Improving Anomalous Sound Detection with Top-M Pseudo-Labeling

Zhang Chen[1], Yucong Zhang[1,2], and Ming Li[1,2⋆]

[1] Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems,
Digital Innovation Research Center, Duke Kunshan University, Kunshan, China
[2] School of Computer Science, Wuhan University, Wuhan, China
ming.li369@dukekunshan.edu.cn

Abstract. Recent years have witnessed the growing popularity of discriminative self-supervised learning methods for learning effective machine sound embeddings in the Anomalous Sound Detection (ASD) task. These methods typically incorporate auxiliary tasks such as machine attribute classification to guide representation learning. In practice, when ground-truth metadata is limited, pseudo-labeling has emerged as an approach to supply supervisory labels for such auxiliary tasks. Existing pseudo-labeling strategies for ASD include approaches that assign a single most confident attribute to each instance. Complementary to these methods, this work explores a Top-M pseudo-labeling strategy to enhance discriminative self-supervised learning for ASD tasks. Instead of relying solely on the only attribute prediction, we incorporated the Top-M most confident attributes after pseudo-label generation to better capture semantic uncertainty. We proposed a categorization-based framework that adapts pseudo-labeling to attribute availability. Experiments were conducted on ASD benchmark datasets (DCASE 2020–2025) and the results on DCASE 2025 development test dataset showed that Top-M labeling consistently outperformed Top-1 across models. These results highlight the general effectiveness of our approach in improving ASD performance under limited supervision.

Keywords: Anomalous Sound Detection · Self-supervised learning · Discriminative learning· Pseudo-labeling

1 Introduction

Anomalous Sound Detection (ASD) is vital for early fault detection in industrial and manufacturing systems. The DCASE Challenge provides a widely used benchmark for ASD research, promoting unsupervised methods which restricts training to normal samples only, simulating real-world anomaly scarcity [11, 16].

In earlier iterations of DCASE Challenge, datasets for DCASE Challenge included rich metadata such as machine type, section ID, and operational attributes, which enabled fine-grained modeling and improved detection sensitivity [4, 10, 11]. In these challenges, self-supervised discriminative learning approaches that utilize additional meta-information achieved strong performance

⋆ Corresponding author: Ming Li

by learning fine-grained representations of normal machine behavior [21–23]. These methods typically train an encoder via auxiliary classification tasks (e.g., predicting machine IDs or attributes) and then use the learned embeddings for anomaly score. For example, in DCASE2023 Task 2, the first-place team utilized attribute information by incorporating machine ID and attribute labels as a joint category to train the classifier [9]; in DCASE 2024, the first-place team adopts the same discriminative learning methodology but using a pre-trained model to fine-tune the model on attribute information for better ASD performance [15].

However, starting in 2024, some of the machine attributes were removed [17]. This trend continues in this year's challenge, where discriminative methods must generalize to unseen machine types without relying on machine- or section-specific labels [18]. These changes also reflect real-world industrial scenarios where complete and reliable metadata may not always be available. As a result, ASD models must increasingly rely on semantic inference—inferring meaningful patterns from data itself—to learn representations that generalize across unseen machines and conditions.

To address the limited availability of labeled data in ASD, prior studies have explored two broad directions: generating pseudo data and assigning pseudo labels. Some works generate pseudo data by applying augmentations and processing original audio to create new samples, thereby improving model generalization [14]. Some studies use generative models (e.g., latent diffusion models [13]) to create new audio samples based on semantic cues, helping to expand the dataset. However, pseudo data generation might introduce discrepancies between synthetic and real-world data, potentially degrading performance on normal samples. As a result, pseudo-labeling is another promising idea as it preserves the integrity of the original data while effectively enriching training supervision.

Pseudo-labeling which involves assigning semantic labels (e.g., machine attributes) to unlabeled audio samples has emerged as a popular strategy [12,20]. However, some methods [12] assign only the single most confident class as the pseudo-label, which may propagate noisy supervision and ignore possible labels, while some may [20] rely on manually chosen confidence thresholds requiring considerable time and effort.

Facing the limitation, we propose a new pseudo-labeling method that integrates multiple candidate attributes into each pseudo-label. Our method fuses the Top-M highest-confidence labels, allowing the model to learn from contextual representations and become more robust to ambiguous or incomplete attribute information. We systematically evaluate this by analyzing model performance across different values of M.

