The APSIPA ASC 2025 Grand Challenge on City and Time-Aware Semi-supervised Acoustic Scene Classification: Summary and Results

Jisheng Bai[1,2], Mou Wang[3], Haohe Liu[4], Bin Xiang[2], Ying Liu[1], Jianfeng Chen[5]
Dongyuna Shi[5], Mark D. Plumbley[4], Susanto Rahardja[6], Woon-Seng Gan[7]
[1]*Xi'an University of Posts & Telecommunications, China*
[2]*Xi'an Lianfeng Acoustic Technologies Co., Ltd., China*
[3]*Institute of Acoustics, Chinese Academy of Sciences, China*
[4]*University of Surrey, UK*
[5]*Northwestern Polytechnical University, China*
[6]*Singapore Institute of Technology, Singapore*
[7]*Nanyang Technological University, Singapore*

***Abstract*—The APSIPA ASC 2025 grand challenge on City and Time-Aware Semi-supervised Acoustic Scene Classification focuses on leveraging geographic and temporal context to improve acoustic scene classification. Building on the ICME 2024 edition, this year's task introduces metadata including city-level location and timestamps for each audio sample. The problem is formulated as a multimodal classification task, where systems integrate both audio signals and contextual information. The challenge maintains a semi-supervised learning framework, which reflects practical scenarios characterized by abundant unlabeled data and limited labeled resources. A total of nine teams from different regions submitted valid results. The best-performing system achieved a macro accuracy of 0.644, higher than the baseline score of 0.582. The outcomes demonstrate that incorporating geographic and temporal metadata can improve the robustness of acoustic scene classification across diverse environments.**

I. Introduction

Acoustic scene classification (ASC) aims to identify the type of acoustic environment from predefined classes, such as bus, streets, and restaurants, combining signal processing and machine learning methods. With the rapid advancement of mobile computing, intelligent wearables, robotics, and smart home technologies, ASC has gained growing attention as a key component for supporting machine perception and adaptation in complex real-world environments [1]–[4].

Over the past decade, ASC has made significant progress, supported by the establishment of standardized evaluation frameworks and the rapid development of deep learning technologies. The Detection and Classification of Acoustic Scenes and Events (DCASE) challenges have provided an important benchmark for the global research community to design and evaluate ASC algorithms [5], [6]. At the same time, advances in deep neural networks, including the use of pre-trained models and sophisticated architectures, have substantially transformed ASC methodologies and contributed to notable improvements in classification performance [5], [7]–[11]. Despite these advances, the deployment of ASC systems in real-world settings continues to face persistent challenges. Variations in recording conditions, background noise, scarcity of labeled data, and the influence of geographic, cultural, and temporal factors on acoustic characteristics all limit the robustness and generalizability of current approaches.

Among the various challenges facing ASC systems, domain shift and data scarcity have emerged as particularly critical issues. Domain shift refers to discrepancies between training and testing data distributions, which commonly occur when models are deployed across different acoustic environments and often lead to performance degradation [7], [12]–[14]. Although collecting labeled data from new domains can help mitigate such effects, this solution is constrained by the high cost and limited availability of manual annotation, especially across diverse global environments [15]. To address data scarcity, semi-supervised learning offers a promising direction by exploiting abundant unlabeled recordings from target domains while reducing dependence on costly labeled resources [16]. To investigate this approach, the ICME 2024 grand challenge (GC) on Semi-supervised Acoustic Scene Classification under Domain Shift was organized [17]. The challenge attracted 13 teams, with 7 surpassing the baseline system, and demonstrated strong potential for semi-supervised learning and domain generalization methods in cross-domain acoustic scene classification.

