MULTI-MODAL ACOUSTIC ANOMALY DETECTION VIA RECONSTRUCTION AND DISCRIMINATIVE LEARNING WITH BEATS REPRESENTATIONS

Technical Report

Pengyuan Zhao[1]*, Zulong Yan*[1]*, Tianju Zhao*[1]*, Yutao Zhang*[1]*, Meng Lei*[1]*,*

[1] China University of Mining and Technology, XuZhou, China
{pyzsuoerhero,zulongyan,tjzhao,04211559,lmsiee}@cumt.edu.cn

ABSTRACT

This technical report focuses on anomalous sound detection (ASD) in DCASE 2025 Task 2, we propose two deep learning approaches based on multimodal feature fusion to enhance robustness and generalization across domains. In the data preparation stage, in order to solve the problem of data complexity, this paper separates the pure sound events and background noise provided by the organizer based on TF-Locoformer, and constructs a more robust data set for model training by reconstructing diversified training samples through random combination. The first approach extracts frame-level waveform features using a fine-tuned BEATs model and aligns them with Mel-spectrogram features extracted by MobileFaceNet. These are fused and passed into an ArcFace classifier for joint attribute and domain classification, enabling discriminative learning and multi-task optimization. The second approach introduces a multimodal autoencoder architecture combining BEATs and TgramNet for hierarchical feature extraction, jointly trained with reconstruction and classification losses. Our best model achieves a pAUC of 0.59.56 on the validation set, demonstrating strong detection performance under multi-source and complex background conditions.highlighting the effectiveness and potential of the proposed methods in real-world, multi-domain ASD scenarios.

***Index Terms*—** Dcase, anomaly detection, BEATs, MobileFaceNet, ArcFace

1. INTRODUCTION

In industrial environments [1], maintaining the health of machinery is critical for ensuring operational continuity and reducing unplanned downtime. Anomalous Sound Detection (ASD) [2] has gained attention as a non-invasive, cost-effective approach for monitoring machine conditions [3] by identifying deviations in acoustic patterns. Unlike traditional supervised approaches that require large amounts of labeled anomalous data—which are often unavailable in real-world applications—DCASE 2025 Task 2 emphasizes *first-shot unsupervised ASD* [4] [5] [6], where models must detect unknown anomalies using only a few normal samples from a target machine. This task poses significant challenges such as domain shift [7], data scarcity, and generalization to unseen conditions, making it a compelling benchmark for developing robust and adaptable models. In this study, we explore ASD methods that aim to improve performance in low-data and cross-domain scenarios [8], aligning with the real-world demands of machine condition monitoring.

Current research in ASD has made significant strides, with numerous studies focusing on supervised learning approaches that rely on large annotated datasets [9]. However, these methods are often limited by the availability of labeled data and may not generalize well to new, unseen anomalies. Recent advancements in unsupervised and semi-supervised learning have shown promise in addressing these limitations [10], but challenges remain in achieving high detection accuracy and robustness across different machine types and environments.

In this work, we propose two complementary approaches tailored for the DCASE 2025 Task 2, which target the challenge of anomalous sound detection under domain shift and limited supervision. Both methods aim to leverage the powerful generalization capability of pre-trained audio encoders BEATs [11], while enhancing model performance through multi-modal feature integration and targeted optimization strategies.Specifically, the two proposed approaches are:

Approach 1: A discriminative multi-modal learning framework that extracts temporal audio embeddings from BEATs and spectral representations via Mel-spectrograms. These are fused and passed through a MobileFaceNet backbone [12], followed by an ArcFace-based [13] classifier to perform machine identity recognition and domain classification jointly. This design encourages the model to learn robust, domain-invariant representations under a multi-task setting.

Approach 2: A reconstruction-based anomaly detection approach that employs a dual-encoder autoencoder structure, where BEATs and TgramNet [14] extract hierarchical features. The model is optimized with both reconstruction loss and auxiliary classification loss. Anomaly scoring is conducted using Mahalanobis distance [15] computed in the latent space, enabling effective detection of unseen anomalous patterns.