2 Semantic Label Assignment by Attribute Availability

To develop targeted pseudo-labeling strategies, we first unify the DCASE Task 2 datasets from 2020 to 2025 [2,10,11,17,18] and design a set of label generation schemes for different machine types. As shown in Table 1, each strategy is chosen based on the structure and completeness of the machine attributes, aiming to generate plausible candidate pseudo-labels.

Table 1. Overview of labeling strategies

Strategy	Applicable Cases	Key Operation and Attribute Generation
Sample-wise	Fully labeled types with consistent semantics	Nearest neighbors in MFCC space; top attributes filtered by Random Forest (RF) confidence scores
Cluster-wise	Fragmented, partially labeled, or unlabeled types	KMeans clustering in latent space; attributes inferred based on proximity to labeled samples or cluster centers
Direct-use	Fully labeled types with clean and reliable annotations	Use ground-truth attributes directly; no pseudo-labeling required

Sample-wise Label Transfer: For machines with a relatively small and consistent set of attributes (e.g., *Bearing*, *Slider*, *ToyTrain*), we directly assign pseudo-labels based on nearest labeled neighbors in the MFCC feature space. Candidate attributes are determined by ranking the semantic labels of nearby labeled samples (via KNN) based on feature distance. Only high-confidence predictions (e.g. with >90% confidence) are retained to ensure quality.

Cluster-wise Label Propagation: When sample-wise transfer is impractical due to label inconsistency or absence of semantic attributes, we apply cluster-wise KMeans strategies using latent feature grouping. Candidate attributes are derived by measuring each sample's proximity to labeled samples or cluster centers within its cluster. Depending on the specific scenario:

- *Label-rich but fragmented types*(e.g., *Fan*): clustering is used to resolve conflicting attribute annotations.
- *Partially labeled types*(e.g., *ToyCar*): clusters propagate dominant labels to unlabeled samples via representative distance-based inference.
- *Fully unlabeled types* (e.g., *AirCompressor*): cluster IDs are directly used as pseudo-labels, where candidate labels are determined based on distance to multiple cluster centroids.

Direct Use of Ground-Truth Labels: For some machines (e.g., *BandSealer*) where annotations are clean and complete, we use the original labels directly without any pseudo-labeling or inference [18].

In summary, each strategy yields a ranked set of candidate attribute labels per sample reflecting varying degrees of semantic certainty. In subsequent sections, we investigate how integrating different numbers of top-ranked candidate labels influences model robustness and performance.

3 Top-M Attribute Fusion for Pseudo-Labeling

3.1 Hyperparameter Introduction

Three hyperparameters are defined to control the label propagation process in our pseudo-labeling framework:

- K_1: the number of nearest neighbors selected from the labeled training set,

- K_2: the number of clusters used in KMeans clustering,
- M: the number of top semantic labels to assign to each unlabeled sample.

Each of these parameters targets a different aspect of the pseudo-labeling framework. K_1 influences the locality and confidence of neighbor-based transfer in sample-wise pseudo-labeling, while K_2 determines the granularity of unsupervised clluster-wise grouping in scenarios when semantic labels are limited or absents. The parameter M controls how much semantic uncertainty is retained when constructing multi-attribute pseudo-labels.

To isolate the effect of M, we fix $K_1 = K_2 = 15$ across all experiments. This ensures that for all machines, we have 15 candidate attributes ranked by either prediction confidence or cluster proximity. This allows us to focus on evaluating how the number of fused attributes impacts the overall effectiveness and robustness of the pseudo-labels.

3.2 Top-M Label Construction

The top-M attributes are concatenated into a composite string to form a multi-attribute pseudo-label. For example, if the top-3 attributes are a, b, and c, the resulting pseudo-label is "a_b_c".

We empirically compare different values of M: $\{1, 3, 5, 7, 10, 13, 15\}$ using two distinct models, BEATS [1] and AudioMAE [8], to assess the consistency of our findings. The generated pseudo-labels are subsequently used as supervision targets in the downstream model training phase, enabling end-to-end learning under unlabeled conditions. This setup lays the foundation for a systematic analysis of how the number of fused attributes (M) influences anomaly detection performance across different machine types.