Building on the success of the ICME 2024 challenge, it is recognized that many existing ASC approaches treat acoustic scenes as static environments and often overlook variations across different cities and temporal contexts. The acoustic characteristics of the same scene category can differ substantially between locations and times. For instance, a public square may exhibit distinct patterns during weekday mornings compared with weekend evenings, or between cities shaped by different cultural practices. To address this limitation, we introduced the APSIPA 2025 GC on City and Time-Aware Semi-supervised Acoustic Scene Classification. The challenge

979-8-3315-7206-8/25/$31.00

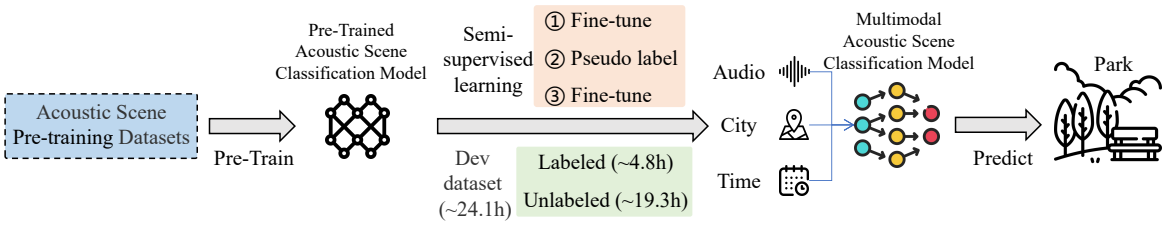

Fig. 1: Overview of the APSIPA 2025 grand challenge on City and Time-Aware Semi-supervised Acoustic Scene Classification.

provides participants with city-level location data and precise timestamps for each audio sample, and is designed to support the development of classification approaches that exploit such contextual information to improve accuracy and generalization.

The APSIPA challenge is based on approximately 24 hours of audio recordings collected from 22 cities in China between April and September 2023. The dataset covers 10 acoustic scene categories and includes both city-level identifiers and precise timestamp metadata. Following a semi-supervised learning paradigm, it contains about 4.8 hours of labeled data and 19.3 hours of unlabeled data. The baseline system[1], which employs a squeeze-and-excitation and a Transformer (SE-Trans)-based architecture, achieves a macro-accuracy of 0.582 [18]. A total of nine teams participated in the challenge, and the best-performing system reached a macro-accuracy of 0.644, highlighting the potential of incorporating spatiotemporal context to improve ASC performance[2].

The organization of this paper is arranged as follows. Section 2 introduces the challenge setup, Section 3 introduces the challenge results, and Section 4 concludes this paper.

II. Challenge Setup

The APSIPA 2025 GC adopts a three-stage semi-supervised learning framework that incorporates both city and time information into ASC. Participants are provided with spatiotemporal metadata, including city-level location and precise timestamps, together with audio recordings. This setup supports the development of models designed to capture contextual variations across diverse urban environments and temporal periods. An overview of the framework is illustrated in Fig. 5.

A. Challenge Dataset

The APSIPA 2025 GC ASC dataset extends the ICME 2024 GC ASC dataset, which itself is a curated subset of the Chinese Acoustic Scene (CAS) 2023 dataset. While the ICME dataset is directly derived from CAS 2023, the APSIPA dataset further incorporates city-level location and timestamp metadata. The CAS 2023 dataset is a large-scale corpus of environmental acoustic scenes, comprising over 130 hours of recordings from 10 common scene categories. Each 10-second audio clip was recorded at a sampling rate of 48 kHz in 22 cities across China between April and September 2023, using a XS-SN-2BE1 binaural sound monitoring device[3]. All recordings are accompanied by city and timestamp annotations to support research on environmental sound analysis and context-aware acoustic scene classification.

TABLE I: The number of labeled recordings and cities for each scene in the development dataset

Scene	Labeled recordings	Cities
Bus	188	2 (Hefei, Shangrao)
Airport	220	2 (Hefei, Xi'an)
Metro	209	2 (Shanghai, Xi'an)
Restaurant	173	2 (Liupanshui, Xi'an)
Shopping mall	147	3 (Luoyang, Shanghai, Xi'an)
Public square	174	1 (Xi'an)
Urban park	148	3 (Luoyang, Chongqing, Xi'an)
Traffic street	143	1 (Xi'an)
Construction site	173	2 (Luoyang, Jinan)
Bar	165	2 (Luoyang, Chongqing)
Total	1740	8 (Hefei, Luoyang, Chongqing, Shanghai, Xi'an, Liupanshui, Jinan, Shangrao)

The APSIPA 2025 GC ASC dataset retains the same audio files as the ICME 2024 GC ASC dataset, where the development set[4] contains approximately 24 hours of audio, including about 4.8 hours of labeled data and 19.3 hours of unlabeled data. The evaluation set[5] comprises about 3.1 hours of audio and is designed to assess system performance under domain shift. In contrast to the ICME 2024 edition, the APSIPA 2025 dataset additionally provides city-level location and timestamp metadata for each audio clip, enabling research on city- and time-aware ASC.

