Learning order matters in class-incremental learning for sound localization and detection

Ruchi Pandey, Manjunath Mulimani, Archontis Politis, and Annamaria Mesaros
Signal Processing Research Centre, Tampere University, Finland
{ruchi.pandey, manjunath.mulimani, archontis.politis, annamaria.mesaros}@tuni.fi

***Abstract*—This study investigates the impact of class learning order in Class-Incremental Learning (CIL) for Sound Event Localization and Detection (SELD) by systematically evaluating class-wise localization error (LE) and F1-score across different class-ordering scenarios. A continual learning model is trained in two stages: initially on nine classes, then incrementally extended with four additional classes that vary in acoustic complexity. The results show that strategically introducing acoustically challenging (difficult to recognize) classes in the incremental learning stage enhances overall SELD performance, leading to increased F1-scores and reduced LE compared to a baseline that learns all the classes simultaneously. Furthermore, this study compares performance across balanced and imbalanced datasets, demonstrating consistent trends and highlighting the critical influence of class order. The study offers insights for designing more robust CIL frameworks for the SELD task.**

***Index Terms*—Class-incremental learning, Sound event detection and localization (SELD), convolutional recurrent neural network (CRNN).**

I. Introduction

Sound Event Localization and Detection (SELD) is a crucial array signal processing task that involves simultaneous detection and spatial localization of sound events when they occur [1]; the task is essential in applications such as robotics, surveillance, and smart environments, where precise detection and localization enhance situational awareness and decision-making [1]–[4]. Recent advancements in deep learning have significantly improved SELD models' accuracy [1], [5]–[7]. However, these models are traditionally trained on a fixed set of sound classes, limiting their adaptability in dynamic environments where new sound classes must be integrated over time. Retraining models from scratch each time new classes are introduced is computationally expensive and impractical. A common alternative is fine-tuning, where a pretrained model is updated on new data [8]. However, fine-tuning often leads to catastrophic forgetting, degrading performance on previously learned classes when incorporating new ones [9], [10].

To address this challenge, Continual Learning (CL) offers a promising framework that enables models to incrementally learn new tasks while preserving previously acquired knowledge [11], [12]. Class-Incremental Learning (CIL), a subtype of CL, specifically focuses on sequentially integrating new classes without full retraining [13], [14]. While CIL has been effectively applied in fields like computer vision [15], natural language processing [16], and audio tasks such as acoustic scene classification and keyword spotting [17]–[19], few studies have explored its impact on SELD tasks [20], [21].

This work leverages CIL to SELD by systematically analyzing how incremental learning affects class-wise localization error (LE) and F1-score. In our previous study, we concentrated solely on developing a class-incremental learning-based SELD model [21]. In this work, we investigate the impact of class training order and their varying levels of training difficulty. A continual learning model is trained in two stages: first on an initial set of nine sound classes, followed by an incremental stage where four new classes–varying in acoustic complexity–are introduced. The CIL-SELD model employs a mean square error (MSE)-based distillation loss, which minimizes discrepancies between the outputs corresponding to the previously learned classes in successive learners. Our experiments examine different class-combination strategies, evaluating their impact on performance retention and knowledge transfer. This approach efficiently expands SELD model capabilities without retraining from scratch, significantly reducing computational costs. By evaluating performance retention and knowledge transfer, this study offers valuable insights into designing adaptive SELD models capable of efficiently learning new sound classes with minimal performance degradation.

The main contributions of this work are as follows:

- A comprehensive study of CIL for SELD, analyzing how incremental learning impacts detection and localization performance across various class-order strategies.
- Analysis of the influence of class difficulty in incremental stages, demonstrating improved performance when introducing acoustically challenging classes in later training stages.
- Performance comparison across balanced and imbalanced datasets, offering insights into how the distribution of classes in the dataset affects CIL performance for SELD.

