Improving Acoustic Scene Classification via City Embeddings and Pseudo-Labeling with SE-Trans

Yuuki Tachioka*
* Denso IT Laboratory, Tokyo
E-mail: tachioka.yuki@core.d-itlab.co.jp

***Abstract*—Acoustic Scene Classification (ASC) identifies the environment of an audio recording, with applications in smart cities and urban planning. Conventional ASC methods often ignore geographic and temporal variability, limiting generalization. The APSIPA ASC 2025 Grand Challenge provides city and timestamp metadata in a semi-supervised setting, enabling context-aware modeling. We propose a geography-aware semisupervised framework that combines the two allowed external datasets through label harmonization and augmentation. For data without city labels, pseudo-city labels are generated by grouping users; synonymous scene labels are normalized, and auxiliary classes are added. The city / pseudo-city embeddings are integrated into the SE-Trans backbone during pre-training. We compare two schedules: `ext-ft1` (pseudo labels after pre-training) and `ext-ft2` (pseudo labels during pre-training). In the official 10-class validation split, `ext-ft2` achieves 0.98 accuracy, a 50% relative error reduction over the strong baseline. Early pseudo-label integration significantly improves ambiguous classes such as *Shopping mall* and *Public square*.**

I. Introduction

Acoustic Scene Classification (ASC) is a fundamental task in environmental sound analysis, with the aim of recognizing the acoustic characteristics of various environments [1]. It has broad applications in smart city surveillance, intelligent devices, and urban planning. Recent ASC research has focused on data-efficient and low-complexity models [2], [3], [4], [5], yet many existing approaches implicitly assume that the acoustic environment is stationary, ignoring variability caused by geographical or temporal factors. In practice, acoustic scenes within the same category can vary substantially depending on the city and recording time. For example, a "public square" on a weekday morning can sound entirely different from the same location on a weekend night, with cultural differences between cities further amplifying this variability [6]. Previous work has shown that incorporating temporal information can improve ASC performance [7], but conventional systems often neglect this context, limiting their generalizability. The ICME 2024 Challenge [8] addressed geographic domain shifts, but did not explicitly utilize city names or timestamps.

The APSIPA ASC 2025 Grand Challenge[1] introduces a new environment where each audio clip is accompanied by precise city and timestamp metadata, explicitly encouraging their use to improve classification. The dataset spans various Chinese cities and recording times, requiring models to handle urban-scale variation, cultural diversity, and time-of-day differences. Similarly to ICME 2024, the challenge adopts a semi-supervised learning framework [9], [10], reflecting real-world scenarios with limited labeled and abundant unlabeled data.

To address this, we propose a *geography-aware semi-supervised* approach that takes advantage of the two external datasets allowed. One dataset provides city labels directly, while the other lacks them; for the latter, we generate pseudo-city labels by grouping users into consistent city clusters based on their IDs. We further augment scene labels by harmonizing synonyms with the challenge taxonomy and adding auxiliary classes absent from the official set. These enriched labels, combined with city/pseudo-city embeddings, enable the model to learn geographic and contextual patterns during pre-training. Two training schedules are explored: `ext-ft1`, which introduces pseudolabels only after pre-training, and `ext-ft2`, which integrates them during pre-training. As shown in our experiments, early pseudo-label integration in `ext-ft2` achieves the highest accuracy (0.98), representing a 50% relative error reduction compared to a strong baseline.

II. SE-Trans baseline model

The baseline system is based on the Squeeze-and-Excitation Transformer (SE-Trans) model proposed by Bai et al. [11], originally developed for acoustic scene classification and other environmental sound recognition tasks. SE-Trans combines two key mechanisms: (i) a channel-wise attention module (Squeeze-and-Excitation block) that dynamically reweights feature channels to emphasize important acoustic cues and (ii) a Transformer encoder that captures long-range temporal dependencies in the input sequence.

Given an input log-mel spectrogram, the SE block first computes channel importance via global average pooling and generates scaling coefficients to reweight each channel. The reweighted features are then aggregated and fed into the Transformer encoder, which models the temporal context using multi-head self-attention. Finally, a classification head outputs scene labels, optionally incorporating location and time embeddings to model city- or time-dependent acoustic patterns.

This architecture highlights salient frequency channels while modeling the temporal structure, leading to improved classification performance with minimal computational overhead.

[1] https://www.apsipa2025.org/wp/grand-challenge/

TABLE I
ENVIRONMENT LABEL MAPPING FOR EXTERNAL DATASETS.

Original label	Mapped label
Subway	Metro
Subway station	Metro station
Street	Traffic street
Park	Urban park

Data augmentation such as FMix is also used to enhance generalization.

