A Semi-Supervised Acoustic Scene Classification Network Based on Multi-Modal Information Fusion

Junkang Yang[1], Hongqing Liu[2], Liming Shi[2], Lu Gan[3], Hiromitsu Nishizaki[1*] and Chee Siang Leow[1]

1 Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences, University of Yamanashi

2 School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications

3 College of Engineering, Design and Physical Science, Brunel University

*Corresponding Author Email: hnishi@yamanashi.ac.jp

UNIVERSITY OF YAMANASHI

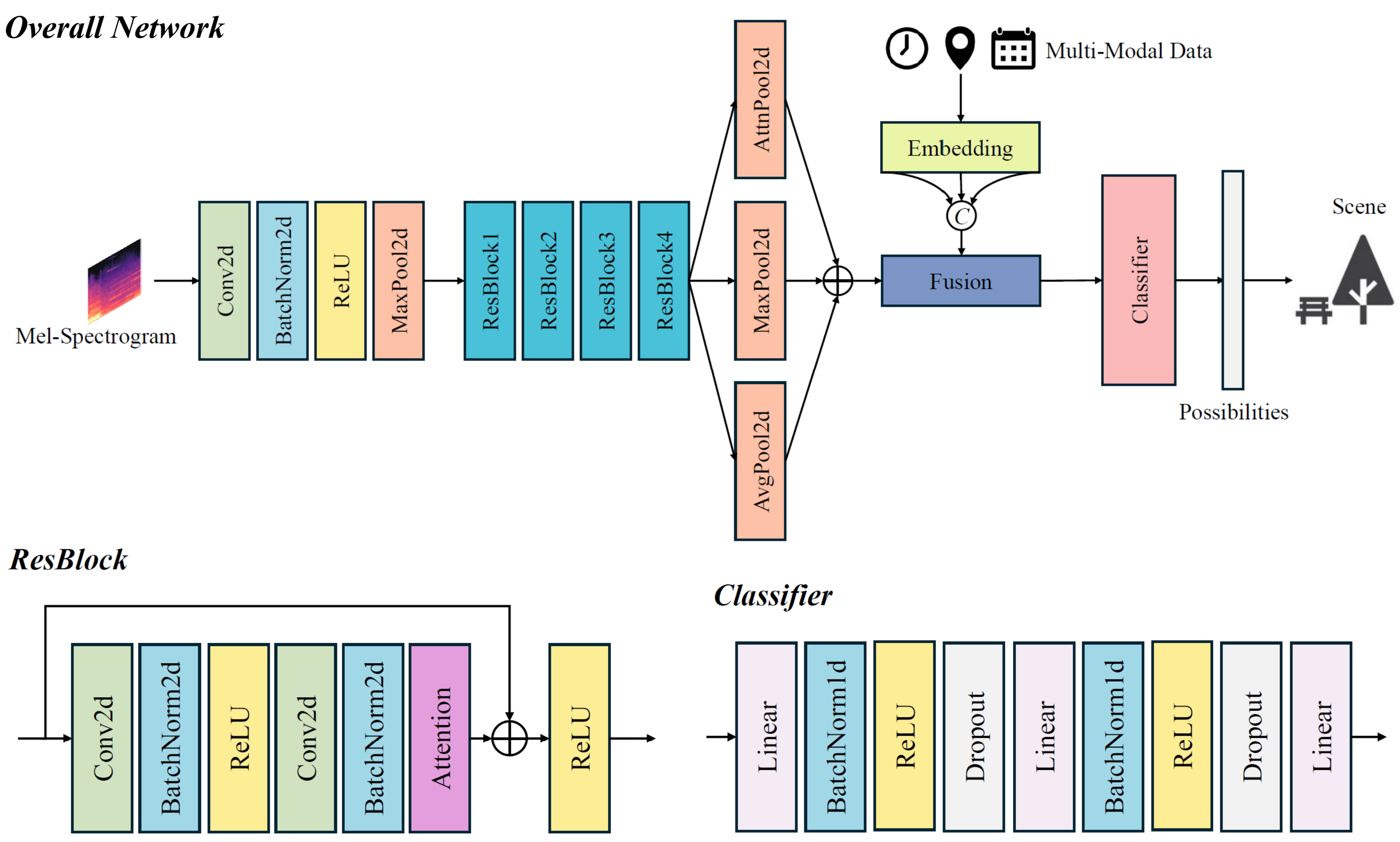
APSIPA ASC 2025

Multimedia Information Processing NISHIZAKI-LEOW LAB.

Introduction

Acoustic scene classification (ASC) plays a crucial role in computational audition, with applications in smart cities and audio devices. Traditional ASC systems often treat scenes as static categories, ignoring variations across locations and time and limiting their real-world effectiveness, as acoustic characteristics differ between cities and temporal contexts. To bridge this gap, the APSIPA ASC 2025 Challenge introduces city-level and timestamp metadata, advancing context-aware ASC solutions. While prior work addressed geographic domain shifts, this challenge uniquely explores city identity and temporal cues in semi-supervised learning, where labeled data is scarce yet crucial for industrial applications like urban sound monitoring. In this paper, we propose a novel approach leveraging spatiotemporal metadata alongside audio features, employing feature representation, domain adaptation, and contextual fusion techniques to enhance ASC accuracy across diverse urban settings and time periods. Our experiments demonstrate significant improvements while maintaining generalizability.

