14	75.69	75.31	75.66
	AUC_t	72.74	72.87	72.70	72.62
	pAUC	60.79	61.27	60.65	61.05
	hmean	68.96	69.35	68.93	69.17

AUC_s and AUC_t are the AUC of the source and target domains, respectively.

with a U-Net architecture serving as the noise prediction network. To enable controllable data generation, the model is conditioned on the working condition attributes, which are embedded and integrated into the network. By undergoing the standard forward noising and reverse denoising processes, the model learns the data distribution for each specific class.

After training, a critical data screening process is applied to the generated samples before they are added to the training set. For this purpose, we first train an auxiliary Xception [28] classification network on the original training set. This classifier is then used to predict the class label and classification confidence for each spectrogram generated by the diffusion model. We selectively retain the generated samples that satisfy two criteria: 1) the predicted class label matches the intended class used as a condition during generation, and 2) the classification confidence is not excessively high. We hypothesize that such samples effectively balance generation quality with data diversity, representing a valuable supplement to the original real data by providing novel yet plausible examples. The filtered synthetic data are then combined with the real data to fine-tune the SSL-based ASD models.

4. MEGA ENSEMBLING

We conduct mega score ensembling for up to 74 models in submitted systems, where anomaly scores of these models are linearly combined into an overall anomaly score. The combination coefficients are obtained by Bayesian optimization which models the mapping from coefficient to detection benchmark as a Gaussian process. Compared with commonly adopted grid search method, Bayesian optimization excels in finer granularity and much lower complexity, enabling the combination of dozens of models.

The ensemble process consists of two steps. In the first step, each model is trained under five different seeds, and we aggregate the scores of these homogeneous models. In the second step, we aggregate the scores of heterogeneous models mentioned in section 2 and section 3. System 1 is the ensemble of SSL models trained without generated samples. System 3 is the ensemble of SSL models trained with original data and samples generated by the small

diffusion model. System 2 and system 4 are ensembles of SSL models trained with original data and all generated samples, with system 2 adopting a more radical set of coefficients.

5. EXPERIMENT

The THUEE system is trained on all six ASD datasets and evaluated on the DCASE25 development set. To provide a fair comparison with previous SOTA models, we also report the results on the DCASE23 dataset and the DCASE24 dataset. Table 1 depicts the performances of four ensemble systems on three test sets, while Table 2 presents the detailed results on the DCASE25 development set. The best performing ensemble system, i.e. system 2, achieves excellent scores of 74.29%, 70.17% and 69.35% on three ASD datasets.

6. CONCLUSION

The THUEE system is developed by three steps. First of all, we build up the basis by fine-tuning three powerful SSL models on all six ASD datasets. Secondly, we train two generative models to generate samples of rare working conditions, which are incorporated in the training set of these SSL models. Finally, we carry out mega score ensembling by Bayesian optimization with two step hierarchy. Our systems are competitive on multiple ASD datasets.

7. REFERENCES

[1] T. Nishida, N. Harada, D. Niizumi, D. Albertini, R. Sannino, S. Pradolini, F. Augusti, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2025 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *In arXiv e-prints: 2506.10097*, 2025.

[2] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Barcelona, Spain, November 2021, pp. 1–5.

[3] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

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

[5] Z. Lv, A. Jiang, B. Han, Y. Liang, Y. Qian, X. Chen, J. Liu, and P. Fan, "Aithu system for first-shot unsupervised anomalous sound detection," DCASE2024 Challenge, Tech. Rep., June 2024.

[6] A. Jiang, X. Zheng, Y. Qiu, W. Zhang, B. Chen, P. Fan, W.-Q. Zhang, C. Lu, and J. Liu, "Thuee system for first-shot unsupervised anomalous sound detection," DCASE2024 Challenge, Tech. Rep., June 2024.

[7] A. Jiang, B. Han, Z. Lv, Y. Deng, W.-Q. Zhang, X. Chen, Y. Qian, J. Liu, and P. Fan, "Anopatch: Towards better consistency in machine anomalous sound detection," in *Interspeech 2024*, 2024, pp. 107–111.

[8] X. Zheng, A. Jiang, B. Han, Y. Qian, P. Fan, J. Liu, and W.-Q. Zhang, "Improving anomalous sound detection via low-rank adaptation fine-tuning of pre-trained audio models," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 969–974.

[9] A. Jiang, X. Zheng, B. Han, Y. Qiu, P. Fan, W.-Q. Zhang, C. Lu, and J. Liu, "Adaptive prototype learning for anomalous sound detection with partially known attributes," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[10] A. Jiang, Y. Shi, P. Fan, W.-Q. Zhang, and J. Liu, "Coopasd: Cooperative machine anomalous sound detection with privacy concerns," 2024. [Online]. Available: https://arxiv.org/abs/2408.14753

[11] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 23–29 Jul 2023, pp. 5178–5193.

[12] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "Eat: Self-supervised pre-training with efficient audio transformer," *arXiv preprint arXiv:2401.03497*, 2024.

[13] C.-Y. Hung, N. Majumder, Z. Kong, A. Mehrish, A. Zadeh, C. Li, R. Valle, B. Catanzaro, and S. Poria, "Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization," 2024. [Online]. Available: https://arxiv.org/abs/2412.21037

[14] J. Jie, "Anomalous sound detection based on self-supervised learning," DCASE2023 Challenge, Tech. Rep., June 2023.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[16] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[17] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The mtg-jamendo dataset for automatic music tagging," in *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, United States, 2019. [Online]. Available: http://hdl.handle.net/10230/42015

[18] I. A. P. Santana, F. Pinhelli, J. Donini, L. Catharin, R. B. Mangolin, V. D. Feltrim, M. A. Domingues, *et al.*, "Music4all: A new music database and its applications," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2020, pp. 399–404.

[19] H. Purohit, R. Tanabe, T. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "MIMII Dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, November 2019, pp. 209–213. [Online]. Available: http://dcase.community/documents/workshop2019/proceedings/DCASE2019Workshop_Purohit_21.pdf

[20] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, November 2019, pp. 308–312. [Online]. Available: https://ieeexplore.ieee.org/document/8937164

[21] R. Tanabe, H. Purohit, K. Dohi, T. Endo, Y. Nikaido, T. Nakamura, and Y. Kawaguchi, "MIMII DUE: Sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due to changes in operational and environmental conditions," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 21–25, 2021.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[23] H. Zhang, Q. Zhu, J. Guan, H. Liu, F. Xiao, J. Tian, X. Mei, X. Liu, and W. Wang, "First-shot unsupervised anomalous sound detection with unknown anomalies estimated by metadata-assisted audio generation," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1271–1275.

[24] J. Yin, Y. Gao, W. Zhang, T. Wang, and M. Zhang, "Diffusion augmentation sub-center modeling for unsupervised anomalous sound detection with partially attribute-unavailable conditions," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[25] J. Yin, W. Zhang, M. Zhang, and Y. Gao, "Self-supervised augmented diffusion model for anomalous sound detection," in *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2024, pp. 1–5.

[26] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022. [Online]. Available: https://arxiv.org/abs/1312.6114

[27] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=St1giarCHLP

[28] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.