II. Class-Incremental Learning for SELD

A signal comprising a mixture of multiple sound events originating from various spatial locations is modelled as a multi-output regression task to perform sound event detection and localization. In this setup, the Activity-Coupled Cartesian DOA (ACCDOA) format is used to jointly represent sound event detection (SED) and direction of arrival (DOA) [5]. ACCDOA encodes sound event activation and spatial localization using three coordinates (x,y,z) representing the event's

This work was partially supported by Jane and Aatos Erkko Foundation under grant number 230048, "Continual learning of sounds with deep neural networks".

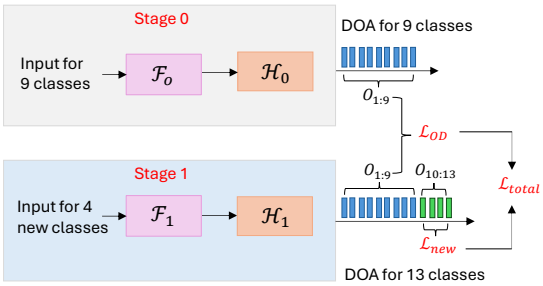

Fig. 1: Class-incremental learning for SELD task

DOA. An event is considered active if the magnitude of this coordinate vector exceeds a predefined threshold, indicating its predicted DOA. We use a convolutional recurrent neural network (CRNN) architecture based on the DCASE 2021 baseline [22]. The model is trained on log-mel spectrograms and acoustic intensity vectors extracted from multichannel audio. The CRNN predicts active sound events and their spatial positions by analyzing consecutive feature frames. The model outputs a single ACCDOA sequence, encoding SED and DOA information, allowing it to localize and detect sound events in complex acoustic environments. Further details on the SELD architecture can be found in [22].

Figure 1 illustrates the class-incremental learning framework for SELD. The model is trained in two stages to systematically evaluate the effects of class order and difficulty on performance:

Stage 0 (Initial training): The model consists of a feature extractor, $\mathcal{F}_0$, and a regression network, $\mathcal{H}_0$, initially trained on a set of nine selected sound classes. The regression network contains 9×3 output neurons corresponding to ACCDOA predictions for the nine sound event classes.

Stage 1 (Incremental learning): The regression network is expanded to accommodate four additional classes, resulting in an updated regression network, $\mathcal{H}_1$, with $(9+4) \times 3$ output neurons. During Stage 1 training, both the feature extractor and regression model parameters are updated using training data of the new classes, while aiming to preserve performance on previously learned classes.

A key challenge during incremental learning is catastrophic forgetting, where the model's performance on previously learned classes deteriorates when new classes are introduced. To address this, we employ a mean square error (MSE)-based output distillation loss ($\mathcal{L}_{\mathcal{OD}}$). This loss calculates the discrepancy between the original model outputs (Stage 0) and the updated model outputs (Stage 1) for the nine previously learned classes, ensuring knowledge retention. The total training loss combines the distillation loss and the MSE loss for new classes:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{\text{MSE}} + \lambda_2 \mathcal{L}_{\mathcal{OD}} \tag{1}$$

Here, $\mathcal{L}_{\text{MSE}}$ represents the loss between the predicted and actual outputs for newly introduced classes, while λ_1 and λ_2 controls the balance between preserving old knowledge and learning new classes. This framework enables systematic investigation into how varying class combinations and difficulty levels introduced in the two different learning stages affect SELD performance, providing insights essential for effective incremental learning in acoustic environments.

III. Experimental Setup

A. Dataset

For this study, we generated a balanced synthetic dataset in which all classes appear nearly equally likely temporally. The generated dataset is balanced in terms of temporal coverage, with each class occupying a roughly equal amount of time (in minutes or timeframes) but not in terms of the number of event instances per class, as each class has a different time duration. The motivation for generating this balanced dataset was to ensure equal representation of each target sound event class throughout the dataset, reducing the bias of deep learning methods in learning the large classes in detriment of the small-size ones. The dataset was created using the data generation pipeline provided by the DCASE Challenge 2022, following the same method as the TAU-NIGENS Spatial Sound Scenes 2021 dataset [22][1] [2].

