Residual View and Prototype Selection for Noise-Aware Anomalous Sound Detection

Technical Report

JeongSik Kim, JongWoo Sung, HyeonJun Bae, BoRyeon Kim, JiAn Lee

LUDO Lab
Fundamental Deep Learning Research
Busan saha-gu, 49407, South Korea
{jskim, uturtle, gull, boryeon, jianee}@ludo-lab.com

ABSTRACT

In this paper we take an in-depth look at noise-aware unsupervised anomalous sound detection in a GenRep-style frozen embedding memory-bank framework using pretrained audio encoders. We propose Residual View, which subtracts a scaled far-channel embedding from the near-channel embedding. Additionally, we use a projection-residual prototype selection branch for the submitted systems. Furthermore, we analyze the effect of the residual coefficient and representation layer. We also benchmark the proposed view with several pretrained audio encoders. Our final submitted systems apply PRPS to Residual View and achieve 63.24 official score with SSLAM on the development set.

***Index Terms*—** anomalous sound detection, noise-aware machine monitoring, residual embedding

1. INTRODUCTION

The DCASE 2026 Task2 challenge focus on noise-aware unsupervised anomalous sound detection for machine condition monitoring. Unlike the previous task setting, this year's task provides two-channel recordings captured at different distances from the target machine. Participants may leverage synchronized recordings captured by microphones placed both near to and far from the target machine to help develop systems that are robust to background noise.

In real-world anomalous sound detection, it is difficult to collect anomalous samples or exhaustive anomaly patterns. Furthermore, the operating states of a machine or environmental noise can change and cause domain shifts. A bigger challenge is that, for unseen machine types, the hyperparameters of the trained model cannot be tuned. Additionally, annotated data are limited [1, 2, 3, 4].

The most typical approaches are reconstruction-based anomaly score system and Outlier Exposure based anomaly score systems [5, 6]. In particular, the latter approach is often used with audio encoder with fine-tuning [7, 8]. Moreover, pretrained audio encoders can be fine-tuned or used only as generic representations for downstream audio tasks [9, 10, 11, 12, 13]. GenRep uses only frozen embeddings with post-embedding processing such as temporal pooling and MemMixup. Even though it uses frozen embeddings, this type of system can achieve good performance compared with fine-tuned models [14, 15].

A GenRep-style frozen-embedding pipeline is effective because it uses generic representations from pretrained audio encoders, but its standard form does not directly handle near and a far two-channel recordings.

DCASE 2026 Task 2 provides two-channel recordings captured at different microphone distances. The far channel is recorded farther from the target machine, and it may contain a relatively different balance between the target machine sound and background noise. Based on this assumption we propose a residual embedding by subtracting a scaled far channel embedding from near channel within GenRep. We also use Projection-Residual Prototype Selection (PRPS) as a score-level memory-bank refinement. Furthermore, we perform an ablation study on the degree of far channel subtraction and check metric at each representation layer Finally, we evaluate the proposed method several pretrained audio encoders.

Our final submitted systems apply Residual View with PRPS. The best system achieves an Official Score of 63.24 on the development set, while the remaining systems examine encoder-ensemble, residual-only, and compact-encoder variants.

2. METHODS

We follow the GenRep-style frozen-embedding memory-bank framework, but replace the original single-channel representation with the proposed two-channel residual embedding.

In the DCASE 2026 Task 2 dataset, synchronized and real two-channel recordings are newly introduced. The near channel is recorded close to the target machine, while the far channel is recorded farther away from the target machine. We define the near channel as a sound that contains both target-machine sound and background noise. For the far channel, we assume that it contains not only background noise but also target-machine sound information, although less than the near channel. Based on this assumption, we examined how to compute the two channels so that useful information can be exploited without fine-tuning the existing generic representation.