The development dataset contains recordings from eight cities: Hefei, Luoyang, Chongqing, Shanghai, Xi'an, Liupanshui, Jinan, and Shangrao. The evaluation dataset includes recordings from twelve cities, six of which are not present in the development set: Changchun, Guangzhou, Nanchang, Shenyang, Taiyuan, and Tianjin. The number of labeled recordings and the cities represented for each scene in the development dataset are summarized in Table I.

We show the distributions of the city, hour and weekday for the development and evaluation dataset in Fig. 2, Fig. 3, and Fig. 4, respectively. The APSIPA 2025 GC dataset covers 14 Chinese cities, with the development set showing over 1,000 recordings in Xi'an and Luoyang, while the evaluation set

[1]Baseline: https://github.com/JishengBai/APSIPA2025GC-ASC
[2]Challenge website: https://ascchallenge.xshengyun.com/
[3]https://www.lfxstek.com
[4]Development Dataset: https://zenodo.org/records/10616533
[5]Evaluation Dataset: https://zenodo.org/records/10820626

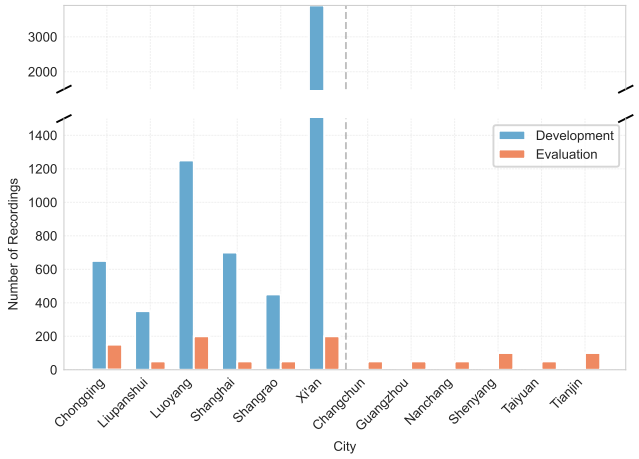

Fig. 2: City distribution comparison between development and evaluation sets.

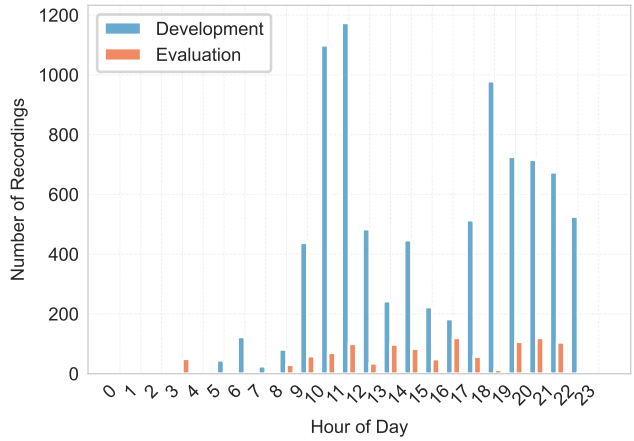

Fig. 3: Hour distribution comparison between development and evaluation sets.

provides a balanced distribution across the cities for domain shift conditions. Regarding hour distribution, both development and evaluation sets contain more recordings during periods of active human activities. For weekday distribution, both sets show higher recording frequencies on weekends compared to weekdays.