4 Experiment

4.1 Dataset

For pseudo-labeling and evaluation, we leverage the official development sets and additional training splits from the 2020–2025 DCASE Task 2 datasets, which contain machine operating sound recordings sourced from the ToyADMOS [7] and MIMII [3] datasets. Pseudo-labels are assigned only to training samples in the development and additional sets from the DCASE datasets. For model evaluation, we use the official test set from the DCASE 2025 dataset. All audio data are resampled to a consistent sampling rate of 16 kHz for feature extraction.

4.2 Model Architecture and Training Setup

In this paper, we use two pre-trained models as the backbone models, namely BEATs [1] and AudioMAE [8]:

BEATs: BEATs is initialized with pre-trained weights trained from AudioSet [6], and finetuned on the machine sound dataset. BEATs encoder consists of 12 Transformer layers, 768-dimensional hidden states, and 8 attention heads, resulting in 90M parameters.

AudioMAE: We re-implement AudioMAE on our own, which contains a 12-layer ViT-B [5] encoder and a 16-layer Transformer decoder. The model is first

Table 2. Selected evaluation results (%) for AudioMAE with $K_1 = K_2 = 15$. Only $M = 1, 3, 15$ are shown here.

M	**Metric**	**Bearing**	**Fan**	**Gearbox**	**Slider**	**ToyCar**	**ToyTrain**	**Valve**	**H. Mean**
1	sAUC	39.18	34.74	57.36	65.28	**72.08**	**65.66**	59.86	52.68
	tAUC	68.46	69.72	62.84	53.28	**64.82**	68.82	68.02	64.64
	ave. AUC	53.82	52.23	60.10	59.28	**68.45**	67.24	63.94	60.16
	pAUC	49.16	**50.11**	52.05	53.58	49.42	54.89	54.68	51.88
	H. Mean	49.61	47.55	57.08	56.87	60.56	62.53	60.37	55.84
3	sAUC	48.36	**38.40**	**62.82**	**66.90**	71.06	64.04	**60.08**	**56.49**
	tAUC	**70.84**	67.42	**71.00**	**64.14**	64.50	**73.26**	**72.54**	**68.92**
	ave. AUC	**59.60**	52.91	**66.91**	**65.52**	67.78	**68.65**	**66.31**	**63.47**
	pAUC	49.47	48.16	**56.26**	**55.47**	**53.11**	**55.47**	**66.37**	**54.39**
	H. Mean	54.54	**48.67**	**62.79**	**61.77**	**61.97**	**63.44**	**65.94**	**59.29**
15	sAUC	**48.82**	33.68	60.50	65.62	64.42	60.40	47.80	51.87
	tAUC	68.74	**72.34**	60.04	61.88	64.56	71.14	70.64	66.74
	ave. AUC	58.78	**53.01**	60.27	63.75	64.49	65.77	59.22	60.47
	pAUC	**53.11**	49.05	53.26	52.79	51.26	54.32	52.32	52.25
	H. Mean	**55.70**	46.95	57.74	59.59	59.38	61.19	55.36	56.18

pre-trained on all machine sounds by recovering the masked patches of the input spectrogram using a decoder. We use a mask ratio of 80% during the pre-training stage. For finetuning, we only take the pre-trained encoder part of the AudioMAE, and finetune the encoder using SpecAugment [19] by classifying the machine attributes.

For both models, the audio wave is converted to a mel-spectrogram with a window size of 25 ms, a hop size of 10 ms, and 128 Mel bins. The spectogram is then divided into 16x16 patches for further modeling. Both models are trained for 40 epochs, and optimized using AdamW with a base learning rate of 2×10^{-4}, weight decay of 1×10^{-5}, batch size 512, and cosine learning rate decay with 4 warm-up epochs. For finetuning, we use cross-entropy loss as the loss function for attribute classification task.

4.3 Evaluation Protocol

We evaluate model performance on DCASE 2025 development sets, strictly following the official scoring metrics [18]. Specifically, we report the following metrics per machine type:

Source AUC (sAUC). Area under the ROC curve for test samples drawn from the same domain as the training data (i.e., source domain).

Target AUC (tAUC). AUC for test samples drawn from target domain.

Average AUC (ave. AUC). Arithmetic mean of Source AUC and Target AUC, reflecting general performance.