2.1. Residual View

First, we extract embeddings for the near and far channels from the training audio clips using a pretrained audio encoder. Let f_{near} and f_{far} denote the temporally pooled embeddings extracted from the same encoder layer for the near and far channels, respectively. The f_{far} is not treated as noise-only representation. Instead, we regard it as an embedding that still contains target-machine information, but with relatively a larger contribution from background

noise components than f_{near}. Therefore, the strength of far-channel subtraction can be important. We define the residual embedding as:

$$f_{residual} = f_{near} - \alpha\, f_{far}$$

Here, alpha is the coefficient that controls how much of f_{far} is subtracted from f_{near}. The resulting $f_{residual}$ embeddings are stored in the memory bank as reference embeddings. For target-domain memory construction, we follow the GenRep strategy and apply MemMixup by mixing source and target $f_{residual}$ embeddings to augment the memory bank. The same residual transformation is applied to both reference training samples and test samples before computing k-nearest-neighbor distances. We do not apply domain normalization to avoid using evaluation-domain statistics at test time.

2.2. Projection-Residual Prototype Selection (PRPS)

After constructing the residual embedding, we add PRPS as a prototype-selection branch. PRPS first computes a projection-residual key:

$$q = f_{near} - P_{far}(f_{near})$$

where

$$P_{far}(f_{near}) = \left(\frac{(f_{near}\, f_{far})}{\left|\left|f_{far}\right|\right|^2}\right) f_{far}$$

The key q is used only to select 128 normal residual prototypes from the training memory. The selected prototype indices from q-space are applied to the corresponding residual embeddings, and kNN distances are still computed in the residual embedding space. We then compute two raw anomaly scores and average them: one from the full residual MemMix memory bank and one from the PRPS-selected residual prototypes. This prototype-selection step is inspired by memory-bank coreset selection [24].

3. EXPERIMENTS

The dataset comprises three subsets: development, additional training, and evaluation datasets. The development dataset includes seven machine types, each with one section containing 990 normal clips from a source domain, 10 normal clips from a target domain, and 200 labeled test clips (100 normal and 100 anomalous) with domain labels. Some machines also include attribute annotations. Each recording is a two-channel audio clip with a duration that varies across machine types. The recording contains both target-machine and environmental sounds captured at different distances from the target machine. Channel 1 is captured near the target machine, and channel 2 is captured farther away.

The additional training dataset introduces five new machine types, each with the same training structure, though attributes are provided for only some machines. The evaluation dataset includes test clips corresponding to the additional training machines, without labels, domain information, or attribute annotations. Participants are required to train models using only one section per machine type, without tuning on the test set or relying on attribute information.

Table 1: Submitted system configurations.

System	Encoder	PRPS	Feature Layers	Parameters
System 1	SSLAM	Yes	Last layer	89.97 M
System 2	BEATs iter3, DaSheng-base, SSLAM	Yes	6, 12 for BEATs, 1 for DaSheng, 6, 12 for SSLAM	265.73 M
System 3	SSLAM	No	Last layer	89.97 M
System 4	AudioMAE++ tiny	Yes	Last layer	21.00 M

Evaluation metrics. Performance under domain shift is evaluated using AUC, partial AUC (pAUC, p=0.1), and the Official Score, which is defined as the harmonic mean of source AUC, target AUC, and mixed pAUC across all machine types.

For implementation, we use 10-second audio clips. Shorter clips are padded, and longer clips are truncated. MemMixup uses $\lambda = 0.9$, and k-nearest-neighbor scoring is performed with $K = 1$. PRPS uses 128 selected residual prototypes as an additional score branch.

For Residual View, alpha is treated as a fixed hyperparameter. We search alpha across representation 1 to 12 layers on the development set using SSLAM as the reference encoder. The residual coefficient α is selected globally, not separately for each machine type. Based on the SSLAM development-set sweep, we fix $\alpha = 0.5$ for all submitted systems.

We use multiple large-scale pretrained audio encoders within the GenRep framework with Residual View and PRPS, denoted as follows: BEATs iter3 for BEATs [16], M2D-CLAP for M2D-CLAP [17], EAT large for EAT [18], SSLAM pretrained for SSLAM [19], CED tiny and CED base for CED [20], FISHER small for FISHER [21], AudioMAE++ tiny for AudioMAE++ [22], and DaSheng base and DaSheng-1.2B for DaSheng [23]. All encoder-level comparisons use the last representation layer. We avoid best-layer selection because it is not available in real evaluation scenarios.

