APSIPA Transactions on Signal and Information Processing, 2025, 14, e31

Original Paper

Joint Analysis of Acoustic Scenes and Sound Events Based on Semi-supervised Training of Sound Events With Partial Labels

Keisuke Imoto*

Kyoto University, Japan; keisuke.imoto@ieee.org

ABSTRACT

Annotating time boundaries of sound events is labor-intensive, limiting the scalability of strongly supervised learning in audio detection. To reduce annotation costs, weakly-supervised learning with only clip-level labels has been widely adopted. As an alternative, partial label learning offers a cost-effective approach, where a set of possible labels is provided instead of exact weak annotations. However, partial label learning for audio analysis remains largely unexplored. Motivated by the observation that acoustic scenes provide contextual information for constructing a set of possible sound events, we utilize acoustic scene information to construct partial labels of sound events. On the basis of this idea, in this paper, we propose a multitask learning framework that jointly performs acoustic scene classification and sound event detection with partial labels of sound events. While reducing annotation costs, weakly-supervised and partial label learning often suffer from decreased detection performance due to lacking the precise event set and their temporal annotations. To better balance between annotation cost and detection performance, we also explore a semi-supervised framework that leverages both strong and partial labels.

*This work was supported by JSPS KAKENHI Grant Numbers 22H03639, 23K16908, and 25H01142.

Received 26 August 2025; revised 09 October 2025; accepted 28 October 2025
ISSN 2048-7703; DOI 10.1561/116.20250080

Moreover, to refine partial labels and achieve better model training, we propose a label refinement method based on self-distillation for the proposed approach with partial labels.

Keywords: Acoustic scene classification, partial label, sound event detection

1 Introduction

Computational analysis of environmental sounds has recently attracted much attention in the field of acoustic signal and speech processing. Environmental sound analysis, which is not limited to speech or musical sound analysis, greatly expands the range of sound-based applications, such as media retrieval, hearing aids, machine condition monitoring, self-driving cars, robot auditions, and biomonitoring systems [7, 23, 22].

In environmental sound analysis, acoustic scene classification (ASC) and sound event detection (SED) are fundamental tasks. Of these tasks, ASC estimates an acoustic scene label most related to an input sound. SED predicts all sound event labels and their corresponding start and end times in an input sound. Recently, various ASC and SED methods based on neural networks have been adopted, and they effected a remarkable improvement in performance. For example, Valenti *et al.* [27] and Ford *et al.* [8] have proposed ASC systems using the convolutional neural network (CNN) and ResNet, respectively. Kong *et al.* [15] proposed an ASC method using a pre-trained model with a large-scale audio dataset. Çakir *et al.* [4] introduced a SED technique incorporating a convolutional recurrent neural network (CRNN). More recently, Kong *et al.* [14] and Miyazaki *et al.* [21] proposed Transformer- and Conformer-based SED methods, respectively, which have been widely employed in many studies.

Acoustic scenes and sound events are mutually related and they are effectively estimated by utilizing mutual information. For instance, in the acoustic scene *home*, the sound events *cutlery* and *door opening/closing* tend to occur, whereas the sound events *car* and *bird singing* are not likely to occur. Taking into account the relationship between acoustic scenes and sound events, Mesaros *et al.* [19] proposed a SED method leveraging knowledge on acoustic scenes. Similarly, Imoto and Ono [12] and Hou *et al.* [9] proposed ASC methods that take into account the association between acoustic scenes and sound events, through the use of Bayesian generative models and graph neural network, respectively. In more recent works, Bear *et al.* [2], Tonami *et al.* [24], and Jung *et al.* [13] proposed the joint analysis of acoustic scenes and sound events using the multitask learning (MTL) framework of ASC and SED, which learns both tasks simultaneously.

