AUDIO DISMAE: UNSUPERVISED ACOUSTIC ANOMALY DETECTION VIA DISENTANGLED MASKED AUTOENCODER

Technical Report

Yuren Bian[1]*,JiayunChen*[2]

Tianjin Key Laboratory of Autonomous Intelligence Technology and Systems, Tiangong University, China

[1]yurenbian@outlook.com

[2]jiayunchen1@outlook.com

ABSTRACT

This technical report presents our submission to DCASE 2025 Task 2, which addresses unsupervised anomalous sound detection under domain shift conditions. We extend the Disentangled Masked Autoencoder (DisMAE), originally proposed for visual domain generalization, to the audio domain. In our approach, machine sounds are first transformed into log-Mel spectrograms and then fed into the DisMAE framework. The semantic branch is designed to reconstruct domain-invariant features, while the variational branch captures domain-specific attributes such as background noise and device variability. By disentangling these two representations, the model achieves robust reconstruction of normal operating sounds. Reconstruction errors from the primary decoder branch are used as anomaly scores. Experimental results demonstrate that the proposed method achieves promising performance on several machine types in the DCASE 2025 dataset.

Index Terms— DCASE, unsupervised anomalous sound detection, disentangled masked autoencoder

1. INTRODUCTION

The DCASE 2025 Task 2 focuses on first-shot unsupervised anomalous sound detection (ASD) under domain generalization conditions. Participants are required to detect machine anomalies without using any anomalous training data or machine-specific adaptation, and models must be capable of generalizing to unseen machine sections operating in previously unencountered environments.Compared to previous editions, the 2025 task introduces several notable updates. In particular, supplementary data are provided for each machine type, including clean machine sounds or background noise recordings, which may be optionally leveraged to enhance anomaly detection robustness in acoustically challenging or noisy conditions. In addition, participants are requested to report the computational complexity of their methods[1].

In this context, we explore the potential of domain-generalizable representation learning to improve anomaly detection under varying acoustic conditions. To this end, we propose Audio DisMAE, an adaptation of the Disentangled Masked Autoencoder (DisMAE)—a model originally developed for visual domain generalization—to the audio domain[2]. Audio DisMAE operates on log-Mel spectrograms derived from machine operating sounds. Its semantic branch aims to extract domain-invariant features, while its variational branch isolates domain-specific characteristics, such as noise or machine condition variability. By disentangling these two factors, the model enables robust reconstruction of normal operating sounds across domains. Anomaly scores are obtained from the main decoder via reconstruction error, and evaluated under DCASE 2025 Task2 protocols.

2. METHODS

2.1. Data Preprocessing

Given time-domain recordings of machine sounds, we first apply noise augmentation by mixing machine sounds with background noise samples provided in the supplemental dataset. This simulates realistic noisy environments and encourages the model to generalize under various acoustic conditions.

The augmented audio signals are then converted into log-Mel spectrograms with 128 Mel bands and a fixed number of frames. These spectrograms serve as the input representation for the Audio DisMAE model.

2.2. Model Architecture: Audio DisMAE

Audio DisMAE is adapted from the Disentangled Masked Autoencoder architecture originally developed for vision tasks. It consists of three branches:

Semantic encoder: Extracts domain-invariant features that are stable across different machine types and acoustic environments, enabling the model to generalize beyond specific domains.

Variational encoder: Models domain-specific variations such as environmental noise, machine-dependent characteristics, and other non-stationary acoustic factors, allowing for disentanglement from core semantic content.

Main decoder: Integrates the outputs from both the semantic and variational encoders to reconstruct the complete input spectrogram, facilitating precise anomaly scoring through reconstruction error.

To adapt to audio inputs, we modify the input tokenization to operate on log-Mel spectrogram patches, using 2D masking across both the time and frequency dimensions. This allows the model to learn feature representations sensitive to localized audio events.

The training objective minimizes the reconstruction loss over masked regions, encouraging the model to learn semantically meaningful latent representations while being invariant to domain-specific noise.

3. EXPERIMENTS

The performance of our system is summarized in Table 1. We employ the Area Under the Receiver Operating Characteristic Curve (AUC) to assess the overall detection capability, and the partial AUC (pAUC) to evaluate performance within a low false-positive rate (FPR) region [0, p], where p is set to 0.1.

