PCL: Partitioned Continual Learning via Unsupervised Latent Experts for Audio Classification

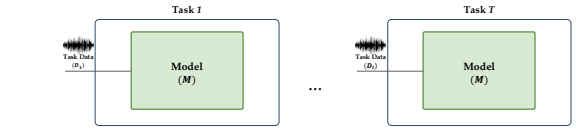
Gautham Krishna Gudur [*1] **Mohit Malu** [*2] **Tanmay Khandait** [2] **Reza Rahimi Azghan** [2] **Anirudh Rayas** [2] **Pavan Turaga** [2] **Joydeep Ghosh** [1] **Hassan Ghasemzadeh** [2] **Edison Thomaz** [1] **Giulia Pedrielli** [2]

Abstract

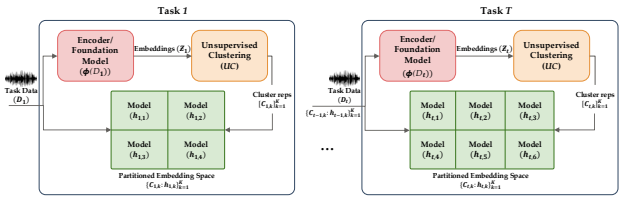
Continual audio classification requires models to learn new sound classes over time without retraining while preserving performance on previously learned classes, balancing *plasticity* and *stability*. Existing continual learning (CL) approaches update a single shared monolithic model across tasks, limiting scalability to evolving task distributions and constraining adaptation to the input space. We propose *PCL*, a representation-space framework that replaces a monolithic model with multiple lightweight experts. Embeddings from pretrained audio foundation models are partitioned via unsupervised clustering, where each homogeneous latent region is assigned an expert. Experiments on ESC-50 and UrbanSound8K using CLAP, AST, and Wav2Vec2 embeddings in exemplar-free class-incremental settings show consistent improvements over monolithic baselines, including a $\sim 17\%$ accuracy gain and a Backward Transfer (BWT) improvement from -0.249 to -0.061 on ESC-50 with minimal overhead.

1. Introduction

Audio classification is an important research area with applications including acoustic event detection (Imoto et al., 2013), environmental sound classification (Mu et al., 2021; Salamon & Bello, 2015), keyword spotting, and wake-word detection (Zhang et al., 2017; Sainath & Parada, 2015). Although modern deep neural networks achieve strong performance on these tasks (Tang & Lin, 2018), they are typically trained in static settings where all data are available simultaneously. In real-world audio systems, however, new classes, devices, and heterogeneities emerge continuously, making retraining computationally prohibitive while naive fine-tuning induces catastrophic forgetting of prior knowledge.



(a) Traditional Continual Learning.



(b) PCL: Partitioned Continual Learning with Latent Experts.

Figure 1. Illustration of our proposed continual learning approach with unsupervised partitions.

Continual learning (CL) addresses this challenge by enabling sequential learning while preserving previously acquired knowledge. Existing CL methods include regularization-based approaches such as Elastic Weight Consolidation (Kirkpatrick et al., 2017) and Learning without Forgetting (Li & Hoiem, 2018), rehearsal-based methods (Rolnick et al., 2019; Chaudhry et al., 2019), and architectural expansion techniques (Rusu et al., 2016). Exemplar-free regularization methods are particularly appealing as they preserve prior knowledge without storing past data. However, most existing approaches update a single shared representation across tasks, causing interference when incremental tasks arise from different underlying distributions.

Recent work has explored class-incremental and few-shot audio classification (Si et al., 2024; 2025; Li et al., 2023; 2024), as well as exemplar-free methods such as AFT (Chen et al., 2026), AnalyticKWS (Xiao et al., 2025), and UCIL (Xiao & Das, 2025). Nevertheless, these approaches still operate within a shared representation space. In many audio applications, distribution shifts arise from variations in devices, acoustic environments, and sound categories. Classical partitioning methods, including

[*]Equal contribution [1]The University of Texas at Austin [2]Arizona State University. Correspondence to: Gautham Krishna Gudur <gauthamkrishna@utexas.edu>.

Learning to Listen: ICML 2026 Workshop on Machine Learning for Audio

Gaussian processes (Malu et al., 2023; Heinonen et al., 2016), non-stationary kernels (Fuentes & Smith, 2001), and tree-based models, address such heterogeneity in the input space. However, audio input features are often highly entangled, where perceptually similar sounds may differ significantly in their raw representations and vice versa.