III. Proposed Approach

We design a semi-supervised ASC pipeline that (i) harmonizes labels across the two allowed external datasets, (ii) injects city metadata (or pseudo-city metadata when unavailable), and (iii) pretrains SE-Trans with these enriched labels before adapting the model to the challenge label space. In the following, we detail the data sources, label construction, input encoding, and the two training schedules (`ext-ft1` and `ext-ft2`).

Only two external datasets are used, as permitted by the challenge. TAU Urban Acoustic Scenes 2020 Mobile Development [12] and CochlScene [13]. The amount of data with valid timestamps is limited; therefore, we focus on expanding and exploiting *geographical (city) information*, which is comparatively easier to standardize across sources.

A. Label harmonization and expansion

Let $\mathcal{Y}^{\mathrm{CAS}}$ denote the 10 scene classes in the target challenge. To maximize reuse of external data, we construct an *expanded* scene label set $\tilde{\mathcal{Y}}$ for pre-training: We normalize synonymous labels in the external corpora to the CAS taxonomy, as shown in Table I. In addition, to increase scene diversity during pre-training, we augment the label set with auxiliary classes $\mathcal{Y}^{\mathrm{aux}}$ that are absent in $\mathcal{Y}^{\mathrm{CAS}}$: {*Metro station*, *Street pedestrian*, *Tram*, *Cafe*, *Car*, *Crowded indoor*, *Elevator*, *Kitchen*, *Residential area*, and *Restroom*}. The pre-training uses $\tilde{\mathcal{Y}} = \{\mathcal{Y}^{\mathrm{CAS}}, \mathcal{Y}^{\mathrm{aux}}\}$; in the adaptation stage, we reconcile the classifier to $\mathcal{Y}^{\mathrm{CAS}}$ (Sec. III-D).

B. City metadata construction

Let $\mathcal{C}^{\mathrm{CAS}}$ denote the 22 cities in the target challenge. We exploit city information using the extended city set $\tilde{\mathcal{C}} = \{\mathcal{C}^{\mathrm{CAS}}, \mathcal{C}^{\mathrm{TAU}}, \mathcal{C}^{\mathrm{Cochl}}\}$ as an input embedding to SE-Trans: TAU provides explicit city tags; we directly attach the 10-city labels $\mathcal{C}^{\mathrm{TAU}}$, {*Barcelona, Helsinki, Lisbon, London, Lyon, Milan, Paris, Prague, Stockholm, Vienna*}, to each clip and learn a city embedding. CochlScene does not have explicit city labels, but recordings are known to be from Korean cities with 831 distinct users. To inject stable geographical priors without external resources, we define *pseudo-city* tokens per user bucket. Concretely, we map users into fixed buckets by the hundreds range of their user ID (e.g., 0–99, 100–199, ...) and assign a consistent pseudo-city string to each bucket, chosen from $\mathcal{C}^{\mathrm{Cochl}}$: {*KOREA-user00, KOREA-user01, ..., KOREA-user08*}. This guarantees that all clips from the same bucketed user share the same pseudo-city embedding, promoting intra-bucket acoustic consistency while preserving privacy.

C. Input encoding: audio, city, and time

The audio is converted to log-mel spectrograms and fed to SE-Trans. The city information is injected via a learned *city embedding* that is concatenated or fused with the audio representation at the input of the encoder (after the initial integration of SE-Trans). Although timestamps are available in the target dataset, we set the *time embedding* to the zero vector for all stages in this work, isolating the impact of geographical priors and avoiding potential timestamp domain mismatch during pre-training ("timestamps with limited coverage").

D. Pretraining and model alignment

We pretrain SE-Trans on the union of TAU and CochlScene with the expanded label space $\tilde{\mathcal{Y}}$ and the city (or pseudo-city) embeddings described above. After pre-training, we *align* the model in the challenge label space; For city embedding reconciliation, we retain only the 22 city embeddings that appear in the CAS dataset and remove the rest. For classifier reconciliation, we replace the final classification layer to output the 10 official CAS classes, removing the auxiliary scene heads that were added for pre-training.

E. Semi-supervised schedules

We consider two training schedules that differ in whether pseudo-labels are already used during the *expanded-data pre-training* stage:

a) `ext-ft1` (three-step training):

1) Expanded-data pre-training (no pseudo-labels): Train on TAU and CochlScene using only available labels with city/pseudo-city embeddings.
2) Model alignment: Apply the reconciliation in Sec. III-D.
3) CAS fine-tuning: Fine-tuning on labeled CAS data; then, following the official baseline, generate pseudo-labels on unlabeled CAS and perform a second fine-tuning on (*CAS labeled* ∪ *CAS pseudo-labeled*).

b) `ext-ft2` (four-step training):

1) Expanded-data pre-training: Same as `ext-ft1`.
2) Expanded-data + CAS pseudo-label joint training: Augment the above with pseudo-labeled CAS clips (generated by the baseline procedure) so that the model is exposed earlier to in-domain acoustics while retaining city/pseudo-city supervision from external data.
3) Model alignment: As in Sec. III-D.
4) CAS fine-tuning with pseudo-labels: Fine-tune in labeled CAS and then in (*CAS labeled* ∪ *CAS pseudo-labeled*) to consolidate the decision boundaries in the domain.