Many methods for environmental sound analysis are based on the supervised learning scheme, which train model parameters using large-scale strongly annotated data. However, annotating labels for environmental sounds, especially annotating time boundaries of sound events, is very laborious. Moreover, there are potential applications where collecting large-scale annotated data itself is difficult. For example, in-home monitoring systems must address privacy concerns, making it difficult to share audio data with unspecified annotators. Similarly, in ecological monitoring, expert knowledge is required to annotate species-specific sounds such as bird calls or amphibian vocalizations, limiting the scalability of manual annotation. To mitigate this challenge, in the context of single-task SED, many methods using weakly-supervised learning have been proposed [16, 26]. In the paradigm of weakly-supervised learning for SED, only information on clip-level activations of sound events is provided in the training stage, whereas sound event labels and their time boundaries are estimated in the inference stage. For the joint analysis of acoustic scenes and sound events, Tsubaki *et al.* [25] and Igarashi *et al.* [10] proposed a method that applies the weakly-supervised SED approach to the MTL framework.

To further reduce the cost of annotation, partial label learning, in which a detection or classification model is trained using a set of possible labels, has also been proposed in image analysis [5]. However, partial label learning for SED has been largely unexplored. To clarify the differences among strong, weak, and partial labels in SED, Figure 1 illustrates each labeling scheme. The partial labels actually are a form of weak labels; however, unlike conventional weak labels that only include true sound event classes, partial labels represent a set of possible sound event classes, which may contain additional labels beyond the ground truth. Although annotating precise weak labels still requires substantial effort, annotating only a set of possible labels can significantly reduce annotation costs and facilitate the creation of large-scale training data. Note that the conventional method of partial label learning [5] has addressed the image classification task, where exactly one true label is assumed to exist in each possible label set. In contrast, our work focuses on SED, where multiple true sound event labels may be present within the partial labels.

In environmental sound analysis, the strong correlation between acoustic scenes and sound events enables the generation of effective partial labels of sound events, as scene information can be used to constrain the candidate set of sound event classes. Since annotating acoustic scene labels requires much less effort than annotating precise weak labels of sound events, leveraging scene information offers a cost-effective way to guide SED model training. Thus, in this work, we propose the MTL framework of SED and ASC using partial labels of sound events, which offers a suitable and practical setting for exploring the use of partial label learning in environmental sound analysis. On

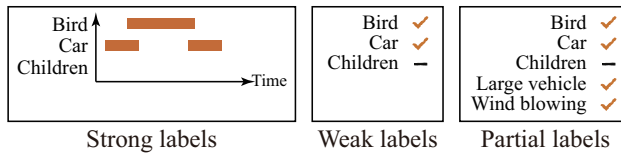

Figure 1: Illustration comparing strong, weak, and partial labels in sound event. Strong labels provide sound event classes and their time stamps, weak labels indicate which event classes occur within an audio clip, and partial labels provide a candidate set of event labels.

the other hand, introducing partial label learning may result in performance degradation compared with methods using strong labels. Thus, we further explore an MTL-based joint analysis of acoustic scenes and sound events using both strong and partial labels of sound events simultaneously, that is, a semi-supervised approach. We then evaluate the performances of ASC and SED in detail and characterize the behavior of the proposed method with partial labels.

The subsequent sections of this paper are as follows. In Section 2, we discuss conventional methodologies for ASC, SED, and the joint analysis of acoustic scenes and sound events by leveraging multitask learning. Section 3 is dedicated to introducing our methods of joint analysis of acoustic scenes and sound events utilizing semi-supervised approaches with partial labels of sound events. The evaluation experiments to validate the detailed performance of scene classification and event detection are presented in Section 4. Finally, in Section 5, we conclude this paper and discuss potential directions for a future work.

2 Conventional Methods

2.1 Acoustic Scene Classification and Event Detection

In this section, we overview basic implementations for ASC and SED using neural networks. Many systems for ASC and SED first extract a time–frequency representation of the acoustic signal $\mathbf{X} \in \mathcal{R}^{D \times T}$ from an audio input. Here, D and T represent the numbers of frequency bins and time frames, respectively. The log mel-band spectrogram or a time series of mel frequency cepstrum coefficients (MFCCs) is typically used for the acoustic feature. The extracted acoustic feature is subsequently fed to the ASC or SED networks, which calculate logits $\mathbf{y}$ for classifying acoustic scenes or detecting sound events, respectively.