For each pretrained audio encoder, we use its official preprocessing pipeline and input format. This is necessary because the compared encoders use different acoustic front ends, such as waveform-based or filter bank-based inputs. For a fair comparison, the same preprocessing, checkpoint, and representation layer are applied to both channels. All encoder parameters are frozen.

Table 1 shows the submitted systems. System 1 uses SSLAM and applies Residual View with PRPS to the last-layer representation. System 2 is a score-level ensemble of BEATs iter3, DaSheng-base, and SSLAM with the same Residual View and PRPS scoring rule. System 3 uses the same SSLAM Residual View as System 1 but disables PRPS as a conservative residual-only anchor. System 4 uses AudioMAE++ tiny with Residual View and PRPS as a compact frozen pretrained-encoder setting.

4. RESULTS

Table 2 compares the submitted systems, including PRPS-based systems, a residual-only anchor, and reference with the MAHALA baseline on the development set. System 1 achieves the best official score and the best target AUC among the the submitted systems. System 2 is included as a multi-encoder and

multi-representation Residual View and PRPS ensemble to reduce dependence on a single encoder and layer, although its official score is slightly lower than System 1. System 3 is the Residual View-only SSLAM anchor without PRPS, and System 4 uses AudioMAE++ tiny with Residual View and PRPS.

Table 2: Overall development-set results.

System	AUC Source	AUC Target	pAUC	Official Score
Baseline (MAHALA)	66.46	54.24	53.91	57.66
System 1	70.34	**68.14**	**57.80**	**63.24**
System 2	**71.00**	67.23	55.38	62.55
System 3	67.21	64.40	56.61	62.41
System 4	67.26	59.83	55.79	59.80

Table 3: Machine-level development-set results.

Machine / Metric	System 1	System 2	System 3	System 4
ToyCarEmu				
AUC source	60.46	62.40	58.14	**70.84**
AUC target	79.50	80.60	**82.82**	71.76
pAUC	49.47	50.00	49.74	**58.11**
Official Score	60.81	61.94	60.75	**66.28**
ToyCar				
AUC source	**84.16**	79.56	82.50	73.50
AUC target	77.84	**85.30**	79.92	70.10
pAUC	**61.79**	59.47	60.26	61.63
Official Score	**73.33**	72.98	72.77	68.03
bearingEmu				
AUC source	65.06	**66.80**	63.28	57.56
AUC target	**64.22**	60.78	60.32	59.54
pAUC	60.16	**60.53**	59.95	59.32
Official Score	**63.07**	62.57	61.15	58.79
fan				
AUC source	51.00	**58.04**	52.56	56.08
AUC target	46.94	45.40	45.52	**54.40**
pAUC	52.89	52.89	52.95	**53.21**
Official Score	50.15	51.58	50.10	**54.54**
gearboxEmu				
AUC source	70.16	74.72	70.70	**75.86**
AUC target	54.76	**54.94**	54.76	48.14
pAUC	52.95	52.95	53.11	**53.74**
Official Score	58.36	**59.44**	58.55	57.07
sliderEmu				
AUC source	**81.82**	80.68	77.74	74.00
AUC target	**68.44**	65.50	64.26	49.34
pAUC	**55.53**	55.05	55.37	51.95
Official Score	**66.90**	65.46	64.54	56.57
valveEmu				
AUC source	**79.74**	74.80	76.94	63.00
AUC target	**85.28**	78.08	85.24	65.52
pAUC	**71.84**	56.74	69.00	52.58
Official Score	**78.56**	68.49	76.49	59.82
All (hmean)				
AUC source	70.34	**71.00**	67.21	67.26
AUC target	**68.14**	67.23	64.40	59.83
pAUC	**57.80**	55.38	56.61	55.79
Official Score	**63.24**	62.55	62.41	59.80

Table 3 reports source AUC, target AUC, pAUC, and official score for each machine type on the development set. The scores

