A Multi-Task Learning Framework for Sound Event Detection using High-level Acoustic Characteristics of Sounds

Tanmay Khandelwal and Rohan Kumar Das

Fortemedia Singapore, Singapore

f20170106p@alumni.bits-pilani.ac.in, rohankd@fortemedia.com

Abstract

Sound event detection (SED) entails identifying the type of sound and estimating its temporal boundaries from acoustic signals. These events are uniquely characterized by their spatio-temporal features, which are determined by the way they are produced. In this study, we leverage some distinctive high-level acoustic characteristics of various sound events to assist the SED model training, without requiring additional labeled data. Specifically, we use the DCASE Task 4 2022 dataset and categorize the 10 classes into four subcategories based on their high-level acoustic characteristics. We then introduce a novel multi-task learning framework that jointly trains the SED and high-level acoustic characteristics classification tasks, using shared layers and weighted loss. Our method significantly improves the performance of the SED system, achieving a 36.3% improvement in terms of the polyphonic sound event detection score compared to the baseline on the DCASE 2022 Task 4 validation set.

Index Terms: sound event detection, multi-task learning, high-level acoustic characteristics

1. Introduction

Humans have the ability to perceive the environment by detecting and segregating various sound events in the surroundings. This can be a challenging task for machines [1] and therefore, sound event detection (SED) research aims to replicate the human auditory system for robust automatic detection of sound events. The SED systems enable machines to recognize the onset and offset of sound events and classify them into different environmental contexts. In addition, SED can provide feedback to other systems, such as triggering an alarm. Thus, there has been significant research dedicated to developing novel methods for comprehending the sounds of everyday life in the context of practical applications such as smart homes [2, 3], and audio surveillance [4, 5].

For the SED models to accurately predict the temporal onset and offset, strongly labeled clips are a requirement to train the system in a supervised manner. The term "strongly labeled" refers to an audio clip in which events are annotated with their respective start and end times. Typically, audio segments outside the annotated start and end time limits are deemed non-target events and are not used as training data. The construction of this set of labeled data is associated with higher annotation costs and is also dependent on the annotator's subjective judgment. Thus, recent advancements have centered on reducing reliance on strongly labeled sets while achieving an effective system. These vary from simple data augmentation methods to semi-supervised learning methods [6, 7].

Data augmentation techniques, such as filter augmentation [8] and SpecAugment [9], have been successful in improving the generalization capability of SED models. Filter augmentation weights frequency bands to extract useful information from a wider frequency range, while SpecAugment masks blocks of consecutive frequency channels and time frames in the spectrogram. In addition, a number of studies have demonstrated the efficacy of semi-supervised methods such as interpolation consistency training [10], shift consistency training [11], and extensive self-training for generating pseudo-labels [12–14] to make use of weakly labeled as well as unlabeled sets and thereby reduce the need for strongly labeled sets.

Humans possess another ability to learn multiple tasks simultaneously and transfer knowledge gained from one task to another. Taking inspiration from this innate capability, multi-task learning (MTL) aims to jointly learn multiple related tasks, allowing the knowledge acquired in one task to benefit other tasks. In MTL, all tasks are treated equally, and the primary objective is to enhance the performance of each task. This approach promotes knowledge sharing among tasks, enhances the effectiveness of each task, and mitigates overfitting risk. As a result, MTL has emerged as a powerful tool for addressing the issue of data scarcity by leveraging existing knowledge and reducing the cost of manual labeling for learning tasks.

The characteristics of various sound occurrences that appear in nature are distinct, which can be used to aid the SED system in enhancing its generalization performance. For instance, sounds generated by mechanical or electrical devices tend to have a consistent mean and variance throughout the recording, unlike impulsive noises such as a dog barking. Previous studies, such as [15], have used event duration information to tailor the length of the window size of the median filters to each sound event. Furthermore, [16–19] have proposed analyzing methods of sound events and scenes jointly based on MTL for SED and acoustic scene classification.