Partial AUC (pAUC). AUC computed over a low false positive rate range $[0, 0.1]$ to highlight early detection capability.

We additionally compute the harmonic mean (H. Mean) across all machine types for each metric, providing a summary of performance across different domains. This value serves as the key comparison metric in our study, with a higher value indicating superior generalization across domains.

Table 3. Selected evaluation results (%) for BEATs with $K_1 = K_2 = 15$. Only $M = 1, 7, 15$ are shown here.

M	Metric	Bearing	Fan	Gearbox	Slider	ToyCar	ToyTrain	Valve	H. Mean
1	sAUC	57.98	**68.60**	41.86	65.48	64.10	41.20	38.94	**51.27**
	tAUC	**61.88**	37.44	42.48	51.06	57.96	66.18	61.74	52.02
	ave. AUC	**59.93**	**53.02**	42.17	58.27	61.03	**53.69**	50.34	53.31
	pAUC	50.21	48.32	49.05	49.26	49.74	**49.42**	50.84	49.54
	H. Mean	56.26	48.40	44.24	54.39	56.65	**50.32**	48.74	50.92
7	sAUC	**59.18**	41.36	54.70	**68.68**	**68.02**	24.94	**59.64**	48.12
	tAUC	59.36	**63.06**	**52.86**	61.66	60.02	77.80	**66.24**	**62.25**
	ave. AUC	59.27	52.21	53.78	**65.17**	64.02	51.37	**62.94**	**57.88**
	pAUC	**55.37**	**50.63**	**49.84**	**52.95**	48.42	49.05	**57.89**	**51.82**
	H. Mean	**57.91**	**50.18**	**52.39**	**60.40**	57.68	40.91	**61.05**	**53.43**
15	sAUC	57.58	32.98	**57.98**	68.54	66.06	25.58	50.20	45.55
	tAUC	56.62	61.36	50.52	**63.70**	**63.70**	**80.28**	63.58	61.44
	ave. AUC	57.10	47.17	**54.25**	64.95	**64.88**	52.93	56.89	56.26
	pAUC	54.79	48.47	48.74	49.53	**54.68**	49.05	53.05	51.05
	H. Mean	56.31	44.61	52.12	58.73	**61.07**	41.70	55.05	51.89

4.4 Experimental Results

As shown in Table 2 and 3 (results for more M values can be accessed in supplemental material), our experimental results demonstrate that the optimal number of pseudo-labels M varies across models. Specifically, the AudioMAE-based system achieves its highest performance when $M = 3$, with a highest harmonic mean of 59.29. For BEATs, the optimal result is observed at $M = 7$, where the H. Mean reaches 53.43%.

These findings suggest that directly selecting the Top-1 pseudo-label is not always the most effective strategy. Instead, leveraging top-ranked attributes enables the model to integrate a broader range of semantic representation and improve generalization across machine types and model architectures. Notably, different models prefer different values of M. We also observe that when M becomes too large, the performance may degrade and even fall below that of the Top-1 case. This suggests that incorporating excessive pseudo-labels may introduce noise or conflicting information. Therefore, Top-M labeling might need to be treated as a flexible and tunable strategy.

5 Conclusion

In this paper, our main contribution is to demonstrate the feasibility of using Top-M pseudo-labels for anomalous sound detection, with a focus on leveraging multiple high-confidence semantic attributes.. By integrating the Top-M most probable attributes into each pseudo-label, we aim to better capture attribute uncertainty and improve model robustness when labeled data is limited. Candidate attributes are obtained through a categorization-based labeling strategy, which tailors pseudo-labeling schemes based on the availability of attributes.

This framework allows us to control the number of top-ranked attributes included in each pseudo-label and examine how M impacts detection performance across different machine types and model backbones. We validate this approach on the DCASE 2025 development dataset, and our results confirm that Top-M approach achieves consistently higher harmonic mean AUC than Top-1 labeling. These improvements highlight the effectiveness of integrating candidate attributes rather than applying singular predictions.