We adopt two separate evaluation strategies for anomaly scoring. The first uses the Mean Squared Error (MSE) between the input and reconstructed Mel-spectrograms to assess reconstruction quality, which is then used as the anomaly score. The second evaluates the cosine distance between the latent representation of each test sample and the training data; specifically, the minimum cosine distance to any training sample in the normalized latent space is used as the anomaly score.

The baseline model used for comparison is the Simple Autoencoder provided by the official DCASE 2025 Task 2 baseline system[3-5].

Based on different model parameter configurations and anomaly scoring strategies, we submitted four systems for evaluation. All models are built upon the DisMAE architecture, with a key modification: the original random masking strategy is replaced by a structured time-frequency masking scheme more suitable for audio spectrograms. The detailed configurations of the submitted systems are summarized in Table 1.The test results are presented in Tables 2 and 3.

Table 1: Parameter settings for systems 1 to 4

	System 1	System 2	System 3	System 4
embed_dim	768	768	192	192
num_heads	12	12	3	3
decoder_depth	8	8	1	1
decoder_num_heads	16	16	8	8
evaluation method	MSE	Cosine	MSE	Cosine

Table 2: Results of System 1 and System 2 on the development set (%).

		Baseline(MSE)	System 1	System 2
ToyCar	AUC-S	71.05	41.51	49.06
	AUC-T	53.32	45.40	36.75
	pAUC	49.79	49.63	49.47
ToyTrain	AUC-S	61.76	50.92	40.29
	AUC-T	56.46	46.42	52.66
	pAUC	50.19	49.47	49.10
bearing	AUC-S	66.53	59.98	57.26
	AUC-T	53.15	61.16	49.84
	pAUC	61.12	65.44	55.66
fan	AUC-S	70.96	52.42	54.90
	AUC-T	38.75	45.10	47.62
	pAUC	49.46	49.94	50.26
gearbox	AUC-S	64.80	57.26	42.46
	AUC-T	50.49	56.32	45.64
	pAUC	52.49	54.05	51.84
slider	AUC-S	70.10	49.46	60.64
	AUC-T	48.77	50.27	54.86
	pAUC	52.32	51.42	52.58
valve	AUC-S	63.53	51.58	47.38
	AUC-T	67.18	58.93	47.74
	pAUC	57.35	49.26	49.10

Table 3: Results of System 3 and System 4 on the development set (%).

		Baseline(MSE)	System 3	System 4
ToyCar	AUC-S	71.05	36.80	62.58
	AUC-T	53.32	59.38	39.06
	pAUC	49.79	48.52	50.42
ToyTrain	AUC-S	61.76	48.62	55.20
	AUC-T	56.46	44.44	52.32
	pAUC	50.19	49.63	49.58
bearing	AUC-S	66.53	51.24	59.70
	AUC-T	53.15	48.56	54.60
	pAUC	61.12	49.16	54.26
fan	AUC-S	70.96	55.70	72.90
	AUC-T	38.75	49.56	23.46
	pAUC	49.46	50.94	49.94
gearbox	AUC-S	64.80	50.02	47.32
	AUC-T	50.49	55.58	58.00
	pAUC	52.49	49.53	49.00
slider	AUC-S	70.10	55.88	47.02
	AUC-T	48.77	54.60	53.28
	pAUC	52.32	51.16	49.84
valve	AUC-S	63.53	45.29	42.70
	AUC-T	67.18	52.30	45.78
	pAUC	57.35	51.74	47.95

4. REFERENCES

[1] Tomoya Nishida, Noboru Harada, Daisuke Niizumi, Davide Albertini, Roberto Sannino, Simone Pradolini, Filippo Augusti, Keisuke Imoto, Kota Dohi, Harsh Purohit, Takashi Endo, and Yohei 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] Zhang, A., Wang, H., Wang, X., Chua, TS. (2025). Disentangling Masked Autoencoders for Unsupervised Domain Generalization. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds) Computer Vision – ECCV 2024. ECCV 2024. Lecture Notes in Computer Science, vol 15128. Springer, Cham. https://doi.org/10.1007/978-3-031-72897-6_8

[3] Noboru Harada, Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Masahiro Yasuda, and Shoichiro 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), 1–5. Barcelona, Spain, November 2021.

[4] Kota Dohi, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, Masaaki Yamamoto, Yuki Nikaido, and Yohei 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.

[5] Noboru Harada, Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, and Masahiro Yasuda. First-shot anomaly detection for machine condition monitoring: a domain generalization baseline. Proceedings of 31st European Signal Processing Conference (EUSIPCO), pages 191–195, 2023.