As for the ASC network, the model parameters are trained using the logits and the cross-entropy (CE) loss function $\mathcal{L}_{\mathsf{scene}}$ as

$$\mathcal{L}_{\mathsf{scene}} = -\sum_{n=1}^{N} \Big\{ z_n \log\big(\sigma\left(y_n\right)\big) \Big\}, \tag{1}$$

where N, $z_n \in \{0, 1\}$, and σ are the number of acoustic scene classes, the acoustic scene label, and softmax function, respectively.

On the other hand, the parameters of the SED network are tuned using the following binary cross-entropy (BCE) loss function as follows:

$$\begin{aligned}\mathcal{L}_{\mathsf{event}} &= -\sum_{t=1}^{T} \Big\{ \mathbf{z}_t \log(\mathbf{y}_t) + (1-\mathbf{z}_t) \log\big(1-s(\mathbf{y}_t)\big) \Big\} \\ &= -\sum_{t,m=1}^{T,M} \Big\{ z_{t,m} \log s(y_{t,m}) + (1-z_{t,m}) \log\big(1-s(y_{t,m})\big) \Big\},\end{aligned} \tag{2}$$

where T, M, $z_{t,m}$, and s indicate the number of time frames in a sound clip, the number of sound event classes, the target event label in the time frame t for the sound event m, and the sigmoid function, respectively.

2.2 Joint Analysis of Acoustic Scenes and Sound Events Based on Multitask Learning

In the realm of environmental sound analysis, numerous methods address scene classification and event detection as individual tasks. Only a few works focus on the idea that information on acoustic scenes and sound events mutually enhances the performance in ASC and SED, and methods that jointly analyze acoustic scenes and sound events, has been proposed [2, 24, 13].

A typical implementation of the joint analysis of acoustic scenes and sound events utilizes an MTL-based neural network, which shares part of the network and information on acoustic scenes and sound events, as shown in Figure 2. The conventional method first extracts a feature embedding common to acoustic scenes and sound events in the shared layers. The resultant feature embedding is subsequently fed to the dedicated layers tailored for ASC and SED. For the dedicated layers for acoustic scenes and sound events, CNN, the recurrent neural network (RNN), and the Transformer encoder are often employed.

To train the model parameters, the conventional methods [24] adopt a loss function represented by the linear combination of Equations (1) and (2) with acoustic scene labels and strong sound event labels.

$$\mathcal{L} = \alpha\mathcal{L}_{\mathsf{scene}} + \beta\mathcal{L}_{\mathsf{event}} \tag{3}$$

Here, α and β are the constant weights for ASC and SED losses, respectively. In this paper, we set $\beta = 1.0$ without loss of generality.

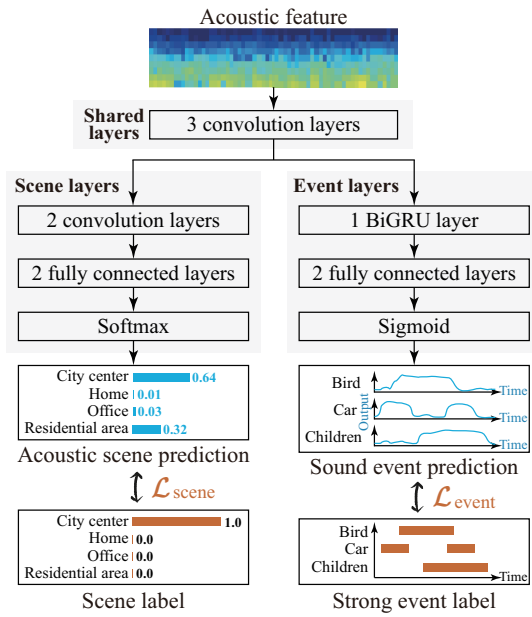

Figure 2: Network structure of conventional MTL-based method [24].