The generated dataset comprises 1200 one-minute synthetic spatial audio mixtures featuring sound events from 13 target classes: *Female speech, Male speech, Clapping, Telephone, Laughter, Domestic sounds, Footsteps, Door open or close, Music, Musical instrument, Water tap, Bell, and Knock*. Each spatial audio mixture is sampled at 24 kHz and presented in a 4-channel first-order Ambisonics (FOA) format. Spatial sound events are synthesized within an angular range of $[-180°, 180°]$ for azimuth and $[-45°, 45°]$ for elevation, ensuring realistic spatial distributions. Mixtures allow for a maximum polyphony of 2 simultaneous sound events, including the possibility of same-class event overlap. The training set includes 900 recordings spatialized across 6 rooms, based on samples from the FSD50K development set, while the testing set comprises 300 recordings from 3 separate rooms sourced from the FSD50K evaluation set. The dataset generator uses real multipoint room impulse responses captured in various rooms to synthesize static and moving events at various configurations.

B. Baseline and evaluation metrics

We adopt the SELD baseline model from the DCASE 2022 Challenge, configured specifically for single-output ACCDOA predictions [23]. The baseline model is simultaneously trained and evaluated on the entire dataset, covering all 13 target sound classes. The model optimizes a Mean Square Error (MSE) loss during training between predicted and ground-truth ACCDOA outputs. This baseline represents the conventional deep learning scenario without continual learning, as it simultaneously learns all classes and thus provides reference performance for classwise analysis. For incremental learning experiments, the

[1]https://github.com/sharathadavanne/ seld-dcase2022
[2]https://github.com/danielkrause/DCASE2022-data-generator

TABLE I: Averaged performance metric across different class-ordering experiments on CIL-SELD

Experiment	CIL						Old model	
	F1 (1:13)	LE (1:13)	F1 C_{old}(1:9)	LE C_{old}(1:9)	F1 C_{new}(10:13)	LE C_{new}(10:13)	F1	LE
Baseline	41.6	19.3	-	-	-	-	-	-
Exp1: 4 easy classes	39.4	19.0	35.5	19.9	**47.7**	16.9	37.9	18.9
Exp2: 1 difficult class	39.8	18.6	38.9	19.3	41.5	**16.6**	42.1	19.0
Exp3: 2 difficult classes	41.7	18.3	42.6	18.6	39.2	17.7	44.0	17.7
Exp4: 3 difficult classes	41.8	17.9	45.2	17.8	34.5	18.0	47.6	16.8
Exp5: 4 difficult classes	**43.0**	**17.5**	**46.8**	**16.6**	34.0	19.3	**48.5**	**16.0**

initial training (Stage 0) incorporates only 9 target classes, treating the remaining 4 as interfering sounds. The model is trained using the Adam optimizer (learning rate of $1e^{-3}$), with a batch size of 128 for 100 epochs.

We evaluate SELD performance using the spatially-thresholded F1-score to check how accurately events are detected and penalize predictions whose directions deviate from the actual source, based on a set angular threshold. Following established practice, we set this threshold to $T = 20°$ [6], [23]. We also calculate the Localization error (LE) individually for each sound class and report the averaged value. LE quantifies the mean angular difference by pairing predicted DOAs to their nearest ground-truth references, providing complementary insights into the localization accuracy that is not captured by threshold-based metrics. Both metrics are computed on one-second non-overlapping frames. Detailed descriptions of SELD evaluation metrics can be found in [6], [24].