In this study, we propose a novel MTL framework that employs the high-level acoustic characteristics of sound events to group them under a few larger classes and use their classification as an additional task to support the primary SED system. The generated categories of classes serve as higher-level abstractions of the original sound classes, making them highly interrelated. To the best of our knowledge, this is the first study to incorporate high-level acoustic characteristics of sound events in an MTL framework, enabling the SED model to apply shared information acquired from an additional classification task. Furthermore, this MTL-based approach can be easily integrated into existing SED frameworks, and the additional branch for high-level acoustic characteristic classification (ACC) can be removed during inference, which will maintain the same number of model parameters during inference. We extend our work in [20, 21] by incorporating the proposed MTL method into the single SED branch to show its effectiveness.
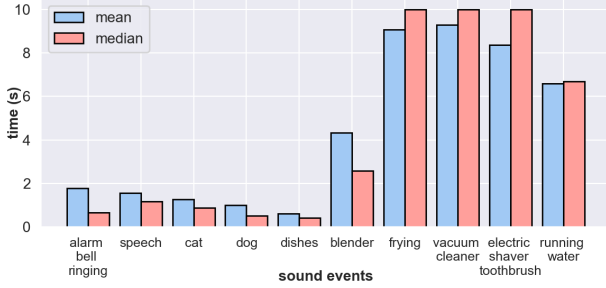
Figure 1: *Bar plot showing mean and median duration for 10-classes in DCASE 2022 Task 4 strongly labeled training set.*

2. Proposed MTL framework

2.1. High-level acoustic characteristics

The world around us is filled with a vast array of sound events that vary in frequency content, duration, and profile. This rich source of acoustic information can be leveraged to train effective SED models. Using this simple idea to aid the existing SED models, we categorize the 10-classes in the DCASE 2022 Task 4 dataset into different categories based on the similarities in their high-level acoustic characteristics in this work. To use this additional information in the existing SED system, we further propose MTL-based joint learning. This allows the SED model to exploit the high-level acoustic characteristics of similar sound signals by adding another classification task to the single SED branch. The use of the MTL framework with joint training has the added benefit that once the MTL-based model is trained, the high-level ACC branch can be removed from the model architecture. As only the 10-class SED branch is used during the inference, the number of parameters remains the same as that of a single SED branch.

In this work, we concentrate on the 10-classes present in the DCASE 2022 Task 4 dataset. Through analysis of the strongly labeled set, we present the mean and median duration for each class in Figure 1. Based on the average duration of the sound events, the plot indicates that these 10-classes can be grouped into two main categories. The 5-sound events on the right side of the bar plot have an average duration greater than 4s and can be referred to as "long duration events", whereas those on the left have an average duration less than 2s and can be referred to as "short duration events". Furthermore, we observe that the median duration for all "short duration events" is less than the mean, showing a positively skewed distribution, indicating that there are more short duration occurrences and fewer long duration occurrences. However, with the exception of "blender", the median duration for the majority of "long duration events" is higher than the mean, showing a negatively skewed distribution, which suggests that there are more long duration occurrences and fewer short duration occurrences.

We further have a closer look at each of the two groups based on the duration statistics discussed above. On analyzing the "long duration events", we observe from Figure 2 (a) that the sound events such as "vacuum cleaner" are nearly stationary over time, exhibiting similar time-frequency trajectories. On the other hand, "short duration events" such as "dishes" fluctuate along the time axis, resulting in much variations along time-frequency trajectories.