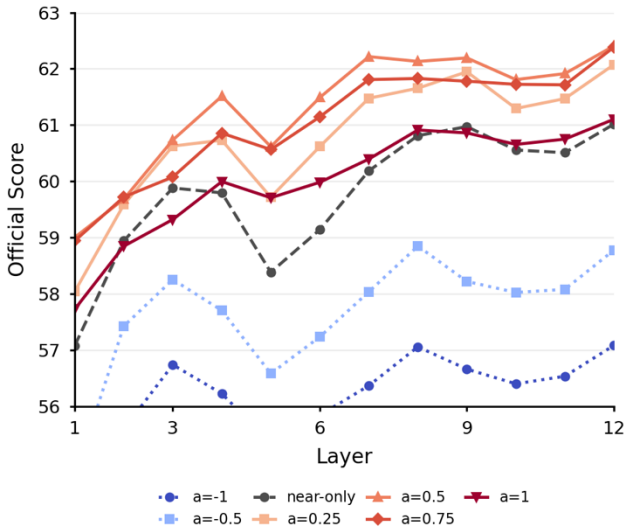

Figure 1: Layer-wise official score of SSLAM under the near-only baseline and Residual View with different residual coefficients.

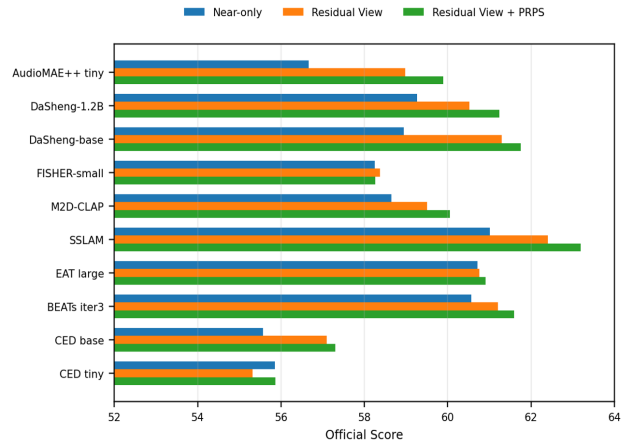

Figure 2: Official score comparison between near-only, Residual View, and Residual View with PRPS representations using the last representation layer of each pretrained audio encoder.

are not uniform across machines. System 1 is strongest overall and remains strong on ToyCar and valveEmu, while fan remains weak. System 2 improves target AUC on ToyCar and ToyCarEmu but loses pAUC and valveEmu score compared with System 1. System 3 is the Residual View-only SSLAM system without PRPS, retained as a conservative anchor for comparison. System 4 is lower overall, but it is retained as a compact AudioMAE++ tiny pretrained audio encoder.

Figure 1 compares the near-only representation and Residual View within the same GenRep-style framework. The α sweep shows that moderate far-channel subtraction, especially around α = 0.25–0.75, improves most SSLAM representation layers. At the last layer, Residual View improves the official score by 1.39 points and target-domain AUC by 2.13 points over the near-only System, reaching an official score of 62.4. When α is negative, the operation effectively adds the far-channel embedding, and

performance generally decreases. Therefore, we use $\alpha = 0.5$ as a fixed coefficient for the submitted systems.

Figure 2 shows the development-set official score when applying near-only, Residual View, and Residual View with PRPS to each pretrained audio encoder. For a consistent comparison, this figure uses the last representation layer. SSLAM gives the highest score among the tested encoders, improving from 61.02 with near-only to 62.41 with Residual View and 63.19 with Residual View with PRPS. AudioMAE++ tiny shows the largest gain from near-only, improving from 56.67 to 59.89. Overall, Residual View and PRPS improves the residual-memory score for most listed encoders, but the gain remains encoder dependent.

5. CONCLUSION

We introduced Residual View and PRPS, training-free two-channel methods for a GenRep-style frozen-embedding memory-bank framework. Residual View subtracts a scaled far-channel embedding from the near-channel embedding, and PRPS adds a projection-residual prototype-selection branch while preserving the full residual memory score. On the development set, System 1 with SSLAM achieves an Official Score of 63.24. However, the benefit still depends on the pretrained encoder, representation layer, and machine type.