IV. Results and Discussions

We trained the continual learning model in two stages to analyze the effect of class order and acoustic difficulty on CIL performance for SELD. In each experiment, the first stage involves training on 9 classes, followed by incremental training on 4 additional classes. The acoustic difficulty of each class was determined based on its baseline performance, specifically characterized by the lowest F1-score and highest LE from the SELD baseline model (*Clapping, Telephone, Door open/close, and Bell*). This systematic variation allows a comprehensive analysis of incremental learning behavior under different class-order scenarios. The experiments were designed by varying the ratio of difficult to easy classes introduced in Stage 2, as follows:

- Experiment 1: Stage 2 with 4 easy classes.
- Experiment 2: Stage 2 with 1 difficult and 3 easy classes.
- Experiment 3: Stage 2 with 2 difficult and 2 easy classes
- Experiment 4: Stage 2 with 3 difficult and 1 easy class
- Experiment 5: Stage 2 with 4 difficult classes

A. Overall performance analysis

Table I summarizes the overall performance metrics across various class-ordering scenarios evaluated in the CIL experiments.The overall F1-score, computed by averaging scores across classes, consistently improves as more difficult classes are introduced in Stage 2. This shows that adding acoustically challenging classes later during training boosts detection performance, resulting in a 3.37% improvement. Similarly, the average LE decreases by 9.33% as more difficult classes are added in Stage 2, indicating better overall localization accuracy. Models trained incrementally with difficult classes achieve a higher overall F1-score and lower LE than the baseline, which is trained simultaneously on all classes.

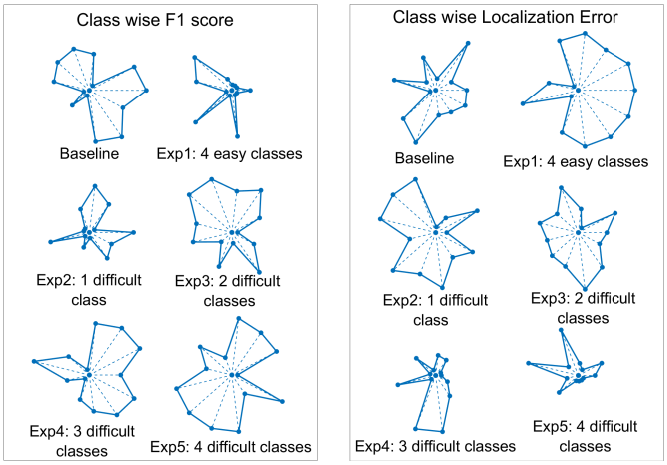

Fig. 2: Glyph plots illustrating class-wise variations in F1-score (left) and LE (right) across baseline and incremental learning experiments. For F1-score, larger glyph shapes indicate improved class-wise detection performance, while for LE, smaller and compressed glyph shapes reflect better localization accuracy.

However, the forgetting phenomenon remains evident, causing performance degradation in both detection and localization for the previously learned classes relative to their initial performance. This can be seen from the same table, where we provide separately calculated metrics for the original 9 classes (C_{old}) and the newly introduced 4 classes (C_{new}). The comparison shows that forgetting occurs more significantly when fewer difficult classes are added in Stage 2. In these cases, the F1-score and localization accuracy (lower LE) notably decline for previously learned classes. This analysis emphasizes that careful ordering and consideration of class difficulty are crucial for effective incremental learning in SELD tasks. It highlights the potential performance benefits and the challenge of managing catastrophic forgetting.

Figure 2 shows glyph plots that illustrate the class-wise variations in F1-score and LE across different incremental learning scenarios and the baseline. Each radial axis represents