In general, the running motor of a "blender" and "vacuum cleaner" dominates over the sound of minor non-stationary noises such as the blender cutting a hard chunk or the vacuum cleaner's head hitting or rolling over other objects. Sim-
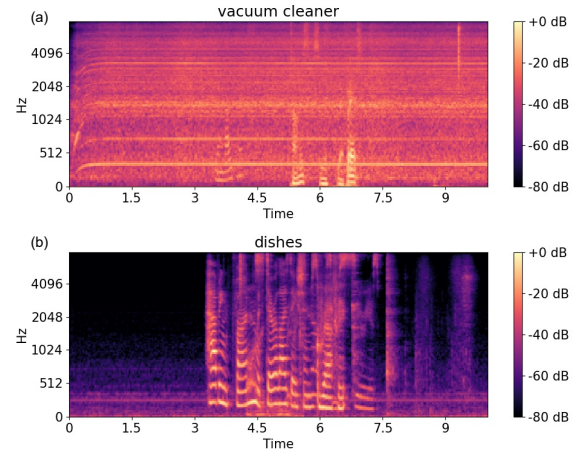

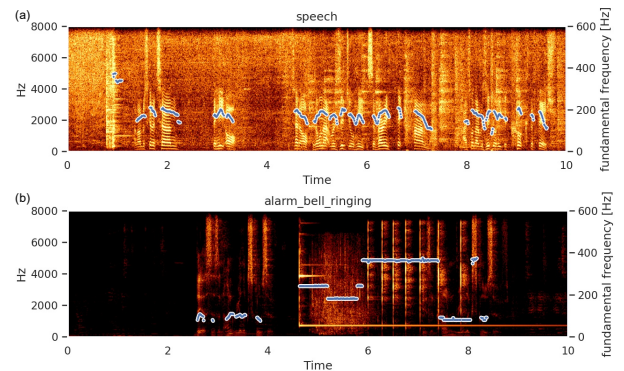Figure 2: *Log-mel spectrogram: (a) vacuum cleaner (b) dishes.*

Figure 3: *Pitch contours on a log-mel spectrogram: (a) speech (b) alarm bell ringing.*

ilarly, the sound of frying is classified as random noise because evaporation happens randomly and continuously, just as raindrops land randomly on the ground. As a result, all of these events have a relatively stable as well as less varying spectral content over time and can be approximated as stationary for a short period of time. Such events are classified as quasi-stationary [8, 22] and we categorize them into the acoustic characteristics-based class "A". Similarly, sound events like "running water" and "electric shaver/toothbrush" can be considered to be quasi-stationary. However, in these instances, the sound produced by the "electric motor" and "running water" is not loud enough to drown out other impulsive sounds. These occurrences are therefore categorized as "semi-quasi-stationary", as they share some characteristics with both stationary and non-stationary sounds, and we assigned them to class "B".

Along a similar direction, we investigate the "short duration events" and find that events like "dog", "cat", and "speech", are all examples of vocalizations produced by living organisms. Thus, these events feature continuous changes in pitch due to the variations of their sound spectra, characterized by abrupt sounds like stops and momentary turbulent sounds like fricatives, as shown in the pitch contour plotted in Figure 3 (a). Using this, we assign these events to class "C". On the other hand, classes such as "dishes" and "alarm bell ringing" are characterized by brief and sudden sounds that exhibit a more consistent and stable pitch with less variability over time, as demonstrated in Figure 3 (b). Additionally, compared to class "C", the pitch of these sounds lies in a higher frequency region. Thus, both of these events are then assigned to class "D". A summary of the

Table 1: *Categorization of the 10 sound events in DCASE Task 4 dataset into 4 classes based on the acoustic characteristics.*

Long duration events	Class	Short duration events	Class
Vacuum cleaner	A	Speech	C
Frying	A	Dog	C
Blender	A	Cat	C
Electric shaver/toothbrush	B	Dishes	D
Running water	B	Alarm bell ringing	D

sub-categorization of the 10 sound events of DCASE 2022 Task 4 based on the high-level acoustic characteristics discussed here is depicted in Table 1.