F. Rationale and implementation notes

- Why pre-train with expanded labels? Auxiliary classes expand acoustic diversity and stabilize representation learning; later removal of auxiliary heads avoids label mismatch at evaluation time.
- Why are city embeddings? City (or pseudo-city) tokens act as structured priors for persistent background characteristics (infrastructure, traffic mix, cultural activity

patterns), helping to disambiguate acoustically similar scenes across locales.
- Why zero-time embedding is used? Given sparse and potentially mismatched timestamp coverage during pre-training, zeroing time embeddings cleanly ablates temporal priors, isolating the gains attributable to geography.
- Pseudo-label generation. We follow the official baseline procedure to create pseudo-labels on unlabeled CAS; no external models beyond the challenge rules are used.

IV. Experimental Setup

A. Dataset

The development dataset provided by the APSIPA ASC 2025 challenge is based on the Chinese Acoustic Scene (CAS) 2023 collection. It contains approximately 24 hours of audio recordings collected between April and September 2023 across 22 cities $\mathcal{C}^{\text{CAS}}$ in China: {*Xi'an, Xianyang, Changchun, Jinan, Hefei, Sanya, Nanning, Haikou, Guilin, Guangzhou, Chongqing, Shenyang, Beijing, Baishan, Taiyuan, Tianjin, Nanchang, Shanghai, Luoyang, Liupanshui, Shangrao, Dandong*}. Each clip is annotated with *city metadata* and a precise *timestamp* (year–month–day–hour–minute–second).

The dataset covers 10 classes of acoustic scene $\mathcal{Y}^{\text{CAS}}$: {*Bus, Airport, Metro, Restaurant, Shopping mall, Public square, Urban park, Traffic street, Construction site, Bar*}. The audio content is identical to that in the ICME 2024 challenge, but the inclusion of city and time metadata enables new context-aware classification strategies. The labeled portion consists of $\sim$4 hours of audio, while the remaining $\sim$20 hours are unlabeled.

B. Challenge constraints

The rules impose the following restrictions:

- Only two external datasets can be used for pre-training: TAU Urban Acoustic Scenes 2020 Mobile Development [12] and CochlScene [13].
- The use of proprietary or non-public data is prohibited.
- Ensemble models are not allowed; evaluation must be based on a single model.
- Large-scale pretrained audio or audio-language models (e.g., Whisper, Qwen-Audio, LTU) are prohibited.
- The primary research focus should be on the use of city and time metadata for performance improvement.

C. Semi-supervised framework

The challenge adopts a multimodal semi-supervised learning framework, where the model processes:

1) Acoustic features of the raw waveform (log-mel spectrograms).
2) City embeddings derived from the provided metadata.
3) Time embeddings derived from the timestamp (not used in our experiments; set to zero vectors).

Following the baseline pipeline, training proceeds as follows:

1) Learning an initial model in the labeled subset ($\sim$4 hours).
2) Generating pseudo-labels for the unlabeled portion ($\sim$20 hours).
3) Re-training the model on the union of labeled and pseudo-labeled data.

The backbone architecture is SE-Trans (Sec. II), which incorporates city/time embeddings into its transformer-based acoustic encoder.

TABLE II
Validation accuracy of baseline and proposed schedules.

Model	Acc.	Description
Baseline (20 ep.)	0.94	Default setting
Baseline (100 ep.)	0.96	Extended training
Ext. + FT (`ext-ft1`)	0.96	No pseudo-labels in pre-training
Ext. + FT (`ext-ft2`)	**0.98**	Pseudo-labels in pre-training, best acc.

V. Results

Metrics follow the official challenge protocol (Sec. IV). The backbone architecture is SE-Trans [11], and the training schedules `ext-ft1` and `ext-ft2` are detailed in Sec. III.