B. Baseline

The APSIPA 2025 GC baseline builds upon the semi-supervised learning framework established in the ICME 2024 GC baseline system[6] by processing and fusing audio, city, and temporal contextual information. The system architecture employs a pre-trained ASC model as its foundation, leveraging pre-training on external acoustic scene datasets to establish robust feature representations. During fine-tuning, labeled data from the development set enables the model to learn the mapping between acoustic patterns and corresponding city and temporal metadata. The framework then processes unlabeled

[6]https://github.com/JishengBai/ICME2024ASC

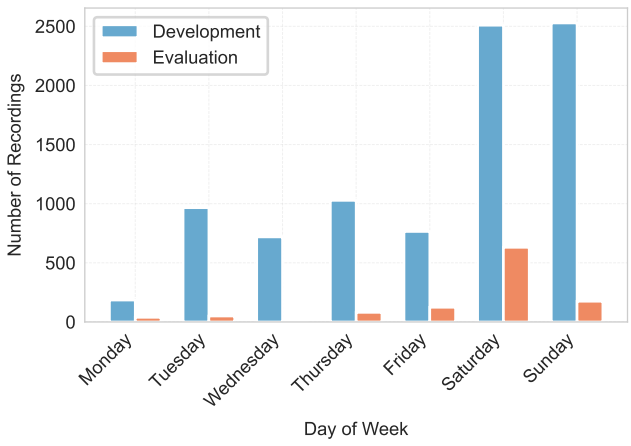

Fig. 4: Weekday distribution comparison between development and evaluation sets.

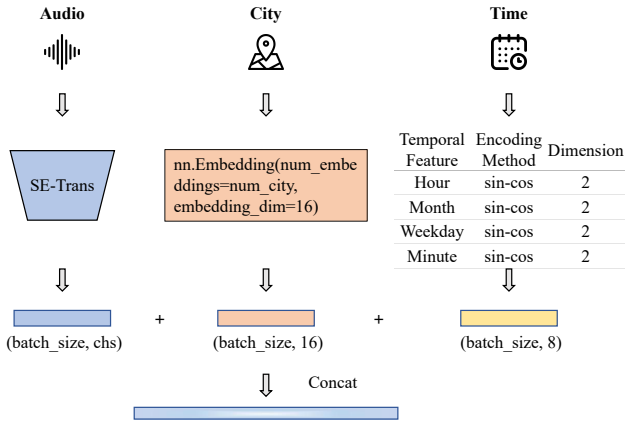

Fig. 5: The processing pipeline of the challenge baseline.

data through pseudo-label generation and constructs the final training dataset. The complete pipeline architecture of the baseline system is illustrated in Fig. 5.

The system employs the SE-Trans architecture, which consists of a hierarchical feature extractor combined with a Transformer-based temporal aggregator [18]. The feature extractor utilizes two convolutional SE blocks with channel dimensions of 64 and 128, where each block incorporates channel-wise attention mechanisms. Temporal features are processed using a single-layer Transformer encoder. The system combines temporal, city and acoustic features using a fusion approach. Temporal information is encoded into an 8-dimensional vector through sin-cos encoding. City information is processed with an embedding layer that maps each city to a 16-dimensional vector. The acoustic, city and temporal features are concatenated into a unified representation that passes through a fully-connected layer for ASC.

Audio recordings are processed after resampling to 44100 Hz, with 64-band log mel spectrograms extracted through STFT using 40 ms Hanning windows and 20ms hop length. Model optimization is based on the Adam optimizer with a learning rate of 10^{-3} and a batch size of 32 during fine-tuning.