2.2. Proposed method

In our proposed MTL framework, we use the high-level acoustic characteristics-based class categorization discussed in the previous section to add another task to the existing single SED framework. In particular, we assign the provided clip section to one of the high-level acoustic characteristics-based classes ("A", "B", "C" or "D") and one of the 10 sound events. In order for the network to acquire low-level features that could enhance the performance of the SED model, these two tasks are built to share some common layers. These shared layers facilitate the extraction of features common to both SED and high-level ACC tasks. It has also been shown that when tasks vary in difficulty, the harder task's performance is improved by using the knowledge from the easier task [23]. Here, we consider SED to be the more difficult task and high-level ACC to be the simpler task, and thus we anticipate that the ACC will aid in the SED performance. Therefore, to conduct the joint training of these two tasks, we use a combined loss function L_{MTL}, which is the weighted loss function from the SED branch and the ACC branch. It can be expressed mathematically as

$$L_{MTL} = \alpha \times L_{SED} + (1 - \alpha) \times L_{ACC} \qquad (1)$$

where α is the trade-off factor that regulates the weighted loss. L_{SED} and L_{ACC} correspond to the 10-class SED loss and the 4-class high-level ACC loss, respectively. Setting $\alpha = 1$ is equivalent to a single SED branch. Furthermore, the ACC branch is removed during inference time. As a result, the number of parameters is the same as in the single SED branch.

2.3. Proposed architecture

We extend the work carried out by us in [20, 21] to transform it into the MTL framework mentioned in the previous subsection. Figure 4 illustrates its two branches, one of which corresponds to the 10-class SED task and the other to the 4-class high-level ACC task. Each of these branches employs the same frequency dynamic (FDY)-convolutional recurrent neural network (CRNN) [8] architecture on both sides after sharing a single convolutional block, which is referred to as "shared layer", as shown in Figure 4.

3. Experimental setup

3.1. Dataset and feature extraction

We use 10-second audio clips from the DCASE 2022 Task 4 dataset taken from AudioSet [24] or synthesized with isolated sound events [25] and backgrounds [26, 27] to model a domestic environment. The development set for this dataset is summarized in Table 2.

We label each entire clip or event segment in the training set with the additional high-level acoustic characteristics-based
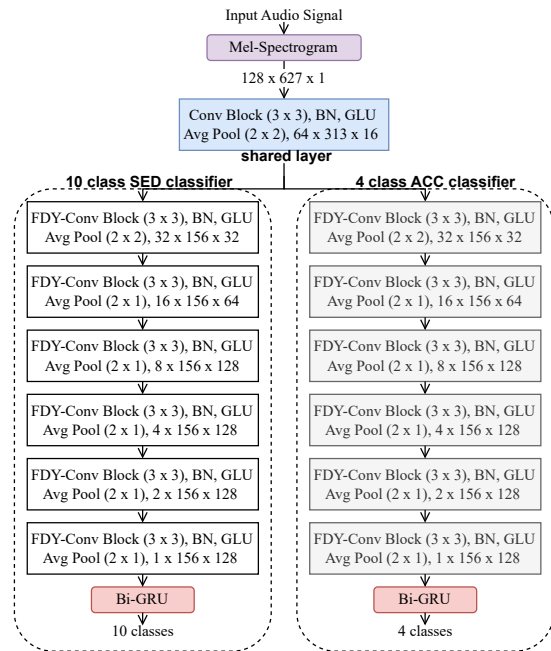

Figure 4: *Proposed MTL framework, with separate branches for 10-class SED task and 4-class high-level ACC task.*

Table 2: *Summary of DCASE 2022 Task 4 development set.*

Set	Clips	Description
Strong	10,000	Synthetic clips
	3,470	Real recordings
Weak	1,578	Real recordings
Unlabeled	14,412	Real recordings in domain
Validation	1,168	Real strongly labeled clips
	2,500	Synthetic strongly labeled clips

classes ("A", "B", "C" or "D"). Then, all these audio clips are re-sampled at 16 kHz to a mono channel using Librosa. From the re-sampled clips, log-mel filter banks are extracted using a window size of 2048 samples with a hop length of 256 samples. In addition, to make feature sequences the same length, we zero-pad shorter sequences and truncate longer sequences. Then the mel-spectrogram of the clip is normalized to zero mean and unit variance is computed over the training set.