2.3 Weakly-supervised Method for Joint Analysis of Acoustic Scenes and Sound Events

Annotating time boundaries of sound events is labor-intensive and time-consuming. To overcome the challenge of annotating strong labels of sound events, in the single-task SED scenarios, many methods apply a weakly-supervised scheme using weak labels of sound events. Here, weak labels of sound events only have information on the presence or absence of sound events in a sound clip. Many weakly-supervised methods for single-task SED employ the multiple-instance learning (MIL) framework [6, 28], which makes a final decision by aggregating small bag-level decisions. In the case of SED, the system outputs a clip-level detection result of a sound event by aggregating frame-level decisions.

Tsubaki *et al.* [25] proposed a framework for the joint analysis of ASC and SED, in which weakly-supervised learning is integrated into the SED task. Figure 3 shows the network structure of the conventional method [25], which has two branches in the event layers. One branch has the pooling layer corresponding to the MIL framework and enables the weakly-supervised training in SED. The other branch only has the sigmoid function to hold temporal information and it enables us to estimate time stamps of sound events in the inference stage.

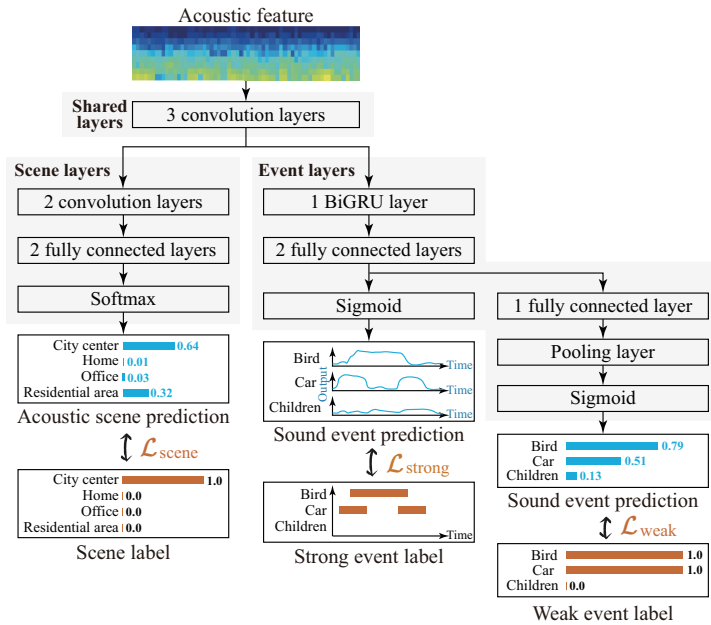

Figure 3: Network structure of MTL-based method using weak labels of sound events.

To train the model parameters of the weakly-supervised method, the linear combination of the ASC and SED losses represented by Equation (3) are also used, whereas we modify the SED loss $\mathcal{L}_{\text{event}}$ as follows:

$$\begin{aligned}\mathcal{L}_{\text{event}} &= \gamma\mathcal{L}_{\text{strong}} + \zeta\mathcal{L}_{\text{weak}} \\ &= -\gamma \sum_{m,t=1}^{M,T} \Big\{ z_{m,t} \log s(y_{m,t}) + (1 - z_{m,t}) \log \big(1 - s(y_{m,t})\big) \Big\} \\ &\quad - \zeta \sum_{m=1}^{M} \Big\{ z_m \log s(y_m) + (1 - z_m) \log \big(1 - s(y_m)\big) \Big\},\end{aligned} \tag{4}$$

where z_m and y_m are the weak label for the sound event m and the clip-level prediction of the sound event m, respectively. γ and ζ are the constant weights for the losses for the frame- and clip-level predictions, respectively. Here, in the training stage, the strong event label $z_{m,t}$ is prepared from the weak label z_m as a pseudo-sound event label as

$$\mathbf{Z}_{\mathsf{pseudo\_strong}} = \overbrace{\begin{pmatrix} z_1 & \cdots & z_1 & \cdots & z_1 \\ \vdots & \ddots & \vdots & & \vdots \\ z_m & \cdots & z_m & \cdots & z_m \\ \vdots & & \vdots & \ddots & \vdots \\ z_M & \cdots & z_M & \cdots & z_M \end{pmatrix}}^{T}. \tag{5}$$