6. REFERENCES

[1] T. Nishida, N. Harada, D. Takeuchi, D. Niizumi, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and discussion on dcase 2026 challenge task 2: Noise-aware unsupervised anomalous sound detection for machine condition monitoring," arXiv preprint arXiv:2606.01578, 2026.

[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," arXiv preprint arXiv:2106.02369, 2021.

[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," arXiv preprint arXiv:2205.13879, 2022.

[4] 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.

[5] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," arXiv preprint arXiv:1812.04606, 2018.

[6] T. Fujimura, I. Kuroyanagi, and T. Toda, "Improvements of discriminative feature space training for anomalous sound detection in unlabeled conditions," in ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025, pp. 1–5.

[7] 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.

[8] B. Han, Z. Lv, A. Jiang, W. Huang, Z. Chen, Y. Deng, J. Ding, C. Lu, W.-Q. Zhang, P. Fan, 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.

[9] 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," arXiv preprint arXiv:2406.11364, 2024.

[10] K. Wilkinghoff, H. Yang, J. Ebbers, F. G. Germain, G. Wichern, and J. Le Roux, "Keeping the balance: Anomaly score calculation for domain generalization," in ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025, pp. 1–5.

[11] B. Han, A. Jiang, X. Zheng, W.-Q. Zhang, J. Liu, P. Fan, and Y. Qian, "Exploring self-supervised audio models for generalized anomalous sound detection," IEEE Transactions on Audio, Speech and Language Processing, 2025.

[12] K. Wilkinghoff, S. Yadav, and Z.-H. Tan, "Temporal pooling strategies for training-free anomalous sound detection with self-supervised audio embeddings," arXiv preprint arXiv:2603.04605, 2026.

[13] L. Wang, "Pre-trained model enhanced anomalous sound detection system for DCASE2025 Task 2," DCASE2025 Challenge, Tech. Rep., June 2025.

[14] P. Saengthong and T. Shinozaki, "Deep generic representations for domain-generalized anomalous sound detection," in ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025, pp. 1–5.

[15] ——, "GENREP for first-shot unsupervised anomalous sound detection of DCASE 2025 challenge," DCASE2025 Challenge, Barcelona, Spain, Tech. Rep., 2025.

[16] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," arXiv preprint arXiv:2212.09058, 2022.

[17] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, M. Yasuda, S. Tsubaki, and K. Imoto, "M2D-CLAP: masked modeling duo meets CLAP for learning general-purpose audio-language representation," arXiv preprint arXiv:2406.02032, 2024.

[18] 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.

[19] T. Alex, S. Atito, A. Mustafa, M. Awais, and P. Jackson, "Sslam: Enhancing self-supervised models with audio mixtures for polyphonic soundscapes," in International Conference on Learning Representations, vol. 2025, 2025, pp. 22 608–22 626.

[20] H. Dinkel, Y. Wang, Z. Yan, J. Zhang, and Y. Wang, "CED: Consistent ensemble distillation for audio tagging," in ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024, pp. 291–295.

[21] P. Fan, A. Jiang, S. Zhang, Z. Lv, B. Han, X. Zheng, W. Liang, J. Li, W.-Q. Zhang, Y. Qian, et al., "Fisher: A foundation model for multi-modal industrial signal comprehensive representation," arXiv preprint arXiv:2507.16696, 2025.

[22] S. Yadav, S. Theodoridis, and Z.-H. Tan, “AudioMAE++: learning better masked audio representations with SwiGLU FFNs,” in 2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2025, pp. 1–6.

[23] H. Dinkel, Z. Yan, Y. Wang, J. Zhang, Y. Wang, and B. Wang, “Scaling up masked audio encoder learning for general audio classification,” arXiv preprint arXiv:2406.06992, 2024.

[24] K. Roth, L. Pemula, J. Zepeda, B. Scholkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.