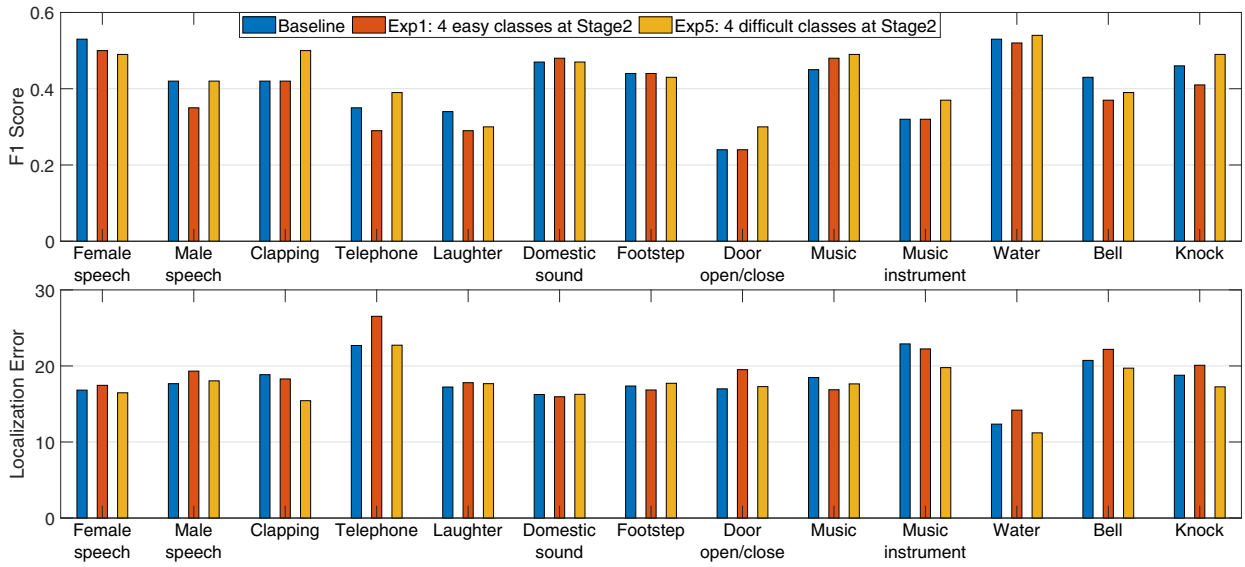

Fig. 3: Classwise comparison of F1-score and LE for the Baseline, Experiment 1, and Experiment 4

one sound class, with its length proportional to the value of the performance metric, forming an area that reflects the overall magnitude across classes. A larger area indicates better performance for the F1-score, while a smaller area signifies lower LE and thus better performance in that metric. The results show that introducing more difficult classes incrementally (such as in Experiment 5 with four difficult classes) leads to consistently higher F1-scores, reflected by larger and more symmetric glyph shapes, and better localization accuracy, indicated by compressed glyph shapes. In contrast, when mostly easy classes are introduced in Stage 2 (Experiment 1), glyph shapes become smaller, with lower overall F1-scores and higher localization errors. This highlights the importance of strategically ordering classes to optimize incremental SELD performance.

B. Classwise performance analysis

Figure 3 illustrates the class-wise performance comparison for F1-score and LE across three experimental conditions: baseline, Experiment 1 and Experiment 5. It can be seen that in most cases (9 out of 13 classes), incrementally introducing difficult classes in Stage 2 yields improved performance, with higher F1-scores and lower LE compared to both the baseline and the incremental scenario with easy classes. This further demonstrates the effectiveness of introducing acoustically challenging classes later in the incremental training process, leading to superior class-wise detection and localization accuracy.

C. Comparison with imbalanced datasets

In this section, we explore the impact of the total number of occurrences of each class within the entire training dataset on the proposed CIL strategy. As explained in Section III-A, the generated dataset is intentionally balanced, ensuring equal occurrence probability for all sound event classes. In this section, we compare the performance of the balanced dataset against two publicly available datasets: the highly imbalanced DCASE 2021 dataset (with 12 classes) and the moderately balanced DCASE 2022 synthetic dataset. Figure 4 (top row) illustrates the temporal distribution of class-wise occurrence for the DCASE 2021, DCASE 2022, and the balanced dataset generated for this study, respectively. To select difficult classes for training in Stage 2, we identified classes that exhibited the lowest F1-score and highest LE in the SELD baseline. Note that the DCASE 2022 dataset has identical classes in the same order as our generated balanced dataset, enabling direct comparisons between the two datasets. However, in our balanced dataset, classes with the lowest F1-score and highest LE in the SELD baseline represent acoustic difficulty. In contrast, the low-performing classes in the DCASE 2022 dataset may reflect acoustic difficulty or insufficient data, as low occurrence inherently limits the model's ability to learn effectively.