2.4 Semi-supervised Method for Joint Analysis of Acoustic Scenes and Sound Events With Weak Labels of Sound Events

The MTL-based method using weak labels of sound events mitigates the challenge inherent in collecting strong labels of sound events to some extent. Nonetheless, joint analysis of ASC and SED using weak labels results in a lower SED performance as compared with that using strongly labeled data. To address this problem, a semi-supervised learning scheme has been proposed for SED, in the context of joint analysis of ASC and SED [10]. In general, methods that amalgamate both supervised and unsupervised training modalities are termed as semi-supervised learning. In this paper, however, we specifically refer to the method leveraging both strong and weak/partial labels for the SED model training as semi-supervised learning.

Let the acoustic feature sets with strong and weak labels be $\mathcal{X}_{\mathsf{strong}}$ and $\mathcal{X}_{\mathsf{weak}}$, respectively. Similarly, we consider that the strong and weak label sets as $\mathcal{Z}_{\mathsf{strong}}$ and $\mathcal{Z}_{\mathsf{weak}}$, respectively. For the semi-supervised approach, we construct the acoustic feature and label sets as

$$\mathcal{X}_{\mathsf{semi}} = \{\mathcal{X}_{\mathsf{strong}}, \mathcal{X}_{\mathsf{weak}}\}, \tag{6}$$
$$\mathcal{Z}_{\mathsf{semi}} = \{\mathcal{Z}_{\mathsf{strong}}, \mathcal{Z}_{\mathsf{weak}}\}. \tag{7}$$

For the semi-supervised approach, we can employ various network architectures once it is designed with four key modules: (i) an acoustic embedding extractor, (ii) acoustic scene classifier, and (iii)(iv) sound event detectors with weak labels and strong labels. In this paper, we illustrate the conventional semi-supervised method with the same network structure as that of the weakly-supervised method as shown in Figure 3.

To train the model parameters, Equations (3) and (4) are also used as the loss function, while γ and ζ are replaced with the following Kronecker delta functions:

$$\delta_\gamma = \begin{cases} 1 & \text{if } \mathbf{z} \text{ is strong label} \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

$$\delta_\zeta = \begin{cases} 1 & \text{if } \mathbf{z} \text{ is weak label} \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

This strategy enables switching between SED networks depending on whether strong labels are available or only weak labels can be used. The semi-supervised method using strong and weak labels is expected to achieve more reliable model training than the weakly-supervised method that relies on pseudo labels of sound events.

3 Joint Analysis of Acoustic Scenes and Sound Events Based on Semi-supervised Approach With Partial Labels of Sound Events

Annotating weak labels for sound events indeed alleviates the cost of labor involved in annotating strong labels for sound events. However, compared with annotating acoustic scene labels, annotating weak labels for sound events is still labor-intensive. Therefore, we propose a method that utilizes acoustic scene labels to generate candidate weak labels for sound events, which can then be employed as partial labels in model training. In particular, this paper explores the use of partial labels in semi-supervised learning for joint analysis of acoustic scenes and sound events.