Acknowledgements This research is funded in part by the Science and Technology Program of Suzhou City (SYC2022051). Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

References

1. Chen, S., Wu, Y., Zhou, C., Liu, S., Tompkins, D., Chen, Z., Che, W., Yu, X., Wei, F.: Beats: Audio pre-training with acoustic tokenizers. In: Proc. of ICML. pp. 5178–5193 (2023)
2. 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: Proc. of DCASE. pp. 31–35 (2023)
3. Dohi, K., Nishida, T., Purohit, H., Tanabe, R., Endo, T., Yamamoto, M., Nikaido, Y., Kawaguchi, Y.: MIMII DG: Sound dataset for Malfunctioning Industrial Machine Investigation and Inspection for Domain Generalization Task. In: Proc. of DCASE 2022 Workshop. pp. 1–5 (2022)
4. 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 DCASE 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques. In: Proc. of DCASE (2022)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. of ICLR (2021), `https://openreview.net/forum?id=YicbFdNTTy`
6. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: Proc. of ICASSP. pp. 776–780. IEEE (2017)
7. Harada, N., Niizumi, D., Takeuchi, D., Ohishi, Y., Yasuda, M., Saito, S.: ToyADMOS2: Another Dataset of Miniature Machine Operating Sounds for Anomalous Sound Detection under Domain Shift Conditions. In: Proc. of DCASE 2021 Workshop. pp. 1–5 (2021)
8. Huang, P.Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F., Feichtenhofer, C.: Masked autoencoders that listen. Advances in Neural Information Processing Systems **35**, 28708–28720 (2022)
9. Jie, J.: Anomalous sound detection based on self-supervised learning. Tech. rep., DCASE2023 Challenge (June 2023)

10. Kawaguchi, Y., Imoto, K., Koizumi, Y., Harada, N., Niizumi, D., Dohi, K., Tanabe, R., Purohit, H., Endo, T.: Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain-shifted conditions. In: Proc. of DCASE. pp. 186–190 (2021)
11. 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 DCASE 2020 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring. In: Proc. of DCASE. pp. 81–85 (2020)
12. Lee, D.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Proc. of ICML (2013)
13. Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., Plumbley, M.D.: Audioldm: Text-to-audio generation with latent diffusion models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proc. of ICML. pp. 21450–21474 (2023)
14. Liu, Y., Guan, J., Zhu, Q., Wang, W.: Anomalous sound detection using spectral-temporal information fusion. In: Proc. of ICASSP. pp. 816–820. IEEE (2022)
15. Lv, Z., Jiang, A., Han, B., Liang, Y., Qian, Y., Chen, X., Liu, J., Fan, P.: Aithu system for first-shot unsupervised anomalous sound detection. Tech. rep., DCASE2024 Challenge (June 2024)
16. Mesaros, A., Heittola, T., Virtanen, T.: A multi-device dataset for urban acoustic scene classification. In: Proc. of DCASE. pp. 9–13 (2018)
17. Nishida, T., Harada, N., Niizumi, D., Albertini, D., Sannino, R., Pradolini, S., Augusti, F., Imoto, K., Dohi, K., Purohit, H., Endo, T., Kawaguchi, Y.: Description and discussion on DCASE 2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring. In: Proc. of DCASE. pp. 111–115 (2024)
18. Nishida, T., Harada, N., Niizumi, D., Albertini, D., Sannino, R., Pradolini, S., Augusti, F., Imoto, K., Dohi, K., Purohit, H., Endo, T., Kawaguchi, Y.: Description and discussion on DCASE 2025 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring. arXiv preprint arXiv:2506.10097 (2025)
19. Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D., Le, Q.V.: Specaugment: A simple data augmentation method for automatic speech recognition. In: Proc. of Interspeech. pp. 2613–2617 (2019). https://doi.org/10.21437/Interspeech.2019-2680
20. Rizve, M.N., Duarte, K., Rawat, Y.S., Shah, M.: In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In: Proc. of ICLR. pp. 1–20 (2021)
21. Wilkinghoff, K.: Self-supervised learning for anomalous sound detection. In: Proc. of ICASSP. pp. 276–280. IEEE (2024)
22. Zhang, Y., Liu, J., Tian, Y., Liu, H., Li, M.: A dual-path framework with frequency-and-time excited network for anomalous sound detection. In: Proc. of ICASSP. pp. 1266–1270. IEEE (2024)
23. Zhang, Y., Suo, H., Wan, Y., Li, M.: Outlier-aware inlier modeling and multi-scale scoring for anomalous sound detection via multitask learning. In: Proc. of Interspeech. pp. 5381–5385 (2023)