Instead, we leverage partitions in the latent representation space learned by self-supervised audio foundation models. Embeddings from models such as Audio Spectrogram Transformer (AST) (Gong et al., 2021), Contrastive Language–Audio Pretraining (CLAP) (Elizalde et al., 2023), and Wav2Vec2 (Baevski et al., 2020) often exhibit locally separable regions corresponding to distinct classes or acoustic conditions, forming a structured geometry where semantic similarity between sounds and environmental factors induces meaningful subspaces, unlike highly entangled input-space features (see Appendix A Fig. 2). We identify these regions via unsupervised partitioning to align continual learning with naturally occurring, locally homogeneous sub-distributions. Concretely, we (i) map inputs to a fixed latent space using a pretrained encoder, (ii) identify latent partitions through unsupervised clustering and assign lightweight expert models, and (iii) train the experts sequentially with partition-aware consolidation to preserve prior knowledge while adapting to new data. Restricting updates to partition-specific experts reduces cross-task interference and unnecessary parameter updates. At first glance, this approach may resemble mixture-of-experts models (Shazeer et al., 2017; Jacobs et al., 1991); however, our method uses unsupervised latent partitions rather than learned gating and is designed for continual adaptation rather than scaling static model capacity. The scientific contributions of this work are as follows:

(1) We propose *PCL*: a framework for exemplar-free continual audio classification based on unsupervised partitioning of latent representations from audio foundation models. **(2)** We develop a clustering mechanism to identify latent partitions and assign lightweight expert models to each region. **(3)** We provide extensive empirical evaluation on two public audio benchmarks, demonstrating consistent improvements over strong class-incremental continual learning baselines.

2. Proposed Approach

We propose a partition-aware continual learning framework that replaces each task into locally homogeneous regions in a frozen latent space and assigns specialized experts to each region. The framework comprises four components: (1) frozen representation extraction, (2) memory-aware weighted clustering, (3) one-to-one expert assignment, and (4) partitioned continual adaptation. Algorithm 1 provides an overview of the framework. We first formalize the problem setting before describing the method in detail.

Algorithm 1 Partitioned Continual Learning (PCL)

1: **Input.** Tasks T, data stream $\{\mathcal{D}_t\}_{t=1}^{T}$, encoder ϕ, similarity metric $S(\cdot,\cdot)$, similarity threshold τ, clustering routine $\textsc{KMedoids}(\cdot,\tau)$.
2: Initialize: (Cluster reps) $\mathcal{C}_0 \leftarrow \varnothing$, (experts) $H_0 \leftarrow \varnothing$
3: **for** $t = 1$ **to** T **do**
4: **Learn latent representations.**
5: $Z_t \leftarrow \{\phi(x) \mid (x,y) \in \mathcal{D}_t\}$
6: **Memory Aware Weighted Clustering.**
7: $\tilde{Z}_t \leftarrow Z_t \cup \mathcal{C}_{t-1}$
8: $\mathcal{C}_t = \{c_{t,j}\}_{j=1}^{K_t} \leftarrow \textsc{KMedoids}(\tilde{Z}_t, \tau)$ (Weighted)
9: **Assign Experts.**
10: Initialize $\mathcal{K}_{t-1} = \{\}$ (index of the assigned experts)
11: **for** $k = 1$ to K_t **do**
12: $k' \leftarrow \arg\max_{j \in [K_{t-1}] - \mathcal{K}_{t-1}} S(c_{t,k}, c_{t-1,j})$
13: $\mathcal{K}_{t-1} \leftarrow \mathcal{K}_{t-1} \cup \{k'\}$
14: **if** $k' \neq$ None **then**
15: $h_{t,k} \leftarrow h_{t-1,k'}$
16: **else**
17: All previous experts are assigned
18: Initialize new expert $h_{t,k}$
19: **end if**
20: **end for**
21: **Partitioned Continual Learning.**
22: **for** $k = 1$ **to** K_t **do**
23: Train $h_{t,k}$ on $\mathcal{D}_{t,k}$ by minimizing $\mathcal{L}_{t,k}$
24: **end for**
25: $H_t \leftarrow \{h_{t,k}\}_{k=1}^{K_t}$
26: **end for**
27: **Output:** $H_T, \mathcal{C}_T$
28: **Inference:** For a test sample x, let $z = \phi(x)$, $k^* = \arg\max_{k \in [K_T]} S(z, \mu_{T,k})$, $\hat{y} = h_{T,k^*}(x)$

Problem Formulation. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a dataset of audio features $x_i \in \mathbb{R}^L$ with labels $y_i \in 1, \ldots, C$. We consider a class-incremental continual learning setting with a sequential stream of tasks $\{\mathcal{D}_t\}_{t=1}^{T}$, where each task arrives incrementally and data from previous tasks are not retained. The objective is to learn new tasks while preserving performance on previously learned tasks.