A. Overview

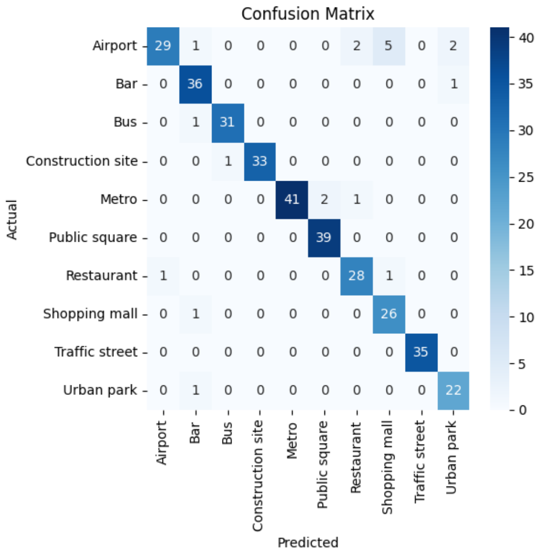
Table II summarizes the validation accuracy in the official 10-class split. The strongest baseline, trained for 100 epochs, achieves 0.96 accuracy, while the `ext-ft2` schedule reaches 0.98, reducing the error rate from 4% to 2%, corresponding to a 50% relative error reduction. The greatest improvements are observed in acoustically ambiguous classes such as *Shopping mall*, *Restaurant*, and *Public square*. A detailed class-wise analysis is provided in the Appendix.

B. Confusion Matrices

The confusion matrices in Figs. 1–4 reveal clear trends. For baseline models (Figs. 1 and 2), most misclassifications occur in categories with overlapping background acoustics, such as *Urban park*, *Bar*, and *Shopping mall*. Extending training to 100 epochs improves recall for *Airport*, *Restaurant*, and *Shopping mall*. With the `ext-ft1` schedule (Fig. 3), decision boundaries become sharper and most classes achieve approximately 0.95 in both precision and recall, although minor confusion remains between scenes of nature and open public space. The `ext-ft2` schedule (Fig. 4) further reduces off-diagonal errors, producing an almost perfectly diagonal matrix and near-perfect classification for *Shopping mall*, *Restaurant*, and *Public square*.

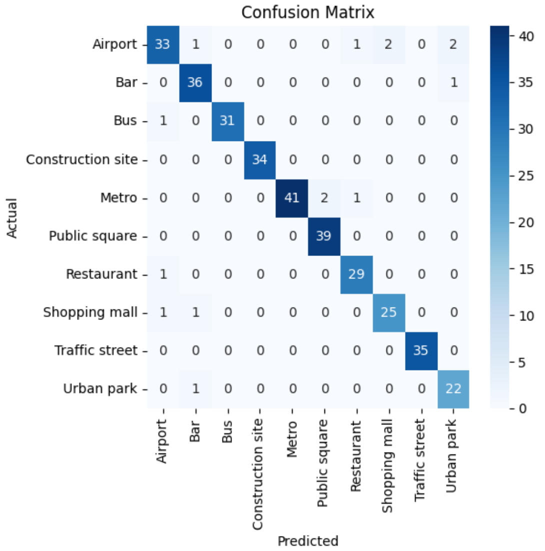
C. Learning Dynamics

Figures 5–8 present the validation accuracy and training loss curves. In the baseline (Fig. 5), Phase 1 training on labeled data yields steady loss reduction and accuracy improvement, but Phase 2 pseudo-label fine-tuning offers minimal additional gain, likely due to the limited reliability of pseudo-labels without stronger priors. For the expanded-data pre-training setup (Fig. 6), both phases show continuous improvement, suggesting that geographic priors from city and pseudo-city embeddings enhance pseudo-label stability. In `ext-ft1`



	Airport	Bar	Bus	Construction site	Metro	Public square	Restaurant	Shopping mall	Traffic street	Urban park
Airport	29	1	0	0	0	0	2	5	0	2
Bar	0	36	0	0	0	0	0	0	0	1
Bus	0	1	31	0	0	0	0	0	0	0
Construction site	0	0	1	33	0	0	0	0	0	0
Metro	0	0	0	0	41	2	1	0	0	0
Public square	0	0	0	0	0	39	0	0	0	0
Restaurant	1	0	0	0	0	0	28	1	0	0
Shopping mall	0	1	0	0	0	0	0	26	0	0
Traffic street	0	0	0	0	0	0	0	0	35	0
Urban park	0	1	0	0	0	0	0	0	0	22

Fig. 1. Confusion matrix for the baseline model after 20 training epochs. Performance is strong for acoustically distinctive classes such as *Construction site* and *Traffic street*, but recall is lower for ambiguous classes like *Urban park* and *Bar*.