These two strategies complement each other by integrating both discriminative and generative perspectives, thereby improving the model's ability to detect anomalous sounds across diverse acoustic domains.

2. METHODOLOGY

This section presents two approaches developed for DCASE 2025 Task 2: a discriminative multimodal model and a reconstruction-based self-supervised anomaly detection model. These methods address the anomalous sound detection task from complementary angles—discriminative learning and reconstruction-driven anomaly scoring—aiming to improve anomaly detection accuracy and cross-domain generalization. To further enhance model robustness, we incorporate a data augmentation strategy based on TF-Locoformer, which separates audio into foreground sound events and back-

ground noise. By randomly recombining these components across different domains, we generate diverse synthetic training samples, significantly improving the models' adaptability to complex and unseen acoustic environments.

2.1. Discriminative Multi-Modal Framework (Approach 1)

The Approach 1 network is a multi-modal deep learning framework designed to enable efficient audio classification and feature learning by integrating various audio feature extraction methods. It combines the BEATs model, TgramNet, and MobileFaceNet to extract hierarchical semantic information from waveforms, spectrograms, and frame-level representations. The multi-modal fusion strategy effectively enhances the overall classification performance.

Initially, the network utilizes a pre-trained BEATs model to extract frame-level features from input audio waveforms, producing representations of shape [`batch_size`, `seq_len`, `hidden_dim`]. To tailor the model for the target task, the last four encoder layers and the positional convolutional layer of BEATs are unfrozen for fine-tuning, thereby improving task-specific performance while preserving pre-trained knowledge. Concurrently, the TgramNet module employs 1D convolutions and a deep convolutional encoder to derive spectral features directly from the raw waveform, capturing rich time-frequency structures. Meanwhile, MobileFaceNet serves as the backbone, employing lightweight depthwise separable convolutions and bottleneck blocks to deeply encode the fused multi-modal features, producing both classification outputs and feature embeddings.

For feature fusion, the frame-level features from BEATs are interpolated to align temporally with the spectral features, then concatenated with the TgramNet-extracted features and the input Mel-spectrogram to form a unified multi-modal representation. This representation is processed by MobileFaceNet's deep convolutional layers, allowing the network to effectively integrate and abstract multi-modal features into global semantic representations.

Additionally, the ArcFace module is employed to introduce an angular margin, thereby enhancing the discriminative power of the learned features and improving generalization. To further improve the network's robustness, a multi-task learning objective is incorporated. Specifically, the model jointly optimizes an attribute prediction loss ($\mathcal{L}_{attr}$), and a domain classification loss ($\mathcal{L}_{dom}$). The overall training objective is defined as:

$$\mathcal{L}_{attr} = \text{CrossEntropy}(f_{attr}(x), y_{attr}) \quad (1)$$

$$\mathcal{L}_{dom} = \text{CrossEntropy}(f_{dom}(x), y_{dom}) \quad (2)$$

where $f_{attr}(x)$ and $f_{dom}(x)$ are the output logits of the attribute and domain classifiers, respectively, and y_{attr} and y_{dom} are the corresponding ground truth labels. The overall training objective is:

$$\mathcal{L}_{total} = \mathcal{L}_{id} + \lambda_{attr} \cdot \mathcal{L}_{attr} + \lambda_{dom} \cdot \mathcal{L}_{dom}, \quad (3)$$

where $\mathcal{L}_{id}$ is the ArcFace-based identity classification loss, and λ_{attr} and λ_{dom} are hyperparameters that balance the contributions of the auxiliary tasks.

In summary, this framework captures multi-level semantic information from audio through coordinated multi-modal feature extraction, fusion, and classification. Its modular design enables flexible adaptation to various audio analysis tasks. By incorporating fine-tuned pre-trained models, deep multi-modal fusion, and multi-task learning, the model achieves significantly improved classification performance and enhanced feature representation capability.