Frozen Latent Representation. For each task $\mathcal{D}_t = \{(x_i, y_i)\}$, we compute latent embeddings using a pretrained audio foundation model $\phi(\cdot)$ (e.g. AST, CLAP, or Wav2Vec2; see Sec. 3.2), as shown in line 4 of Algorithm 1. The encoder remains fixed across all tasks. Each input is mapped as $z_i = \phi(x_i)$, where $z_i \in \mathcal{Z} \subset \mathbb{R}^d$. Operating in this frozen latent space stabilizes partition structure across tasks and prevents representation drift.

Memory-Aware Weighted Clustering. For task t (line 5 in Algorithm 1), we form the latent set $\mathcal{Z}_t = \{\phi(x) \mid (x,y) \in \mathcal{D}_t\}$. Unlike standard task-wise clustering, we retain the pre-

vious cluster representatives $\{c_{t-1,k}\}_{k=1}^{K_{t-1}}$ and their sample counts $\{n_{t-1,k}\}_{k=1}^{K_{t-1}}$, which serve as compact memories of previously discovered latent regions. We then perform weighted clustering (e.g. weighted k-medoids) over (i) new embeddings $z_k \in \mathcal{Z}_t$ with unit weight and (ii) previous representatives $c_{t-1,k}$ weighted proportional to $n_{t-1,k}$. We use k-medoids following prior work on detecting new audio classes (Gudur & Perepu, 2021; Shuyang et al., 2017). This biases clustering toward preserving prior regions while allowing new clusters to emerge in novel latent regions.

We initialize with a large number N_{clusters} and iteratively merge clusters whose representatives satisfy $S(c_{t,i}, c_{t,j}) > \tau$. The resulting K_t clusters are denoted $\{\mathcal{Z}_{t,k}\}_{k=1}^{K_t}$ with representatives $c_{t,k}$. Each cluster induces a subset $\mathcal{D}_{t,k} = \{(x, y) \in \mathcal{D}_t \mid \phi(x) \in \mathcal{Z}_{t,k}\}$. The updated statistics $\{c_{t,k}, n_{t,k}\}$ are stored for the next task.

Assigning Experts. (Lines 9–18 of Algorithm 1) Let $\{h_{t-1,k}\}_{k=1}^{K_{t-1}}$ denote the experts after task $t-1$, each associated with representative $c_{t-1,k}$. We compute similarity scores $S(c_{t,i}, c_{t-1,j})$ and perform one-to-one matching between new clusters and previous experts. If the similarity between $c_{t,i}$ and its closest previous representative exceeds a threshold τ, the corresponding expert is reassigned to cluster k. Each expert is matched to at most one cluster, while unmatched clusters are assigned newly instantiated experts. Consequently, the number of experts may grow over time as novel latent regions emerge.

Partitioned Continual Learning. Each expert $h_{t,k}$ is trained independently on its corresponding subset $\mathcal{D}_{t,k}$ (lines 21–25 in Algorithm 1) using,

$$\mathcal{L}_{t,k} = \sum_{(x,y)\in\mathcal{D}_{t,k}} \ell_{\text{CE}}\big(h_{t,k}(x), y\big) + \lambda\, \ell_{\text{CL}}\big(\theta^{(t,k)}\big) \quad (1)$$

where ℓ_{CE} denotes cross-entropy loss and ℓ_{CL} could be any continual learning regularizer, with specific choices described in Sec. 3.3. Since experts are updated independently, consolidation remains localized to relevant partitions, reducing interference across unrelated latent regions. The framework is agnostic to the choice of regularizer.

Partition-Aware Inference. At inference time, a test sample x is embedded as $z = \phi(x)$. The nearest cluster representative $c_{t,k}$ in $\mathcal{Z}$ is identified, and the corresponding expert produces the prediction $\hat{y} = h_{t,k}(x)$. Only a single expert is activated per input, ensuring computational efficiency while preserving structural specialization.