Figure 4 (bottom row) compares the class-wise F1-scores obtained from the SELD baseline and the incremental learning scenario with four difficult classes introduced at Stage 2, across the three datasets. Incremental training on difficult classes substantially improved the F1-scores for those classes in all three datasets, demonstrating that strategically introducing difficult classes (either due to fewer occurrences in the dataset or being acoustically complex) effectively enhances detection accuracy. These results show the effectiveness of the incremental learning strategy across datasets with varying levels of class imbalance.

V. Conclusions

In this study, we explored CIL for the SELD task, systematically analyzing how incremental learning affects detection and localization across different class-ordering scenarios. Although CIL is traditionally used when future classes are unknown, our study illustrates that incremental learning can be deliberately leveraged to handle challenging scenarios. The

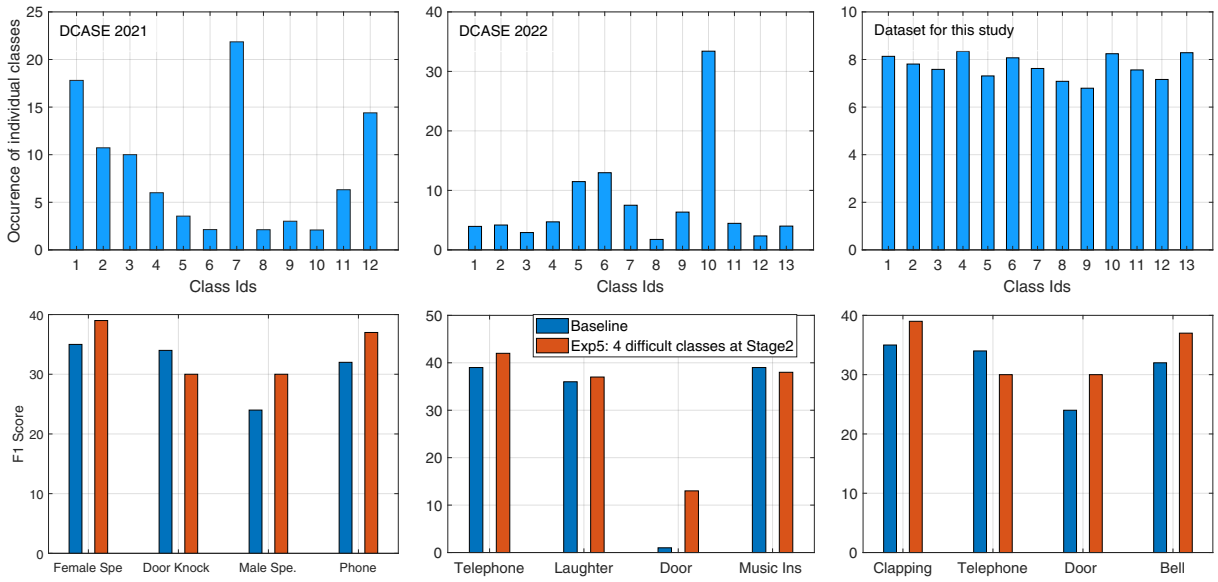

Fig. 4: Comparison of class-wise occurrence distributions (top row) and corresponding F1-scores (bottom row) between SELD baseline and incremental learning (4 difficult classes in Stage 2) across highly imbalanced (DCASE 2021), moderately balanced (DCASE 2022), and balanced (this study) datasets.

results show that strategically introducing difficult classes later during incremental stages, whether due to acoustic complexity, limited data availability, or other factors, improves overall SELD performance. This approach leads to improved overall results and more balanced class-wise performance compared to training in all classes simultaneously.