3.2. Two-stage system (TSS) for SED

We used our two-stage method designed in [20, 21] for DCASE 2022 Task 4. This system has two stages: Stage-1 concentrates on audio-tagging (AT), and Stage-2 improves SED performance by using the reliable pseudo-labels generated by Stage-1. In Stage-1, the CNN-14-based pre-trained audio neural network [28] is utilized as the feature extractor, whereas in Stage-2, FDY-convolutions are utilized. Both of these feature extractors are followed by 2 layers of bidirectional gated recurrent units [29] with 1024 and 128 hidden units, respectively. The AT system (Stage-1) is used to make predictions on the unlabeled set and generate a pseudo-weak set. During the training of Stage-2, the weak and pseudo-weak sets are combined, and the strongly labeled and combined weakly (weak and pseudo-weak sets) labeled samples are given a weight of 1. In our study, we extend Stage-2 employing FDY-CRNN to construct the MTL

Table 3: *Performance comparison showing the importance of the proposed method on the DCASE 2022 Task 4 validation set.*

System	PSDS1	PSDS2	PSDS1+PSDS2
Baseline: CRNN [31]	0.351	0.552	0.903
Two-stage system (TSS) [20, 21]	0.472	0.721	1.193
TSS + MTL(α=0.5)	0.476	0.751	1.227
TSS + MTL(α=0.6)	0.457	0.740	1.197
TSS + MTL(α=0.7)	0.479	0.738	1.217
TSS + MTL(α=0.8)	0.480	0.751	**1.231**
TSS + MTL(α=0.9)	0.490	0.729	1.219
TSS + R-MTL(α=0.8)	0.461	0.713	1.174

framework described in Section 2.3, and assigning the high-level acoustic characteristics-based classes to each segment in the strongly labeled and combined weakly labeled set.

3.3. Training and post-processing

With a similar training setting as in [20, 21], we adopt a batch size of 48 for all experiments. Similarly, the model is trained using the Adam optimizer, with a learning rate ramp-up occurring over the first 50 optimization epochs and a maximal learning rate of 0.001. Stage-1 is trained using a total of 100 epochs, and Stage-2 is trained using a total of 200 epochs. Additionally, we also use the same data-augmentation techniques outlined in [20, 21] at each stage to enhance the model's generalizability. Since each event class has a different duration length, we also adopt the class-wise median filter [15] and search for the optimal filter length. The system was developed using Py-Torch Lightning and trained using an NVIDIA Quadro RTX 5000 GPU.

3.4. Evaluation metric

We used polyphonic sound event detection scores (PSDS) [30] as a performance metric to evaluate the systems because it can surmount the limitations of conventional event F-scores based on collars. This is computed based on two distinct scenarios, with Scenario-1 denoted by PSDS1 concentrating on temporal localization and Scenario-2 denoted by PSDS2 focusing on class confusion prevention rather than reaction time. By using PSDS as our performance metric, we aim to provide a more nuanced evaluation of our SED models that accounts for both temporal localization and class confusion prevention.

4. Results and analysis

This section presents the results and analysis of our various studies. Our experimental evaluation begins with an assessment of the SED system's effectiveness using the MTL framework. We use the two-stage system (TSS) defined in Section 3.2 and extend the single SED branch in Stage-2 to construct the proposed MTL framework. Since the primary task is the 10-class SED task, we assigned more weight to L_{SED} by varying the trade-off factor, α, from 0.5 to 0.9. The PSDS1 and PSDS2 obtained for each setup are shown in Table 3. It can be observed that the baseline adopted from DCASE 2022 Task 4 [31] and the TSS [20, 21] developed by us achieves a total PSDS (PSDS1 + PSDS2) of 0.903 and 1.193, respectively. We find an improvement in the performance of the TSS by incorporating our proposed MTL framework, with the best performance of total PSDS of 1.231 when $\alpha = 0.8$. Furthermore, with this method of joint training, the total parameters during the time of inference remain the same at 2.8M, as for the single SED branch.

