Abstract—In this technical report, a low-complexity deep learning system for acoustic scene classification (ASC) is presented. The proposed system comprises two main phases: (Phase I) Training a teacher network; and (Phase II) training a student network using distilled knowledge from the teacher. In the first phase, the teacher, which presents a large footprint model, is trained. After training the teacher, the embeddings, which are the feature map of the second last layer of the teacher, are extracted. In the second phase, the student network, which presents a low complexity model, is trained with the embeddings extracted from the teacher. Our experiments conducted on DCASE 2023 Task 1 Development dataset have fulfilled the requirement of low-complexity and achieved the best classification accuracy of 57.4%, improving DCASE baseline by 14.5%.

Clinical relevance—Mixup augmentation, Convolutional Neural Network (CNN), spectrogram, late fusion.

I. INTRODUCTION

To deal with the ASC challenge of mismatched recording devices, the state-of-the-art systems mainly leverage ensemble techniques: Ensemble of spectrogram inputs [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12] or ensemble of different classification models [13], [14], [15]. Although these approaches prove effective to deal with the issue of mismatched recording devices and achieve potential results, they present large model complexity. This lead to challenges to apply ASC components on edge-devices. Recently, DCASE 2021 Task 1A challenge [16] focuses on dealing the issue of high-complexity model. The challenge requires the maximum model complexity of 128 KB. Furthermore, the next challenges of DCASE 2022 Task 1 and DCASE 2023 Task 1 do not allow to use pruning techniques as the pruning parameters still occupy the memory and cost the computation on edge-devices. These challenges also require the maximum MACs (Multiply-Add cumulation) of 30 M.

In this technical report, a low-complexity deep learning frameworks using teacher-student scheme and multiple spectrograms for ASC task is presented.

II. THE TEACHER NETWORK ARCHITECTURE

As Fig. 1 shows, the proposed teacher network can be separated into three main steps: the front-end feature extraction, the online data augmentation, and the convolutional neural network (CNN) based network. Initially, a raw audio signal is firstly transformed into three spectrograms of $128 \times 132$ by using MEL filter [17], Gammatone filter [18], or CQT [17] with the FFT number, Hanna window size, hop size, and the filter number set to 4096, 2048, 326, and 128. Next, we apply delta and delta-delta on these spectrograms to generate three-dimensional spectrograms of $128 \times 128 \times 3$. The original spectrogram, delta, and delta-delta). We then apply the Mixup [19], [20] augmentation method on the spectrograms. We finally feed the augmented spectrograms into back-end deep learning networks for classification, referred to as the teachers. As we use three spectrogram input, we train three individual teachers.

Regarding the teacher architecture, it comprises two main parts: a CNN-based backbone followed by a dense block. The CNN-based backbone, which presents a residual-inception based architecture, is reused from [21], [10]. The dense block comprises two dense layers (Dense Layer 01 and Dense Layer 02), which is shown in the lower part of Fig. 2. After training the teachers, the embeddings, which are the feature map at the first fully connected layer of the dense block (FC (64)), are extracted for training the student networks. The teachers are trained using Entropy loss ($Loss_1$) as shown in Fig. 2.

III. THE STUDENT NETWORK ARCHITECTURE

A student network architecture is presented in Table 1. As we use three spectrograms, we develop three individual students which share the same network architecture. As the configuration shows in Table 2 three student presents 22962 trainable parameters, which occupy 88704 Byte (one parameter is presented by 32 bit) and 29267550 MACs on an edge device. Training the students is presented in Fig. 3 with two loss functions ($Loss_2$ and $Loss_3$). The
### IV. Evaluation Setting and Results

**A. TAU Urban Acoustic Scenes 2022 Mobile, development dataset [22]**

This report presents the results on DCASE 2023 Task 1 Development set, which was proposed in DCASE 2023 challenge [23]. In this challenge, the limitation of model size is set to 128 KB of trainable parameters and the maximum MACs is set to 30 M, not allow to use pruning techniques, and evaluate on 1-second audio segment. The dataset is slightly unbalanced, being recorded across 12 large European cities: Amsterdam, Barcelona, Helsinki, Lisbon, London, Lyon, Madrid, Milan, Prague, Paris, Stockholm, and Vienna. It consists of 10 scene classes: airport, shopping mall (indoors), metro station (underground), pedestrian street, public square, street (traffic), traveling by tram, bus and metro (underground), and urban park. The audio recordings were recorded from 3 different physical devices namely A (10215 recordings), B (749 recordings), C (748 recordings).

Additionally, synthetic data for mobile devices was created based on the original recordings, referred to as S1 (750 recordings), S2 (750 recordings), S3 (750 recordings), S4 (750 recordings), S5 (750 recordings), and S6 (750 recordings).