To generate partial labels of sound events, we can utilize acoustic scene labels in several ways: one approach is to pre-construct candidate label lists for each acoustic scene, while an alternative is to generate these candidate lists using a pre-trained model, such as a large language model (LLM). For instance, in our experiments in this study, we created partial weak labels by inputting acoustic scene labels into ChatGPT o3-mini-high.[1,2] The prompt used for generating the partial labels is provided in Appendix, and the resulting constructed partial labels are listed in Table 1. Compared with the actual sound event label list shown in Table 2, the generated partial label set includes a significantly larger number of candidate sound events, such as generic events like *(object) impact*, which commonly appear in various acoustic scenes. On the other hand, we observed no case where actually occurring events were omitted from the generated labels. Given that the partial labels were created using the publicly available LLM, we believe that the quality of partial labels reflects a realistic application scenario, and their reliability is sufficient for practical use.

In this work, we further apply a method that refines sound event labels and generates pseudo strong labels using self-distillation, to mitigate the noise

[1]The label list was generated using ChatGPT o3-mini-high on February 02, 2025.

[2]We have also generated partial labels using the same prompts with several LLMs, including ChatGPT 5 Thinking and Gemini 2.5 Pro. These models produced sound event label sets largely similar to those obtained with ChatGPT o3-mini-high.

Table 1: Partial labels generated by ChatGPT o3-mini-high for TUT Acoustic Scenes 2016 and TUT Sound Events 2016/2017.

	(object) banging	(object) impact	(object) rustling	(object) snapping	(object) squeaking	bird singing	brakes squeaking	breathing	car	children	cupboard	cutlery	dishes	drawer	fan	glass jingling	keyboard typing	large vehicle	mouse clicking	mouse wheeling	people talking	people walking	washing dishes	water tap running	wind blowing
City center	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	-	-	-	-	-	-	-	✓	-	-	✓	✓	-	-	✓
Home	✓	✓	✓	✓	✓	✓	-	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	✓	✓
Office	✓	✓	✓	-	✓	-	-	✓	-	-	✓	✓	✓	✓	✓	-	✓	-	✓	✓	✓	✓	✓	✓	-
Residential area	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	-	-	-	-	-	-	-	✓	-	-	✓	✓	-	-	✓

Table 2: Sound event list of training dataset in TUT Acoustic Scenes 2016 and TUT Sound Events 2016/2017.

	(object) banging	(object) impact	(object) rustling	(object) snapping	(object) squeaking	bird singing	brakes squeaking	breathing	car	children	cupboard	cutlery	dishes	drawer	fan	glass jingling	keyboard typing	large vehicle	mouse clicking	mouse wheeling	people talking	people walking	washing dishes	water tap running	wind blowing
City center	-	-	-	-	-	✓	-	✓	✓	-	-	-	-	-	-	-	✓	-	-	✓	✓	-	-	-
Home	-	✓	✓	✓	-	-	-	-	-	✓	✓	✓	✓	-	✓	-	-	-	-	✓	✓	✓	-
Office	-	✓	✓	-	✓	-	-	✓	-	-	-	-	-	✓	-	✓	-	✓	✓	✓	✓	-	-	-
Residential area	✓	-	-	-	-	✓	-	-	✓	✓	-	-	-	-	-	-	-	✓	-	-	✓	✓	-	-	✓

in the partial label set generated using an LLM. This self-distillation-based approach represents one of the simplest methods for label refinement in semi-supervised learning. To verify the feasibility of model training from partial labels in the multitask learning of sound events and acoustic scenes, we employ this simple label refinement method in this study. The procedure for this partial label learning is shown in Figure 4. First, partial labels are treated as weak ground truth labels, and the joint ASC and SED model is trained using both strong and partial labels according to the method described in Section II-D. Once the model parameters have been trained, the pre-trained model is frozen and the training data with partial labels is fed into the self-distillation module to obtain logits. The posterior probabilities of the sound events are then calculated using a sigmoid function, and the distillated strong event labels are obtained by thresholding them with ϕ. After that, the main module is re-trained using the strong and distillated strong event labels with the conventional MTL-based method described in Section II-B.