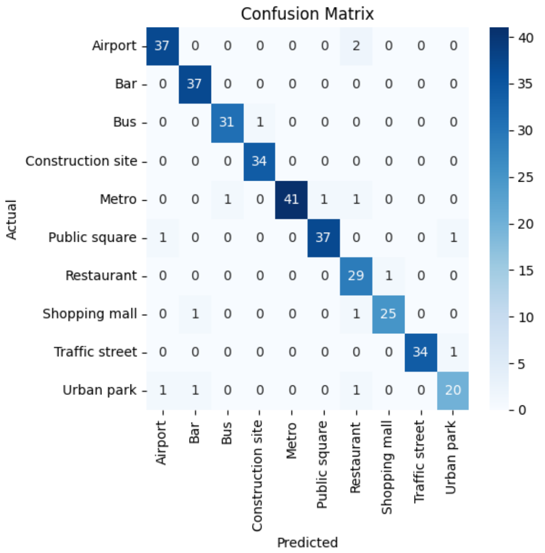


	Airport	Bar	Bus	Construction site	Metro	Public square	Restaurant	Shopping mall	Traffic street	Urban park
Airport	33	1	0	0	0	0	1	2	0	2
Bar	0	36	0	0	0	0	0	0	0	1
Bus	1	0	31	0	0	0	0	0	0	0
Construction site	0	0	0	34	0	0	0	0	0	0
Metro	0	0	0	0	41	2	1	0	0	0
Public square	0	0	0	0	0	39	0	0	0	0
Restaurant	1	0	0	0	0	0	29	0	0	0
Shopping mall	1	1	0	0	0	0	0	25	0	0
Traffic street	0	0	0	0	0	0	0	0	35	0
Urban park	0	1	0	0	0	0	0	0	0	22

Fig. 2. Confusion matrix for the baseline model after 100 training epochs. Increasing training epochs improves recall for *Airport*, *Restaurant*, and *Shopping mall*, while maintaining high accuracy for clearly defined acoustic scenes.

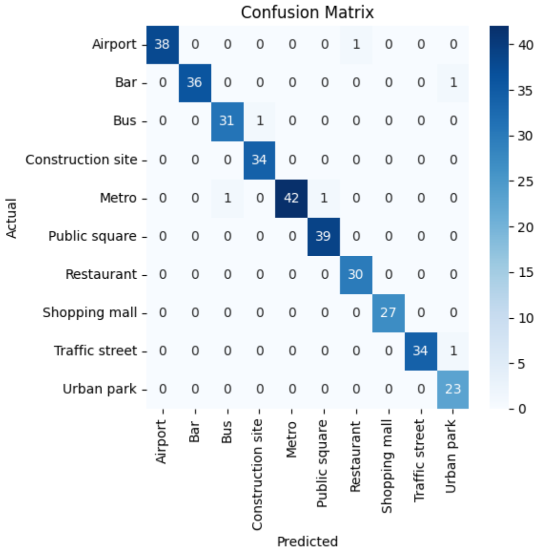
(Fig. 7), the absence of pseudo-labels during pre-training leads to earlier saturation of accuracy gains; while subsequent CAS fine-tuning improves results, the final accuracy remains at 0.96. In contrast, `ext-ft2` (Fig. 8) incorporates pseudo-labels already during the expanded-data stage, enabling early, geography-aware exposure to in-domain acoustics. This approach maintains accuracy improvements through both phases, ultimately achieving the best performance of 0.98.

City and pseudo-city embeddin gs prove effective in disambiguating acoustically similar scenes that vary systematically across locations. Early introduction of pseudo-labels in



	Airport	Bar	Bus	Construction site	Metro	Public square	Restaurant	Shopping mall	Traffic street	Urban park
Airport	37	0	0	0	0	0	2	0	0	0
Bar	0	37	0	0	0	0	0	0	0	0
Bus	0	0	31	1	0	0	0	0	0	0
Construction site	0	0	0	34	0	0	0	0	0	0
Metro	0	0	1	0	41	1	1	0	0	0
Public square	1	0	0	0	0	37	0	0	0	1
Restaurant	0	0	0	0	0	0	29	1	0	0
Shopping mall	0	1	0	0	0	0	1	25	0	0
Traffic street	0	0	0	0	0	0	0	0	34	1
Urban park	1	1	0	0	0	0	1	0	0	20

Fig. 3. Confusion matrix for the proposed `ext-ft1` schedule. Most classes achieve precision and recall around 0.95 or higher, with clear improvements over the baseline; however, slight recall drops remain for *Urban park* and *Restaurant*.



	Airport	Bar	Bus	Construction site	Metro	Public square	Restaurant	Shopping mall	Traffic street	Urban park
Airport	38	0	0	0	0	0	1	0	0	0
Bar	0	36	0	0	0	0	0	0	0	1
Bus	0	0	31	1	0	0	0	0	0	0
Construction site	0	0	0	34	0	0	0	0	0	0
Metro	0	0	1	0	42	1	0	0	0	0
Public square	0	0	0	0	0	39	0	0	0	0
Restaurant	0	0	0	0	0	0	30	0	0	0
Shopping mall	0	0	0	0	0	0	0	27	0	0
Traffic street	0	0	0	0	0	0	0	0	34	1
Urban park	0	0	0	0	0	0	0	0	0	23

Fig. 4. Confusion matrix for the proposed `ext-ft2` schedule. This configuration achieves the highest overall accuracy (0.98), with recall ≥ 0.95 for nearly all classes. Predictions are notably more stable for acoustically ambiguous categories such as *Shopping mall*, *Restaurant*, and *Public square*.