To evaluate, we follow the DCASE 2023 Task 1 challenge [23], use two subsets known as Training (Train.) and Evaluation (Eval.) from the Development set for training and testing processes, respectively. Notably, two of 12 cities and S4, S5, S6 audio recordings are only presented in the Eval. subset for evaluating the issue of mismatched recording devices and unseen samples.

**B. Network Implementation**

All the CNN architectures are conducted by Tensorflow frameworks. Training CNN architecture use Adam algorithm for the optimization. We run all experiments on the GPU GeForce RTX 2080.

**C. Experimental Results**

The experimental results are presented in Table II. As Table II shows, results on GAM and MEL are competitive.
and outperform the records of DCASE baseline and CQT spectrogram. The ensemble of three models without and with using knowledge distillation achieve accuracy of 56.8% and 57.4%, respectively. The best model using ensemble of multiple spectrogram and knowledge distillation improves the DCASE baseline by 14.5%. The log-loss score of this system presents 1.333 which is less than the DCASE score of 1.575. However, this system requires more memory of 88.7 MB compared with DCASE baseline of 46.5 MB.

V. Conclusion

We have presented a low-complexity system for ASC task, which leverages teacher-student scheme and multiple spectrogram inputs. Our proposed low-complexity ASC system achieves an accuracy of 57.4%, a log-loss score of 1.333, 88.7 KB memory occupation, and 29.27 M MACs.

References

[1] L. Pham, I. McLoughlin, Huy Phan, R. Palaniappan, and A. Mertins, “Deep feature embedding and hierarchical classification for audio scene classification,” in International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1–7.
[2] L. Pham, I. McLoughlin, H. Phan, R. Palaniappan, and Y. Lang, “Bag-of-features models based on C-DNN network for acoustic scene classification,” in Proc. AES, 2019.
[3] Huy Phan, Olivier Y Chén, Lam Pham, Philipp Koch, Maarten De Vos, Ian McLoughlin, and Alfred Mertins, “Spatio-temporal attention pooling for audio scene classification,” in Proc. INTERSPEECH, 2019, pp. 3845–3849.
[4] Lam Pham, Ian McLoughlin, Huy Phan, and Ramaswamy Palaniappan, “A robust framework for acoustic scene classification,” in Proc. INTERSPEECH, 2019, pp. 3634–3638.
[5] D. Ngo, Hao Hoang, A. Nguyen, Tien Ly, and L. Pham, “Sound context classification basing on joint learning model and multi-spectrogram features,” ArXiv, vol. abs/2005.12779, 2020.
[6] Truc Nguyen and Franz Pernkopf, “Acoustic scene classification using a convolutional neural network ensemble and nearest neighbor filters,” in Proc. DCASE, 2018, pp. 34–38.
[7] Yuma Sakashita and Masaki Aono, “Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions,” Tech. Rep., DCASE Challenge, 2018.
[8] Huy Phan, Huy Le Nguyen, Olivier Y. Chén, Lam Pham, Philipp Koch, Ian McLoughlin, and Alfred Mertins, “Multi-view audio and music classification,” in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 611–615.
[9] L. Pham, Huy Phan, T. Nguyen, R. Palaniappan, A. Mertins, and I. McLoughlin, “Robust acoustic scene classification using a multi-spectrogram encoder-decoder framework,” Digital Signal Processing, vol. 110, pp. 102943, 2021.
[10] Lam Pham, Khoa Tran, Dat Ngo, Hieu Tang, Son Phan, and Alexander Schindler, “Wider or deeper neural network architecture for acoustic scene classification with mismatched recording devices,” in Proceedings of the 4th ACM International Conference on Multimedia in Asia, 2022.
[11] Lam Pham, Alexander Schindler, Mina Schütz, Jasmin Lampert, Sven Schlarb, and Ross King, “Deep learning frameworks applied for audio-visual scene classification,” arXiv preprint arXiv:2106.06840, 2021.
[12] Lam Pham, Alexander Schindler, Hieu Tang, and Truong Hoang, “Dcase 2021 task 1a: Technique report,” Tech. Rep., DCASE2021 Challenge, 2021.
[13] Sai Phaye, Emmanuel Benetos, and Ye Wang, “SubSpectralNet using sub-spectrogram based convolutional neural networks for acoustic scene classification,” in Proc. ICASSP, 2019, pp. 825–829.
[14] Zhao Ren, Kun Qian, Yebin Wang, Zixing Zhang, Vedhas Pandit, Alice Baird, and Bjorn Schuller, “Deep scalogram representations for acoustic scene classification,” IEEEC/CAA Journal of Automatica Sinica, vol. 5, no. 3, pp. 662–669, 2018.