4 Evaluation Experiments

4.1 Experimental Conditions

We carried out experiments to evaluate the conventional and proposed MTL-based joint analyses of acoustic scenes and sound events. For the evaluation experiments, we constructed a dataset composed of the TUT Acoustic Scene 2016/2017 and TUT Sound Events 2016/2017 [20, 18], which includes four acoustic scenes (*city center*, *home*, *office*, and *residential area*) and 25 sound events (e.g., *bird singing*, *car*, *dishes*, and *keyboard typing*). The dataset con-

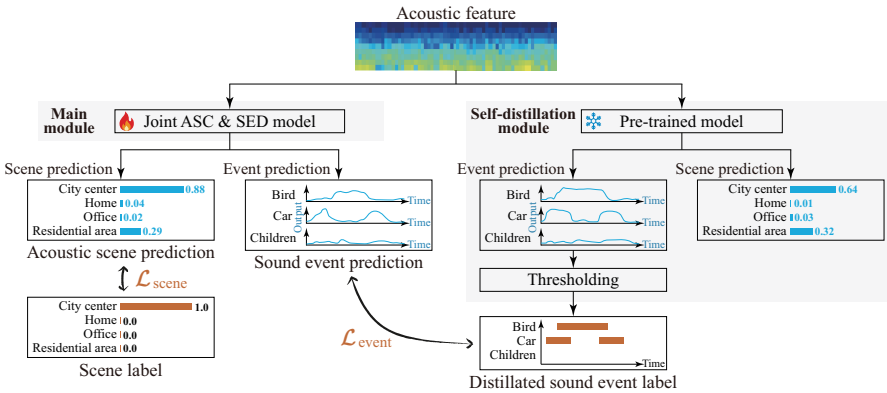

Figure 4: Self-distillation-based model training for semi-supervised method using partial labels of sound events.

tains a total of 266 min of sounds, which includes 192 min of sounds for model training and 74 min of sounds for evaluation. The partial labels were created using ChatGPT o3-mini-high, which was one of the most capable and generally applicable models available at the time of our experiments. All experiments were conducted on a single Intel Xeon Gold 6128 Processor and an NVIDIA RTX 6000 Ada Generation GPU. The details of the dataset and baseline code are available.[3,4]

We calculated the 64-dimensional log mel-band spectrogram with a frame length of 40 ms and a hop size of 20 ms. The model structure used for our experiment is shown in Figures 2, 3, and 4, and Table 3, which are based on conventional works [24]. In our preliminary experiments, we also evaluated other sophisticated model architectures for the SED-specific layers such as the Transformer and Conformer. However, these model architectures showed performance nearly equivalent to that of the CRNN-based method. This may be because we used the dataset with limited size. In this study, we thus adopt the same model architecture as in previous research to enable direct comparisons. The threshold ϕ for self-distillation was determined through the preliminary experiment using cross-validation setup on the training data as shown in Table 4. The other experimental conditions are also found in Table 4. These settings and hyperparameters were determined by referring to [24]. Since the original dataset has strong labels of sound events, we randomly selected samples from the training set and discarded time stamps to create weak labels. We conducted the evaluation experiments 10 times for each experimental condition with random initial values of model parameters.

[3] https://www.ksuke.net/dataset/.
[4] https://github.com/KeisukeImoto/mtl_sed_asc.

Table 3: Detailed structure of MTL network of ASC and SED using weak/partial labels.

Shared layers	
Log-mel energy (500 frames × 64 mel bin)	
3×3 kernel size/128 ch. Batch norm., Leaky ReLU 1×8 Max pooling	
(3×3 kernel size/128 ch. Batch norm., Leaky ReLU 1×2 Max pooling) × 2	
Scene layers	**Event layers**
3×3 kernel size/256 ch. Batch norm., Leaky ReLU 25×1 Max pooling	Transformer Enc. w/ 512 units
3×3 kernel size/256 ch. Batch norm., Leaky ReLU Global max pooling	FC w/ 48 units, Leaky ReLU
FC w/ 32 units, Leaky ReLU	FC w/ 25 units
FC w/ 4 units, Softmax	Sigmoid FC w/ 16 units Leaky ReLU Global max pooling Sigmoid

Table 4: Experimental conditions.