`ext-ft2` synergizes with geography-aware pre-training, leading to more reliable pseudo-labels and stronger downstream fine-tuning performance. Remaining errors are primarily found in nature-like environments such as *Urban park*, indicating potential benefits from incorporating temporal embeddings or explicit scene-attribute modeling in future work.

VI. Conclusion

In this work, we addressed the challenge of acoustic scene classification under geographic and temporal variability using the APSIPA ASC 2025 dataset. We proposed two extended-

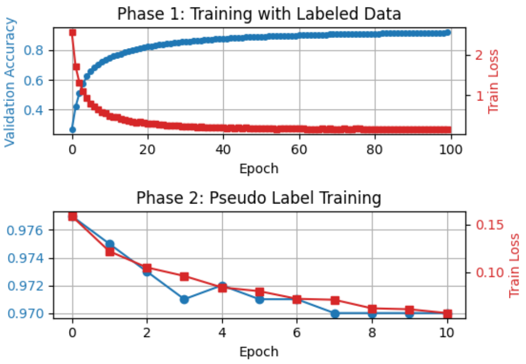


Fig. 5. Validation accuracy and training loss for the baseline model. Accuracy improves steadily in Phase 1 but saturates quickly in Phase 2, indicating limited benefit from pseudo-labels without stronger priors.

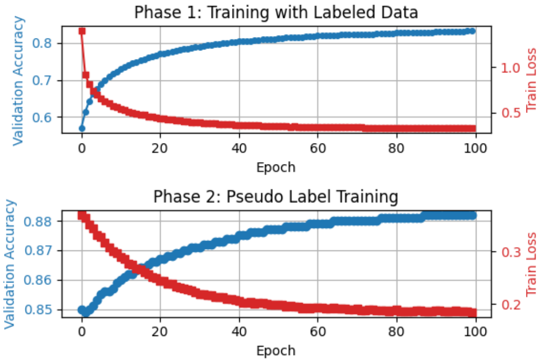


Fig. 6. Validation accuracy and training loss for expanded-data pre-training. Accuracy increases consistently in both phases, demonstrating the stabilizing effect of external data and city/pseudo-city embeddings.

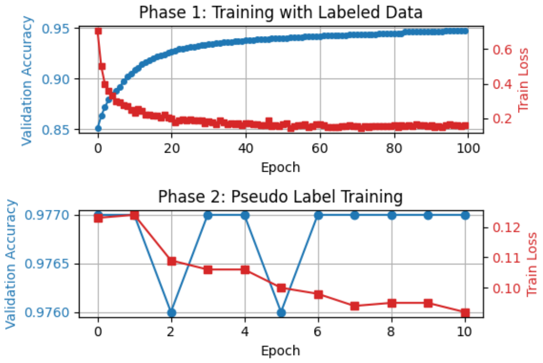


Fig. 7. Validation accuracy and training loss for the `ext-ft1` schedule. Without pseudo-labels in pre-training, Phase 2 gains are modest, and the final accuracy (0.96) is lower than that of `ext-ft2`.

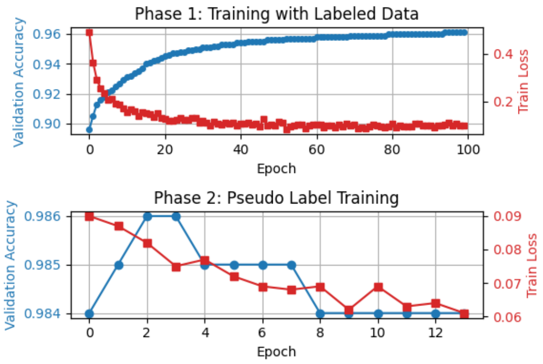


Fig. 8. Validation accuracy and training loss for the `ext-ft2` schedule. Early integration of CAS pseudo-labels during pre-training yields high Phase 1 accuracy and continued improvement in Phase 2, achieving the best final accuracy (0.98).

TABLE III
BASELINE PERFORMANCE (20 EPOCHS).