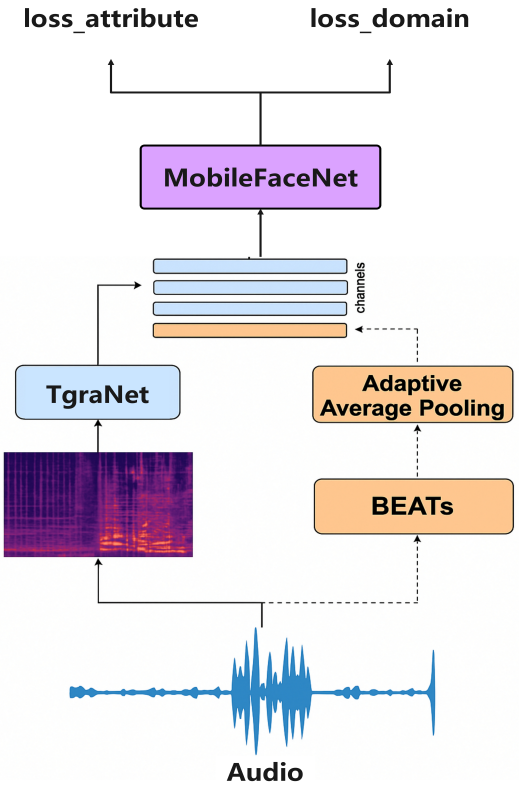

Figure 1: Discriminative Multi-Modal Training Framework

2.2. Reconstruction-based Anomaly Detection (Approach 2)

This approach adopts a reconstruction-based self-supervised paradigm for anomaly detection. The central idea is to learn the normal acoustic patterns of machines during training and identify anomalies during inference based on reconstruction errors or deviations in the latent representation.

The input audio waveform is simultaneously fed into two feature encoders: BEATs and TgramNet. BEATs captures high-level temporal representations using a pretrained transformer-based architecture, while TgramNet extracts complementary low-level spectral features from the raw waveform via convolutional operations. The extracted features are concatenated along the channel dimension to form a joint multi-scale representation.

This fused feature representation is then passed through a lightweight convolutional autoencoder (Conv-AE), which is designed to reconstruct the original input from its compressed latent code. The encoder part of the Conv-AE transforms the input into a low-dimensional latent vector, and the decoder attempts to restore the original feature. In addition to reconstruction, the encoded latent vector is also passed to an auxiliary classifier that predicts the machine identity, which guides the encoder to preserve discriminative semantic information in the compressed space.

The model is trained using a joint loss that combines reconstruction loss and classification loss. The reconstruction objective is formulated as a mean squared error (MSE) between the original and reconstructed features. The classification loss is computed us-

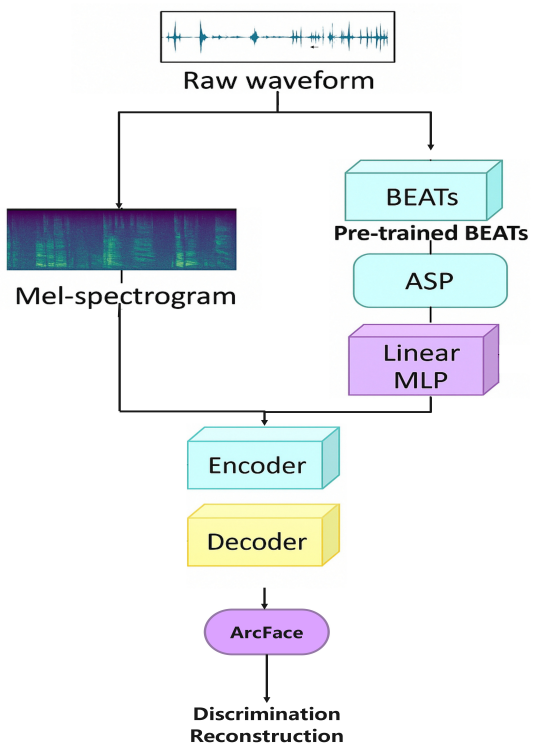

Figure 2: NetMamba Encoder

ing cross-entropy between the predicted and ground-truth machine IDs. The total training loss is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{recon} + \alpha \cdot \mathcal{L}_{CE}, \tag{4}$$

where $\mathcal{L}_{recon}$ is the MSE loss, $\mathcal{L}_{CE}$ is the cross-entropy loss, and α is a balancing hyperparameter.