TABLE II: Challenge results of all participating teams and baseline system

Rank	Team Name	Score	Bus	Airport	Metro	Restaurant	Shopping mall	Public square	Urban park	Traffic street	Construction site	Bar
1	DeepTone	0.644	0.670	0.753	0.910	0.650	0.500	0.170	0.520	0.780	0.580	0.910
2	Kawamura_TMU	0.628	0.430	0.800	0.860	0.600	0.440	0.680	0.613	0.660	0.440	0.760
3	ALPS	0.613	0.440	0.693	0.920	0.750	0.580	0.040	0.700	0.650	0.510	0.850
4	Masayuki-sera-TMU	0.586	0.420	0.667	0.830	0.690	0.400	0.240	0.627	0.540	0.760	0.690
*	*Baseline*	0.582	0.580	0.513	0.770	0.580	0.480	0.260	0.640	0.610	0.500	0.890
5	HAI-LAB	0.539	0.060	0.860	0.910	0.520	0.490	0.370	0.440	0.600	0.500	0.640
6	ditlab	0.522	0.530	0.640	0.830	0.520	0.430	0.190	0.700	0.730	0.560	0.090
7	gisp	0.500	0.140	0.460	0.670	0.640	0.520	0.480	0.413	0.650	0.500	0.530
8	Li_NTU	0.493	0.450	0.533	0.600	0.690	0.530	0.030	0.533	0.640	0.490	0.430
9	Audio_IIT Mandi	0.341	0.020	0.113	0.620	0.170	0.050	0.400	0.780	0.500	0.460	0.300

TABLE III: Comparison of methods proposed by the top four performing teams

Team	Data augmentation	Features	External data sources	Classifier	City and Temporal Data Usage	Strategy	Model Complexity
DeepTone	-	Log-mel spectrogram	TAU UAS 2020 Mobile CochlScene	DenseEncoder, Dual-path Mamba, SE-Trans	Text Embedding and Conditional Layer Normalization	Two-step pseudo-labeling	Param 1.6M MACs 20.0G
Kawamura _TMU	-	Log-mel spectrogram with low-pass filtering	-	SE-Trans with average pooling	Embedding, sin-cos encoding	City-disjoint cross-validation	Param 0.4M MACs 1.1G
ALPS	SpecAugment, Mixup	Mel spectrogram	TAU UAS 2020 Mobile CochlScene	Residual CNN with attention	Embedding, sin-cos encoding + linear projection	The same as baseline	Param 21.7M MACs 2.3G
Masayuki-sera-TMU	Waveform pitch shift	Log-mel spectrogram	-	SE-Trans with average pooling	Embedding, sin-cos encoding	City-disjoint cross-validation	Param 0.4M MACs 3.0G

The evaluation adopts macro-average accuracy as the primary metric, consistent with previous ASC challenges [5], [7], [17].

III. Challenge Results

The challenge received submissions from nine participating teams. All entries were evaluated using macro-accuracy across ten acoustic scene classes, as shown in Table II.

Four teams achieved performance superior to the baseline system with a score above 0.582, while five teams scored below this threshold. DeepTone achieves the leading performance with an overall accuracy of 0.644, exceeding the baseline by 0.062. Among acoustic scene categories, Metro environments achieved the highest average recognition rates across all submissions, with most teams reaching accuracies above 0.8. Bar environments also demonstrated strong classification performance, with several teams achieving scores exceeding 0.8. Conversely, Public Square scenes proved most challenging, with many teams recording accuracies below 0.4. The approaches adopted by the four highest-ranking teams are introduced as follows.

DeepTone utilized log-mel spectrograms and external ASC datasets for pre-training. The proposed ASCMamba model combines a DenseEncoder with dual-path Mamba blocks and employs conditional layer normalization to fuse city and time data. A two-step pseudo-labeling strategy is adopted to enhance the semi-supervised learning.

Kawamura_TMU processed low-pass filtered log-mel spectrograms using an SE-Trans model with average pooling. City-disjoint cross-validation ensured robust evaluation across unseen urban environments.

ALPS applied SpecAugment and Mixup to mel spectrograms and incorporated external datasets. Their residual CNN with attention fused city and time data through embeddings and linear projections.

Masayuki-sera-TMU enhanced log-mel spectrograms with waveform-level pitch shifting. They also modified SE-Trans with average pooling and used city-disjoint cross-validation to improve model generalization.

Table III summarizes the proposed methods employed by these teams. DeepTone introduced the dual-path Mamba architecture alongside DenseEncoder and SE-Trans components, achieving the highest performance. For city and temporal information processing, DeepTone proposed text embedding and a novel conditional layer normalization method. Two teams, Kawamura_TMU and Masayuki-sera-TMU, adopted city-disjoint cross-validation strategies, which potentially enhance model generalization by ensuring geographic diversity in validation sets and reducing overfitting to specific urban environments. In terms of model complexity, Kawamura_TMU and Masayuki-sera-TMU remained close to the baseline with only minor adjustments, DeepTone added relatively few parameters but incurred a larger increase in MACs (multiply–accumulate operations), while ALPS introduced a substantially larger parameter size with only modest growth in computational cost.