Table 4: *Categorization of the 10 sound events in DCASE Task 4 dataset into 4 classes to keep similar events in separate high-level acoustic characteristics classes.*

Sound events	Class	Sound events	Class
Alarm bell ringing	A	Dishes	C
Blender	A	Frying	C
Electric shaver/toothbrush	A	Speech	C
Vacuum cleaner	B	Running water	D
Dog	B	Cat	D

Table 5: *Comparison with top-ranked single systems (without ensemble) from DCASE Task 4 2022 on the validation set.*

System	PSDS1	PSDS2	PSDS1+PSDS2	#Parameters
Ebbers-UPB-task4 [32]	0.505	0.807	1.312	15.4M
TSS + MTL (α=0.8)	**0.483**	**0.728**	**1.231**	**2.8M**
Zhang-UCAS-task4 [33]	0.459	0.672	1.131	11M
Kim-GIST-task4 [34]	0.455	0.670	1.125	1M
Dinkel-XiaoRice-task4 [35]	0.425	0.644	1.069	37M

To test whether the proposed method of classifying classes based on high-level acoustic characteristics works, we run another experiment in which we classify highly different acoustic events into the same high-level classes and term it randomized-MTL (R-MTL). Table 4 summarizes our categorization to retain similar acoustic classes into different high-level classes. According to the findings in Table 3, the total PSDS for TSS+R-MTL at the same $\alpha = 0.8$ value as the best-performing system decreases significantly from 1.231 to 1.174. As a result, we can conclude that the method of categorizing sound events has a substantial impact on performance, and our method described in Section 2.1 aids in the SED performance using MTL.

We then demonstrate the effectiveness of our proposed approach on the development set by comparing it to the top teams from DCASE 2022 Task 4 with single systems (no ensembling). The reported performances of other systems are taken directly from their respective technical reports. Table 5 shows the comparison of our proposed system to those top-performing team's single systems in a sorted order based on total PSDS value. We observe that our proposed MTL framework outperforms most of the systems and comes really close to the best single system with a larger model size of 15.4M parameters and multiple stages of self-training. It is also worth noting that our proposed framework can be easily integrated into other SED systems with just an additional branch for the high-level ACC task, without increasing the number of parameters during inference.

5. Conclusion

In this paper, we propose a novel MTL framework that leverages high-level acoustic characteristics to enhance the SED performance, without usage of additional annotated data. Our proposed framework performs joint learning with an additional high-level ACC task branch using a weighted loss, with both tasks sharing common layers to leverage cross-task information by sharing parameters (weights). The studies revealed that the MTL framework with additional high-level ACC task improves the primary SED system's performance. We also demonstrate the effectiveness of our proposed sound event categorization by separating similar sound events into distinct high-level acoustic characteristics classes and comparing our system's performance with the top teams in DCASE 2022 Task 4 on the validation set. In the future, we intend to incorporate an adaptive trade-off factor to control the weighted loss.

6. References

[1] T. Virtanen, M. Plumbley, and D. Ellis, "Computational analysis of sound scenes and events," *Springer International Publishing*, pp. 1–422, 2017.

[2] J. P. Bello, C. Mydlarz, and J. Salamon, "Sound analysis in smart cities," *Springer International Publishing*, pp. 373–397, 2018.

[3] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, "Monitoring activities of daily living in smart homes: Understanding human behavior," *IEEE Signal Processing Magazine*, pp. 81–94, 2016.

[4] A. Harma, M. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," *IEEE International Conference on Multimedia and Expo*, pp. 634–637, 2005.

[5] T. Khandelwal, R. K. Das, and E. S. Chng, "Is your baby fine at home? Baby cry sound detection in domestic environments," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 275–280, 2022.

[6] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *International Conference on Neural Information Processing Systems*, pp. 1195–1204, 2017.