During inference, anomaly detection can be performed using two complementary strategies. First, the Mahalanobis distance between the encoded latent feature and its corresponding class center is computed to quantify distributional deviations, which serves as an anomaly score. Second, the frame-level reconstruction error itself can be used directly as an alternative anomaly score. These dual scoring mechanisms provide robustness against both structural and distributional anomalies, enabling the model to effectively detect unexpected machine sounds even under domain shifts or noisy conditions.

2.3. Data Augmentation via Foreground–Background Separation

To improve the generalization ability of anomaly detection models in the presence of complex and rare background conditions, a data augmentation method based on foreground–background separation using the TF-Locoformer model [16] is adopted. This approach leverages the model's strong source separation capabilities and its dual-path structure that captures both global and local patterns in the time–frequency (TF) domain. By recombining machine-related foreground sounds with diverse background recordings, it is possible to generate a wider variety of training data that remain acoustically realistic.

The TF-Locoformer is a Transformer-based network that integrates self-attention for long-range dependency modeling and convolutional feedforward modules for local structure learning. It performs separation of the input signal into two components: the foreground, which typically contains operational or anomalous machine sounds, and the background, which includes environmental noise, reverberation, and other irrelevant acoustic elements. The training of this separation model is conducted in a self-supervised manner, relying on pseudo-labels derived from domain-specific priors and reconstruction-based losses.

Once separation is completed, new training samples can be constructed by recombining background and foreground segments. Background components are collected from a wide range of machine types and environments to ensure diversity, while foreground segments are randomly selected from machine recordings containing either normal or anomalous events. These two components are mixed at various signal-to-noise ratios (SNRs) to simulate different operational conditions. In some cases, a mixup-inspired interpolation between different domains is applied to further expand the data distribution and improve robustness to domain shifts.

This data augmentation pipeline helps enrich the training dataset without requiring additional manual annotation, and contributes to improved anomaly detection performance, especially under mismatched or noisy background conditions. The TF-domain separation model ensures that critical spectral and temporal features are preserved during augmentation.

3. EXPERIMENTS

3.1. Data Preparation

We use the official training dataset provided by the DCASE challenge. For Approach 1, audio signals are first converted into log-Mel spectrograms using a sampling rate of 16 kHz, an FFT window size of 1024, a hop length of 512, and 128 Mel filter banks.For Approach 2, the same preprocessing pipeline is applied to extract log-Mel features. The model is trained for 50 epochs with a batch size of 2048 and a learning rate of 0.001.

3.2. Submitted Systems

We submit four systems in total, based on two different modeling approaches. System 1 and System 2 are built upon Approach 1, which employs a discriminative multi-modal structure that takes both raw waveform and Log-Mel spectrogram as input. The waveform and spectrogram are processed by BEATs and MobileFaceNet respectively, and their extracted features are fused into a 256-dimensional representation. This fused feature is used jointly for machine ID and domain classification, with a domain alignment module implemented via MMD loss to mitigate feature distribution discrepancies across modalities. The key difference between the two systems lies in the BEATs backbone: System 1 fine-tunes only the last four Transformer blocks and the positional convolution layer, while System 2 fully unfreezes the entire BEATs model for joint training. During inference, anomaly scores are computed based on GMM or Mahalanobis distance applied to the fused feature representation.

System 3 and System 4 are based on Approach 2, which adopts a reconstruction-based strategy. In these systems, audio features are extracted using BEATs ,then fed into a lightweight convolutional autoencoder. The training objective includes both reconstruction loss and classification loss to ensure that the learned latent

Table 1: Comparison of AUC and pAUC performance across different systems and machine types.