3. Experimental Setup

3.1. Datasets

In this paper, we evaluate the proposed framework on public audio classification benchmarks for acoustic event detection and environmental sound classification. **ESC-50** contains 2,000 environmental audio clips spanning 50 classes, with 40 clips per class and a duration of 5 seconds each. **UrbanSound8K** comprises 8,732 labeled urban sound excerpts (≤ 4 seconds) across 10 classes.

3.2. Foundation Models for z-space embeddings

Audio Spectrogram Transformer (AST) is a transformer-based model that learns audio representations from spectrogram patches through self-attention (Gong et al., 2021).
Contrastive Language-Audio Pretraining (CLAP) is a multimodal foundation model that learns joint audio–text representations through large-scale contrastive pretraining (Elizalde et al., 2023).
Wav2Vec2 is a self-supervised model that learns contextual speech representations from raw waveforms via masked prediction (Baevski et al., 2020).

3.3. Baselines

Joint training trains a single model on aggregated data from all tasks simultaneously, serving as an upper-bound reference without continual learning constraints.

Finetune sequentially trains a single model on each task without preserving prior knowledge, resulting in catastrophic forgetting and serving as a lower bound.

Elastic Weight Consolidation (EWC) mitigates forgetting by regularizing updates to parameters important for previous tasks using a Fisher information–based penalty (Kirkpatrick et al., 2017).

Learning from Forgetting (LwF) mitigates forgetting by distilling predictions from the previous model while training on new tasks without storing past data (Li & Hoiem, 2018).

3.4. Metrics

Accuracy (ACC). We report the final test accuracy across all tasks after training on the final task.

Backward Transfer (BWT). Backward transfer quantifies retention (or forgetting) on earlier tasks after learning subsequent ones. It is computed as the mean change in accuracy on past tasks between their completion and the end of training.

$$\text{BWT} = \frac{1}{T-1} \sum_{t=1}^{T-1} \big(a_{T,t} - a_{t,t}\big)$$

where, $a_{t,t}$ is the accuracy on task t immediately after finishing training on t, and $a_{T,t}$ is the final accuracy on t after all T tasks. Positive BWT indicates beneficial transfer, and negative values indicate forgetting.

Table 1. Class-incremental results on ESC-50 and UrbanSound8K using CNN-2-layer (CNN-2L) and CNN-4-layer (CNN-4L) models with different CL methods and z-spaces. Higher ACC and BWT indicate better performance (↑).

z-space	CL Technique	ESC-50						UrbanSound8K					
		CNN-2L			CNN-4L			CNN-2L			CNN-4L		
		ACC↑	BWT↑	Clusters	ACC↑	BWT↑	Clusters	ACC↑	BWT↑	Clusters	ACC↑	BWT↑	Clusters
–	Finetune	23.254	-0.295	–	23.848	-0.286	–	23.240	-0.379	–	23.850	-0.368	–
–	EWC	37.682	-0.252	–	38.416	-0.249	–	27.948	-0.351	–	28.420	-0.348	–
	LwF	38.275	-0.238	–	38.142	-0.235	–	28.726	-0.345	–	29.285	-0.341	–
–	Joint (Upper bound)	**68.750**	–	–	**69.284**	–	–	**86.320**	–	–	**87.204**	–	–
CLAP	EWC	50.846	-0.082	24	52.392	-0.076	24	64.850	-0.124	8	66.420	-0.119	8
	LwF	52.518	-0.073	24	52.174	-0.073	24	65.370	-0.118	8	67.187	-0.118	8
AST	EWC	**54.924**	**-0.066**	28	**55.318**	**-0.061**	28	67.940	-0.118	9	68.524	-0.116	9
	LwF	54.285	-0.064	28	54.436	-0.065	28	**68.634**	**-0.115**	9	**70.249**	**-0.114**	9
Wav2Vec2	EWC	42.713	-0.198	15	42.286	-0.187	15	53.180	-0.187	11	55.260	-0.181	11
	LwF	43.052	-0.191	15	44.917	-0.204	15	54.057	-0.184	11	54.420	-0.184	11

3.5. Implementation Details

We evaluate our method using 2-layer CNN and 4-layer CNN experts across 10 random seeds, training each task for 50 epochs. Experiments are conducted under a class-incremental setting with 6 sequential tasks on UrbanSound8K (10 classes: 5 base + 1 per task) and ESC-50 (50 classes: 25 base + 5 per task). Latent embeddings are extracted from pretrained CLAP, AST, and Wav2Vec2 encoders. We use sweep-selected regularization scales for EWC ($\lambda = 10^7$) and LwF ($\lambda = 3$), and perform clustering using Euclidean distance with sweep-selected thresholds of 1, 30, and 4 for CLAP, AST, and Wav2Vec2, respectively.