REFERENCES

[1] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Euro. Sig. Proces. Conf. (EUSIPCO)*, 2018, pp. 1462–1466.

[2] L. Wan *et al.*, "The application of DOA estimation approach in patient tracking systems with high patient density," *IEEE Trans. Ind. Electron.*, vol. 12, no. 6, pp. 2353–2364, 2016.

[3] M. Farmani *et al.*, "Maximum likelihood approach to "informed" sound source localization for hearing aid applications," in *IEEE Inter. Conf. Acous., Spe., Sig. Proces.*, 2015, pp. 16–20.

[4] P. A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *J. Acoust. Soc. Am.*, vol. 152, no. 1, pp. 107–151, 2022.

[5] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *IEEE Inter. Conf. Acous., Spe., Sig. Proces.*, 2021, pp. 915–919.

[6] A. Mesaros *et al.*, "Joint measurement of localization and detection of sound events," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 333–337.

[7] Y. Cao *et al.*, "An improved event-independent network for polyphonic sound event localization and detection," in *IEEE Inter. Conf. Acous., Spe., Sig. Proces.*, 2021, pp. 885–889.

[8] S. Jung, J. Park, and S. Lee, "Polyphonic sound event detection using convolutional bidirectional LSTM and synthetic data-based transfer learning," in *IEEE Inter. Conf. Acous., Spe., Sig. Proces.*, 2019, pp. 885–889.

[9] L. Xu, L. Wang, S. Bi, H. Liu, and J. Wang, "Semi-supervised sound event detection with pre-trained model," in *IEEE Inter. Conf. Acous., Spe., Sig. Proces.*, 2023, pp. 1–5.

[10] Y. Xiao and R. K. Das, "Dual knowledge distillation for efficient sound event detection," *preprint arXiv:2402.02781*, 2024.

[11] S. Hou, X. Pan, C. Loy, Z. Wang, and D. Lin, "Learning a unified classifier incrementally via rebalancing," in *IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 831–839.

[12] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural networks*, vol. 113, pp. 54–71, 2019.

[13] S. A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental classifier and representation learning," in *IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.

[14] G. M. V. de Ven and A. S. Tolias, "Three scenarios for continual learning," *preprint arXiv:1904.07734*, 2019.

[15] P. Garg, R. Saluja, V. Balasubramanian, C. Arora, A. Subramanian, and C. Jawahar, "Multi-domain incremental learning for semantic segmentation," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 761–771.

[16] R. Bhatt, P. Kumari, D. Mahapatra, A. E. Saddik, and M.Saini, "Characterizing continual learning scenarios and strategies for audio analysis," *preprint arXiv:2407.00465*, 2024.

[17] M. Mulimani and A. Mesaros, "Class-incremental learning for multi-label audio classification," in *IEEE Inter. Conf. Acous., Spe., Sig. Proces.*, 2024, pp. 916–920.

[18] Y. Huang, N. Hou, and N. F. Chen, "Progressive continual learning for spoken keyword spotting," in *IEEE Inter. Conf. on Acous., Spe. and Sig. Process. (ICASSP)*. IEEE, 2022, pp. 7552–7556.

[19] Y. Xiao *et al.*, "Continual learning for on-device environmental sound classification," *preprint arXiv:2207.07429*, 2022.

[20] Y. Xiao and R. K. Das, "Configurable DOA estimation using incremental learning," *preprint arXiv:2407.03661*, 2024.

[21] R. Pandey, M. Mulimani, A. Politis, and A. Mesaros, "Class-incremental learning for sound event localization and detection," *preprint arXiv:2411.12830*, 2024.

[22] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," *preprint arXiv:2106.06999*, 2021.

[23] A. Politis *et al.*, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *preprint arXiv:2206.01948*, 2022.

[24] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in DCASE 2019," *IEEE/ACM Trans. on Aud., Spe., and Lang. Process.*, vol. 29, pp. 684–698, 2020.