Class	Precision	Recall	F1-Score	Support
Airport	0.97	0.74	0.84	39
Bar	0.90	0.97	0.94	37
Bus	0.97	0.97	0.97	32
Construction site	1.00	0.97	0.99	34
Metro	1.00	0.93	0.96	44
Public square	0.95	1.00	0.97	39
Restaurant	0.90	0.93	0.92	30
Shopping mall	0.81	0.96	0.88	27
Traffic street	1.00	1.00	1.00	35
Urban park	0.88	0.96	0.92	23
Accuracy			**0.94**	340
Macro avg	0.94	0.94	0.94	340
Weighted avg	0.95	0.94	0.94	340

TABLE IV
BASELINE PERFORMANCE (100 EPOCHS).

Class	Precision	Recall	F1-Score	Support
Airport	0.92	0.85	0.88	39
Bar	0.92	0.97	0.95	37
Bus	1.00	0.97	0.98	32
Construction site	1.00	1.00	1.00	34
Metro	1.00	0.93	0.96	44
Public square	0.95	1.00	0.97	39
Restaurant	0.94	0.97	0.95	30
Shopping mall	0.93	0.93	0.93	27
Traffic street	1.00	1.00	1.00	35
Urban park	0.88	0.96	0.92	23
Accuracy			**0.96**	340
Macro avg	0.95	0.96	0.95	340
Weighted avg	0.96	0.96	0.96	340

data training schedules, `ext-ft1` and `ext-ft2`, which leverage city/pseudo-city embeddings and harmonized scene labels from two allowed external corpora. Our results show that incorporating pseudo-labels *during* the expanded-data pre-training stage (`ext-ft2`) yields the highest validation accuracy (0.98), representing a 50% relative error reduction compared to a strong 100-epoch baseline. Confusion matrix analysis confirms that the largest gains occur in acoustically ambiguous classes such as *Shopping mall*, *Restaurant*, and *Public square*, while learning-curve trends demonstrate the stabilizing effect of early, geography-aware pseudo-label integration.

Although residual errors remain in nature-like ambiences (e.g., *Urban park*), our findings indicate that explicitly modeling temporal context or additional scene attributes could further improve performance. Overall, the proposed geography-aware semi-supervised framework provides a simple yet effective way to enhance ASC generalization across cities and time, and can be extended to other domain-shift scenarios in environmental sound recognition.

TABLE V
EXTENDED-DATA FINE-TUNING (EXT-FT1), WITHOUT PSEUDO-LABELS IN PRE-TRAINING.

Class	Precision	Recall	F1-Score	Support
Airport	0.95	0.95	0.95	39
Bar	0.95	1.00	0.97	37
Bus	0.97	0.97	0.97	32
Construction site	0.97	1.00	0.99	34
Metro	1.00	0.93	0.96	44
Public square	0.97	0.95	0.96	39
Restaurant	0.85	0.97	0.91	30
Shopping mall	0.96	0.93	0.94	27
Traffic street	1.00	0.97	0.99	35
Urban park	0.91	0.87	0.89	23
Accuracy			**0.96**	340
Macro avg	0.95	0.95	0.95	340
Weighted avg	0.96	0.96	0.96	340

TABLE VI
EXTENDED-DATA FINE-TUNING (EXT-FT2), WITH PSEUDO-LABELS IN PRE-TRAINING.

Class	Precision	Recall	F1-Score	Support
Airport	1.00	0.97	0.99	39
Bar	1.00	0.97	0.99	37
Bus	0.97	0.97	0.97	32
Construction site	0.97	1.00	0.99	34
Metro	1.00	0.95	0.98	44
Public square	0.97	1.00	0.99	39
Restaurant	0.97	1.00	0.98	30
Shopping mall	1.00	1.00	1.00	27
Traffic street	1.00	0.97	0.99	35
Urban park	0.92	1.00	0.96	23
Accuracy			**0.98**	340
Macro avg	0.98	0.98	0.98	340
Weighted avg	0.98	0.98	0.98	340

APPENDIX

Tables III–VI present the precision, recall and F1 scores per class for the baseline and proposed training schedules. Table III corresponds to the official baseline script with 20 epochs, while Table IV shows the result after increasing the training epochs to 100. Tables V and VI present the results for `ext-ft1` (no pseudo-labels in pre-training) and `ext-ft2` (pseudo-labels in pre-training), respectively.

a) Additional observations from per-class results.: From Tables III and IV, extending the baseline training from 20 to 100 epochs yields notable recall improvements for *Airport* (+0.11), *Restaurant* (+0.04), and especially *Shopping mall* (+0.12), confirming that longer training helps the model better fit acoustically ambiguous categories. However, even with extended training, precision for *Airport* decreases slightly (0.97 → 0.92), suggesting possible overfitting or increased confusion with other transportation-related scenes.