IV. Conclusion

The APSIPA 2025 GC aims to enhance ASC performance by incorporating spatiotemporal context through a semi-supervised learning framework. The challenge attracted nine participating teams and established an evaluation benchmark for city- and time-aware ASC. Four teams exceeded the baseline performance of 0.582, with DeepTone achieving the top performance of 0.644. The teams adopted novel architectures such as Mamba, new normalization methods for processing city and time information, and new strategies within semi-supervised frameworks, which hold significance for advancing city- and time-aware ASC research and improving classification performance.

Acknowledgement

We acknowledge the support of Xi'an Lianfeng Acoustic Technologies Co., Ltd. for providing the recording devices in data collection and computational resources. This work was partly supported by a PhD scholarship from the Centre for Vision, Speech and Signal Processing (CVSSP), Faculty of Engineering and Physical Science (FEPS), University of Surrey, the British Broadcasting Corporation Research and Development (BBC R&D), and the Engineering and Physical Sciences Research Council [grant number EP/T019751/1]. For the purpose of open access, the authors have applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising.

References

[1] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[2] B. Ding, T. Zhang, C. Wang, *et al.*, "Acoustic scene classification: A comprehensive survey," *Expert Systems with Applications*, vol. 238, p. 121 902, 2024, ISSN: 0957-4174.

[3] I.-Y. Jeong and J. Park, "CochlScene: Acquisition of acoustic scene data using crowdsourcing," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, IEEE, 2022, pp. 17–21.

[4] J. Bai, H. Yin, M. Wang, *et al.*, "AudioLog: LLMs-powered long audio logging with hybrid token-semantic contrastive learning," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2024.

[5] A. Mesaros, T. Heittola, E. Benetos, *et al.*, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, 2017.

[6] A. Mesaros, R. Serizel, T. Heittola, T. Virtanen, and M. D. Plumbley, "A decade of DCASE: Achievements, practices, evaluations and future challenges," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025.

[7] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in DCASE 2020 challenge: Generalization across devices and low complexity solutions," *arXiv preprint arXiv:2005.14623*, 2020.

[8] F. Schmid, T. Morocutti, S. Masoudian, K. Koutini, and G. Widmer, "CP-JKU Submission to DCASE23: Efficient Acoustic Scene Classification with CP-Mobile," DCASE2023 Challenge, Tech. Rep., May 2023.

[9] M. Wang and R. Wang, "Ciaic-ASC System for DCASE 2019 Challenge Task1," DCASE2019 Challenge, Tech. Rep., Jun. 2019.

[10] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[11] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[12] A. Mesaros, T. Heittola, and T. Virtanen, "Acoustic Scene Classification in DCASE 2019 Challenge: Closed and Open Set Classification and Data Mismatch Setups," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop*, 2019, pp. 164–168.

[13] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," *arXiv preprint arXiv:1807.09840*, 2018.

[14] Y. Tan, H. Ai, S. Li, and M. D. Plumbley, "Acoustic scene classification across cities and devices via feature disentanglement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1286–1297, 2024.

[15] S. Singh, H. L. Bear, and E. Benetos, "Prototypical networks for domain adaptation in acoustic scene classification," in *International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2021, pp. 346–350.

[16] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," *arXiv preprint arXiv:1807.10501*, 2018.

[17] J. Bai, M. Wang, H. Liu, *et al.*, "Description on IEEE ICME 2024 grand challenge: Semi-supervised acoustic scene classification under domain shift," *arXiv preprint arXiv:2402.02694*, 2024.

[18] J. Bai, J. Chen, M. Wang, M. S. Ayub, and Q. Yan, "A Squeeze-and-Excitation and Transformer-Based Cross-Task Model for Environmental Sound Recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 3, pp. 1501–1513, 2023.