4. Results

Table 1 presents class-incremental continual learning results for the proposed partition-aware framework on ESC-50 and UrbanSound8K. Across both datasets, latent-space partitioning consistently outperforms regularization-based baselines (EWC, LwF), particularly with stronger foundation model embeddings such as CLAP and AST. On ESC-50, AST-based partitioning improves ACC from $\sim$38% under standard EWC/LwF to $\sim$55%, while reducing forgetting from strongly negative BWT values (~ -0.25) to near-zero values (~ -0.06). Similar trends are observed on UrbanSound8K, where AST- and CLAP-based partitioning consistently achieve higher ACC and improved BWT. These results suggest that semantically structured latent spaces naturally decompose into reusable sub-regions that can be effectively exploited through expert specialization.

The effectiveness of the proposed framework strongly depends on representation quality. AST embeddings consistently yield the highest ACC and lowest forgetting across both datasets, indicating that transformer-based audio representations induce more separable and transferable latent partitions. In contrast, Wav2Vec2 embeddings exhibit weaker gains and more negative BWT, likely reflecting weaker semantic alignment between speech-oriented representations and environmental sound structure, thereby limiting effective clustering and expert reuse. Notably, CNN-4L models do not consistently outperform CNN-2L models, indicating that representation partitioning contributes more to continual learning performance than increased classifier capacity, which may overfit under limited partition-specific data. The moderate number of threshold-driven clusters discovered (e.g., 8–11 for UrbanSound8K and 24–28 for ESC-50) further suggests that PCL captures stable, locally homogeneous regions reusable across tasks, while balancing specialization with modest expert growth. Overall, these findings demonstrate that partition-aware continual learning improves both plasticity and stability while providing a scalable alternative to monolithic continual learning for audio classification.

5. Conclusion

We introduced *PCL*, a partition-aware continual learning framework that leverages latent representations from pretrained audio foundation models to partition the representation space into locally homogeneous regions and assign specialized experts to each region. By performing continual updates within these partitions, the framework reduces cross-task interference while preserving prior knowledge without storing past data. Experiments on ESC-50 and UrbanSound8K using CLAP, AST, and Wav2Vec2 embeddings demonstrate consistent improvements over strong exemplar-free baselines in class-incremental settings. These findings highlight the potential of exploiting latent-space structure for scalable continual audio learning, and suggest promising directions for extending partition-aware learning to broader multimodal and real-world streaming scenarios.

References

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P. K., Torr, P. H., and Ranzato, M. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

Chen, X., Chen, X., Weng, Z., and Xiao, Y. Aft: An exemplar-free class incremental learning method for environmental sound classification. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 15577–15581, 2026.

Elizalde, B., Deshmukh, S., Ismail, M. A., and Wang, H. Clap learning audio concepts from natural language supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

Fuentes, M. and Smith, R. L. A new class of nonstationary spatial models. Technical report, North Carolina State University. Dept. of Statistics, 2001.

Gong, Y., Chung, Y.-A., and Glass, J. AST: Audio Spectrogram Transformer. In *Interspeech 2021*, pp. 571–575, 2021.

Gudur, G. K. and Perepu, S. K. Zero-Shot Federated Learning with New Classes for Audio Classification. In *Interspeech 2021*, pp. 1579–1583, 2021. doi: 10.21437/Interspeech.2021-2264.

Heinonen, M., Mannerström, H., Rousu, J., Kaski, S., and Lähdesmäki, H. Non-stationary gaussian process regression with hamiltonian monte carlo. In *Artificial intelligence and statistics*, pp. 732–740. PMLR, 2016.

Imoto, K., Shimauchi, S., Uematsu, H., and Ohmuro, H. User activity estimation method based on probabilistic generative model of acoustic event sequence with user activity and its subordinate categories. In *INTERSPEECH*, pp. 2609–2613, 2013.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Li, Y., Cao, W., Xie, W., Li, J., and Benetos, E. Few-shot class-incremental audio classification using dynamically expanded classifier with self-attention modified prototypes. *IEEE Transactions on Multimedia*, 26:1346–1360, 2023.