[7] H. Dinkel, X. Cai, and Z. Yan, "The smallrice submission to the DCASE 2021 Task 4 challenge: A lightweight approach for semi-supervised sound event detection with unsupervised data augmentation," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2021.

[8] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, "Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection," *Interspeech*, pp. 2763–2767, 2022.

[9] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *Interspeech*, pp. 2613–2617, 2019.

[10] X. Zheng, H. Chen, and Y. Song, "Zheng USTC team's submission for DCASE 2021 Task 4 semi-supervised sound event detection," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2021.

[11] J. Ebbers and R. Haeb-Umbach, "Self-trained audio tagging and sound event detection in domestic environments," *Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, pp. 226–230, 2021.

[12] C.-Y. Koh, Y.-S. Chen, Y.-W. Liu, and M. R. Bai, "Sound event detection by consistency training and pseudo-labeling with feature-pyramid convolutional recurrent neural networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 376–380, 2021.

[13] C. Park, D. Kim, and H. Ko, "Sound event detection by pseudo-labeling in weakly labeled dataset," *Sensors*, vol. 21, 2021.

[14] N. K. Kim and H. K. Kim, "Self-training with noisy student model and semi-supervised loss function for DCASE 2021 challenge Task 4," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2021.

[15] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Convolution-augmented transformer for semi-supervised sound event detection," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2020.

[16] K. Imoto, N. Tonami, Y. Koizumi, M. Yasuda, R. Yamanishi, and Y. Yamashita, "Sound event detection by multitask learning of sound events and scenes with soft scene labels," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 621–625, 2020.

[17] H. L. Bear, I. Nolasco, and E. Benetos, "Towards joint sound scene and polyphonic sound event recognition," *Interspeech*, pp. 4594–4598, 2019.

[18] N. Tonami, K. Imoto, M. Niitsuma, R. Yamanishi, and Y. Yamashita, "Joint analysis of acoustic events and scenes based on multitask learning," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 338–342, 2019.

[19] R. Mars and R. K. Das, "A device classification-aided multi-task framework for low-complexity acoustic scene classification," *Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2022.

[20] T. Khandelwal, R. K. Das, A. Koh, and E. S. Chng, "FMSG-NTU submission for DCASE 2022 Task 4 on sound event detection in domestic environments," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2022.

[21] ——, "Leveraging audio-tagging assisted sound event detection using weakified strong labels and frequency dynamic convolutions," *IEEE Statistical Signal Processing Workshop*, 2023.

[22] J. S. Bendat and A. G. Piersol, "Random data: Analysis and measurement procedures," *Wiley*, pp. 8–12, 2000.

[23] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, pp. 41–75, 1997.

[24] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "AudioSet: An ontology and human-labeled dataset for audio events," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017.

[25] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[26] G. Dekkers, S. Lauwereins, B. Thoen, M. Adhana, H. Brouckxon, B. V. den Bergh, T. van Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," *Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, pp. 1–5, 2017.

[27] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," *European Signal Processing Conference (EUSIPCO)*, pp. 1128–1132, 2016.

[28] 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.

[29] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014.

[30] C. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulovic, "A framework for the robust evaluation of sound event detection," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 61–65, 2020.

[31] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for DCASE 2019 Task 4 technical report," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2019.

[32] J. Ebbers and R. Haeb-Umbach, "Pre-training and self-training for sound event detection in domestic environments," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2022.

[33] S. Xiao, "Pretrained models in sound event detection for DCASE 2022 challenge Task 4," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2022.

[34] J. W. Kim, G. W. Lee, H. K. Kim, Y. S. Seo, and I. H. Song, "Semi-supervised learning-based sound event detection using frequency-channel-wise selective kernel for DCASE challenge 2022 Task 4," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2022.

[35] H. Dinkel, Z. Yan, Y. Wang, M. Song, J. Zhang, and W. Wang, "A large multi-modal ensemble for sound event detection," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2022.