Proposed Network



The proposed model processes audio spectrograms (**[batchsize, 1, frames, bins]**) via an initial **7×7 convolution**, **BN+ReLU**, and **3×3 max pooling**, followed by 4 residual blocks with **3×3 convolutions** and channel-spatial attention. Each block expands channels (**[64→128→256→512]**) while performing **stride-2** downsampling for deep time-frequency feature extraction. Next, an innovative pooling fusion stage combines spatial-attention-weighted summation, global average pooling, and max pooling into a robust **512D** audio representation. For multimodal input, this vector is fused with location embeddings and processed temporal features, then compressed to **256D** via a fusion layer (BN+ReLU+dropout). Finally, a 3-layer classifier (**128D→64D→output**) with BN, ReLU, and dropout (**0.4**) generates class probabilities.

code is released on github

Data & Augmentation

- Data

For pre-training, we utilize the **TAU Urban Acoustic Scenes 2020 Mobile** and **CochlScene datasets**, converting them to match the challenge's format by standardizing audio clips to **44.1 kHz** and **10-second** segments. Since the scene categories differ across datasets, we manually relabel the pre-training data to align with the challenge's taxonomy (details in table 1). The dataset is split into 80% training and 20% validation subsets.

- Augmentation

SpecAugment and Mixup (α=1.0).

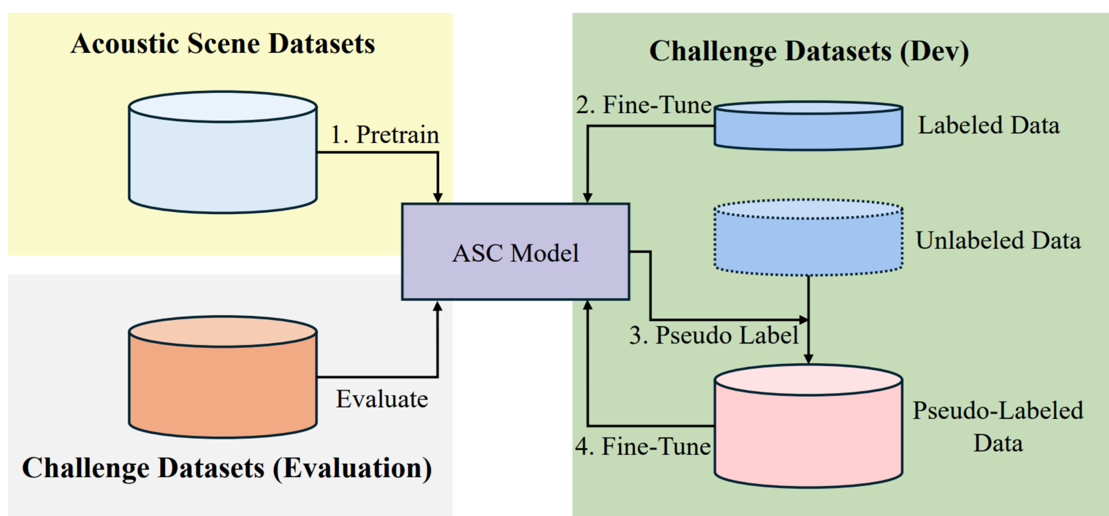
TABLE I
RELATIONSHIP BETWEEN THE LABEL TYPE OF CHALLENGE DATA AND OUR PRE-TRAINING DATA

	Challenge Data	TAU Urban Acoustic Scenes 2020	CochlScene
Labels	bus	bus	bus
	airport	airport	-
	metro	metro station, metro	subway, subway station
	restaurant	-	restaurant
	shopping mall	shopping mall	-
	public square	public square	-
	urban park	park	park
	traffic street	street traffic	street
	construction site	-	-
	bar	-	cafe

Training Details

The model is trained for up to 1000 epochs (early stopping after 20 epochs without validation improvement) using Adam optimizer (LR=5×10⁻⁴, decayed by 0.9 every 2 epochs) and cross-entropy loss. Data is split 80:20 (train/val) with fixed random seed 1234, batch size 64.

Our semi-supervised training consists of four stages:
(1) Pre-training on labeled data with SpecAugment and Mixup augmentation;
(2) Supervised fine-tuning on task-specific labeled data, saving the best checkpoint;
(3) Pseudo-labeling, where the fine-tuned model predicts labels for unlabeled data;
(4) Pseudo-label training, retraining the model on the combined labeled and pseudo-labeled data using the same hyperparameters.



Results

TABLE II
TRAINING ACCURACY ON VALIDATION DATA OF DIFFERENT STAGES.

Stage	Accuracy (Average)
Pre-Training	93.70%
First Round Fine-Tuning	87.00%
Second Round Fine-Tuning	87.60%

TABLE IV
COMPAXITY ANALYSIS OF PROPOSED MODEL.

Item	Value
#Params	21.65M
MACs	2.34G
CPU Inference Time	40ms

TABLE III
FINAL RESULTS ON EVALUATION DATA.

Item	Accuracy
Bus	0.440
Airport	0.693
Metro	0.920
Restaurant	0.750
Shoppingmall	0.580
Public square	0.040
Urban park	0.700
Traffic street	0.650
Construction site	0.510
Bar	0.850
Macro-accuracy	0.613

As TABLE II shows, during pre-training, model achieved **93.70%** average accuracy on the validation set. After supervised fine-tuning, accuracy initially dropped to **87.00%**. A second fine-tuning round improved performance slightly to **87.60%**.

On challenge evaluation data the macro accuracy reached **61.3%** (detail in TABLE III), ranking **3rd** among all teams.

TABLE IV illustrates the model contains **21.65M** parameters and requires **2.34G MACs**. CPU inference time is only 40ms, the CPU platform is Intel(R) Core(TM) i9-10940X CPU @ 3.30GHz and the input include 3 parts: a 44.1 kHz audio lasts 10 seconds and two character strings representing city and time information.

Cooperated with:

Brunel University London