Machine Type	System 1		System 2		System 3		System 4	
	AUC(s/t)	pAUC	AUC(s/t)	pAUC	AUC(s/t)	pAUC	AUC(s/t)	pAUC
ToyCar	63.84 / 58.2	52.28	52.6 / 68.88	52.2	72.48 / 47.8	48.84	73.46 / 41.84	49.89
ToyTrain	73.28 / 64.28	51.98	73.84 / 60.96	52.14	46.78 / 50.36	48.74	40.76 / 49.08	48.26
bearing	71.76 / 71.12	55.16	68.88 / 67.72	55.16	59.94 / 62.58	63.21	59.28 / 60.52	61.84
fan	58.72 / 50.6	52.51	55 / 55.28	52.63	78.72 / 32.98	50.32	59.1 / 45.94	51.84
gearbox	73.24 / 82.64	54.02	72 / 78.24	53.88	74.3 / 49.62	54.53	70.92 / 50.48	52.68
slider	66.44 / 59.04	54.07	68.84 / 61.72	53.5	74.04 / 49.82	50.36	70 / 51.32	53.53
valve	96.68 / 79.04	53.71	94 / 76.6	54.49	52.58 / 55.98	51.31	53.1 / 56	51.68
All	69.21 / 56.65	56.66	68.18 / 56.4	59.56	60.94 / 50.74	54.84	61.33 / 47.69	51.19

representation captures both low-level signal consistency and high-level semantic discriminability. The primary difference lies in the anomaly scoring method: System 3 uses mean squared error (MSE) between input and reconstruction as the anomaly score, while System 4 combines MSE with Mahalanobis distance computed in the latent space between encoded features and their corresponding class centers. This hybrid scoring method enhances robustness to subtle anomalies.

4. REFERENCES

[1] D. G. Bhatt, P. U. Kyada, R. S. Rathore, M. Nallakaruppan, R. H. Jhaveri, *et al.*, "Enhancing anomaly detection in industrial control systems through supervised learning and explainable artificial intelligence." *Journal of Cybersecurity & Information Management*, vol. 15, no. 1, 2025.

[2] 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.

[3] B. Chen, W. A. Smith, Y. Cheng, F. Gu, F. Chu, W. Zhang, and A. D. Ball, "Probability distributions and typical sparsity measures of hilbert transform-based generalized envelopes and their application to machine condition monitoring," *Mechanical Systems and Signal Processing*, vol. 224, p. 112026, 2025.

[4] 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.

[5] 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.

[6] 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.

[7] Z. Wang, Z. Wang, X. Fan, and C. Wang, "Federated learning with domain shift eraser," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 4978–4987.

[8] H. Xu, S. Zhi, S. Sun, V. Patel, and L. Liu, "Deep learning for cross-domain few-shot visual recognition: A survey," *ACM Computing Surveys*, vol. 57, no. 8, pp. 1–37, 2025.

[9] 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.

[10] Y. Wang, Q. Zhang, W. Zhang, and Y. Zhang, "A lightweight framework for unsupervised anomalous sound detection based on selective learning of time-frequency domain features," *Applied Acoustics*, vol. 228, p. 110308, 2025.

[11] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," 2022.

[12] S. Xue, "Facial recognition for surveillance videos based on retinaface and mobilefacenet," in *The International Conference Optoelectronic Information and Optical Engineering (OIOE2024)*, vol. 13513. SPIE, 2025, pp. 878–883.

[13] S. F. Lima, E. Portmann, and L. Terán, "Fuzzyarcloss: Dynamic margin adjustment for robust recognition across domains," *Expert Systems With Applications*, p. 127477, 2025.

[14] S. Huang, Y. Zhang, Z. Fang, M. Tang, R. Xu, and L. He, "Anomalous sound detection using time-frequency feature and mixbatch," *Journal of Shanghai Jiaotong University (Science)*, pp. 1–8, 2025.

[15] Y. Chen, X. Ma, H. Qin, Y. Wang, H. Huang, and C. Xue, "Mahalanobis distance-based grey correlation analysis method for madm under q-rung orthopair hesitant fuzzy information on the lung cancer screening," *Expert Systems with Applications*, vol. 270, p. 126515, 2025.

[16] K. Saijo, G. Wichern, F. G. Germain, Z. Pan, and J. Le Roux, "Tf-locoformer: Transformer with local modeling by convolution for speech separation and enhancement," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept. 2024.