Li, Y., Li, J., Si, Y., Tan, J., and He, Q. Few-shot class-incremental audio classification with adaptive mitigation of forgetting and overfitting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2297–2311, 2024.

Li, Z. and Hoiem, D. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

Malu, M., Pedrielli, G., Dasarathy, G., and Spanias, A. Class gp: Gaussian process modeling for heterogeneous functions. In *International Conference on Learning and Intelligent Optimization*, pp. 408–423. Springer, 2023.

Mu, W., Yin, B., Huang, X., Xu, J., and Du, Z. Environmental sound classification using temporal-frequency attention based convolutional neural network. *Scientific Reports*, 11(1):21552, 2021.

Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., and Wayne, G. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., and Hadsell, R. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Sainath, T. N. and Parada, C. Convolutional neural networks for small-footprint keyword spotting. In *Interspeech*, pp. 1478–1482, 2015.

Salamon, J. and Bello, J. P. Unsupervised feature learning for urban sound classification. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 171–175. IEEE, 2015.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Shuyang, Z., Heittola, T., and Virtanen, T. Active learning for sound event classification by clustering unlabeled data. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 751–755, 2017.

Si, Y., Li, Y., Li, J., Tan, J., and He, Q. Fully few-shot class-incremental audio classification using expandable dual-embedding extractor. In *Interspeech 2024*, pp. 4788–4792, 2024.

Si, Y., Li, Y., Tan, J., He, Q., and Kwak, I.-Y. Fully few-shot class-incremental audio classification using multi-level embedding extractor and ridge regression classifier. In *Interspeech 2025*, pp. 1318–1322, 2025.

Tang, R. and Lin, J. Deep residual learning for small-footprint keyword spotting. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5484–5488, 2018.

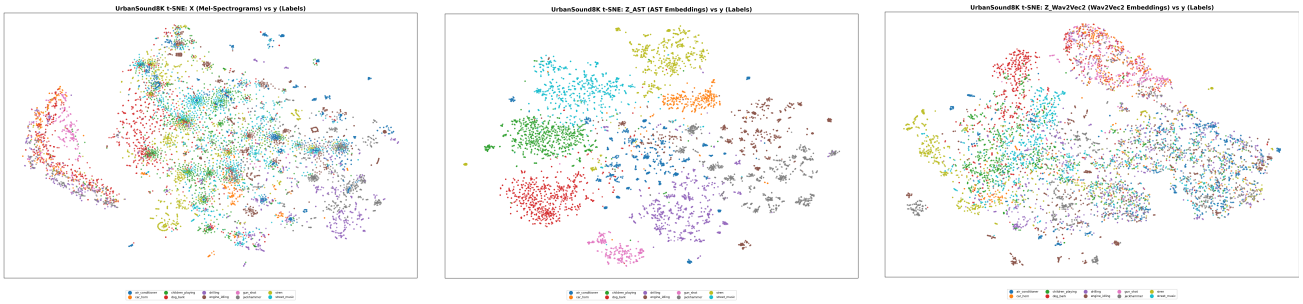
Xiao, Y. and Das, R. K. Ucil: An unsupervised class incremental learning approach for sound event detection. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025.

Xiao, Y., Tianyi, P., Das, R. K., Hu, Y., and Zhuang, H. Analytickws: towards exemplar-free analytic class incremental learning for small-footprint keyword spotting. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14147–14158, 2025.

Zhang, Y., Suda, N., Lai, L., and Chandra, V. Hello edge: Keyword spotting on microcontrollers. *arXiv preprint arXiv:1711.07128*, 2017.

A. t-SNE visualizations of mel-spectrograms and embeddings

Figure 2 visualizes the representation geometry of UrbanSound8K using t-SNE projections as an example. Compared to the input Mel-spectrogram space (X), embeddings from pretrained foundation models exhibit substantially improved class separability and more compact semantic structure, with AST showing slightly clearer separation than Wav2Vec2. This observation suggests that pretrained latent representations induce locally homogeneous regions that are well suited for partition-aware continual learning.



(a) X (Mel-spectrograms) vs Labels (y) *(b)* Z-space (AST) vs Labels (y) *(c)* Z-space (Wav2Vec2) vs Labels (y)

Figure 2. t-SNE visualizations of UrbanSound8K representations. Foundation model embeddings (Z spaces) exhibit clearer class separability than the input Mel-spectrogram space (X), with AST showing the most compact structure. Labels (y) are shown by color.