Comparing the baseline (Table IV) with `ext-ft1` (Table V), improvements are evident for *Bar* (+0.03 F1), *Shopping mall* (+0.01 F1), and *Public square* (+0.01 F1). Yet, `ext-ft1` underperforms in *Restaurant* (F1 drop from 0.95 to 0.91) and *Urban park* (0.92 → 0.89), probably due to the absence of pseudo-label guidance during pre-training.

The `ext-ft2` schedule (Table VI) delivers consistent gains in almost all classes. Large improvements appear in *Airport* (0.88 → 0.99 F1), *Shopping mall* (0.93 → 1.00 F1), and *Restaurant* (0.95 → 0.98 F1) compared to the strong baseline. In particular, `ext-ft2` also corrects the recall drop in *Urban park*, reaching 1.00 recall while increasing precision to 0.92. These gains confirm that early integration of in-domain pseudo-labels during pre-training stabilizes decision boundaries for acoustically overlapping classes and yields near-ceiling performance across the board.

Overall, the per-class analysis reinforces the conclusions in Sec. V: city/pseudo-city embeddings and early pseudo-label integration (`ext-ft2`) not only boost macro-level accuracy but also systematically enhance recognition of the most challenging scene categories.

REFERENCES

[1] B. Ding, T. Zhang, C. Wang, G. Liu, J. Liang, R. Hu, Y. Wu, and D. Guo, "Acoustic scene classification: A comprehensive survey," *Expert Syst. Appl.*, vol. 238, no. PB, Mar. 2024. [Online]. Available: https://doi.org/10.1016/j.eswa.2023.121902

[2] F. Schmid, P. Primus, T. Heittola, A. Mesaros, I. Martín-Morató, K. Koutini, and G. Widmer, "Data-efficient low-complexity acoustic scene classification in the dcase 2024 challenge," 2024. [Online]. Available: https://arxiv.org/abs/2405.10018

[3] Y. Li, J. Tan, G. Chen, J. Li, Y. Si, and Q. He, "Low-complexity acoustic scene classification using parallel attention-convolution network," 2024. [Online]. Available: https://arxiv.org/abs/2406.08119

[4] B. Han, W. Huang, Z. Chen, A. Jiang, P. Fan, C. Lu, Z. Lv, J. Liu, W.-Q. Zhang, and Y. Qian, "Data-efficient low-complexity acoustic scene classification via distilling and progressive pruning," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2025, p. 1–5. [Online]. Available: http://dx.doi.org/10.1109/ICASSP49660.2025.10890296

[5] J. J. S. Yeo, E.-L. Tan, J. Bai, S. Peksi, and W.-S. Gan, "Data efficient acoustic scene classification using teacher-informed confusing class instruction," 2024. [Online]. Available: https://arxiv.org/abs/2409.11964

[6] J. Abeßer, Z. Liang, and B. Seeber, "Sound recurrence analysis for acoustic scene classification," *EURASIP J. Audio Speech Music Process.*, vol. 2025, no. 1, Jan. 2025. [Online]. Available: https://doi.org/10.1186/s13636-024-00390-2

[7] W. Wang, W. Wang, M. Sun, and C. Wang, "Acoustic scene analysis with multi-head attention networks," 2020. [Online]. Available: https://www.amazon.science/publications/acoustic-scene-analysis-with-multi-head-attention-networks

[8] J. Bai, M. Wang, H. Liu, H. Yin, Y. Jia, S. Huang, Y. Du, D. Zhang, D. Shi, W.-S. Gan, M. D. Plumbley, S. Rahardja, B. Xiang, and J. Chen, "Description on ieee icme 2024 grand challenge: Semi-supervised acoustic scene classification under domain shift," 2024. [Online]. Available: https://arxiv.org/abs/2402.02694

[9] W. Huang, A. Jiang, B. Han, X. Zheng, Y. Qiu, W. Chen, Y. Liang, P. Fan, W.-Q. Zhang, C. Lu, X. Chen, J. Liu, and Y. Qian, "Semi-supervised acoustic scene classification with test-time adaptation," in *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2024, pp. 1–5.

[10] Y. Liang, W. Chen, A. Jiang, Y. Qiu, X. Zheng, W. Huang, B. Han, Y. Qian, P. Fan, W.-Q. Zhang, L. Cheng, J. Liu, and X. Chen, "Improving acoustic scene classification via self-supervised and semi-supervised learning with efficient audio transformer," in *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2024, pp. 1–6.

[11] 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.

[12] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 9–13. [Online]. Available: https://dcase.community/documents/workshop2018/proceedings/DCASE2018Workshop_Mesaros_8.pdf

[13] I.-Y. Jeong and J. Park, "Cochlscene: Acquisition of acoustic scene data using crowdsourcing," 2022. [Online]. Available: https://arxiv.org/abs/2211.02289