Acoustic feature	Log-mel energy (64 dim.)
Frame length/shift	40 ms/20 ms
Length of sound clip	10 s
Optimizer	RAdam [17]
SED detection threshold	0.5
α, β, γ, ζ	0.001, 1.0, 1.0, 0.01
ρ_{GTC}, ρ_{DTC}	0.1, 0.1
Threshold ϕ for self-distillation	0.2

4.2 Experimental Results

4.2.1 Overall Performance Characteristics of ASC and SED

Table 5 shows the overall performance of ASC and SED in terms of Fscore, especially in the segment-based and intersection-based (IS-based) metrics [3] for SED. In our experiments, we refer to the methods using strong and weak labels of sound events as strong MTL and weak MTL, respectively. The semi-supervised methods using weak and partial labels are referred to as semi-MTL w/ weak labels and semi-MTL w/ partial labels, respectively. For the semi-MTL conditions, we conducted the experiments using 30% of the strongly labeled data and 70% of the weakly/partially labeled data. We also conducted experiments under a condition where data without strong labels were excluded from training. This setting is referred to as Strong MTL w/ reduced data.

Table 5: Overall performance characteristics of ASC and SED. We conducted the experiments with 30% of the strongly labeled data and 70% of the weakl/partial labeled data under the semi-MTL condition.

Method	Scene		Event (Segment-based)		Event (IS-based)	
	Micro-Fscore	Macro-Fscore	Micro-Fscore	Macro-Fscore	Micro-Fscore	Macro-Fscore
Strong MTL	91.42% ±3.00	91.68% ±3.09	**53.91% ±0.94**	**24.09% ±0.83**	**26.22% ±1.50**	**16.81% ±1.32**
Weak MTL	90.51% ±2.97	90.57% ±3.31	22.74% ±18.99	10.60% ±7.37	8.66% ±1.26	7.18% ±1.00
Strong MTL w/ reduced data	91.78% ±2.19	91.86% ±2.38	49.01% ±2.03	15.95% ±1.77	20.90% ±2.74	10.01% ±1.88
Semi-MTL w/ weak labels	91.76% ±2.70	92.08% ±2.81	52.11% ±1.98	21.58% ±1.35	23.57% ±2.21	14.55% ±1.75
Semi-MTL w/ partial labels (proposed)	**92.12% ±2.59**	**92.58% ±2.43**	51.77% ±1.76	21.51% ±1.24	23.96% ±1.69	14.87% ±1.47

The results show that the semi-MTL-based methods achieve reasonable micro- and macro-Fscores for ASC that are similar to those of the conventional strong and weak MTL methods. In particular, the proposed semi-supervised approach using partial labels outperformed conventional MTL methods in ASC. This is because the partial labels for sound events, which were generated using acoustic scene labels from an LLM, contain information on acoustic scenes, and they may have enhanced scene classification.

For SED, the proposed semi-supervised methods with partial labels achieves the detection performance equivalent to that of the conventional semi-supervised method using weak labels in terms of both segment- and IS-based metrics. This result indicates that the proposed method can further reduce the annotation costs for sound events compared to the conventional semi-supervised method with promising SED results.

4.2.2 Performance Characteristics of ASC and SED at Various Proportion of Weak/partial Labels

To investigate the detailed behavior of strong, weak, and semi-MTL approaches, we show the evaluation performance of ASC and SED as the proportion of strongly labeled sound event data varies in Figures 5–7. Figure 5 shows that the ASC performance of the proposed semi-MTL approaches remains nearly equivalent to that of the strong MTL approach, even as the proportion of weak/partial labels increases. This result indicates that ASC does not necessarily require temporal information on sound events, but requires only clip-level information on sound events in acoustic signals.

For the SED performance, Figures 6–7 show that the F-score does not decrease considerably until the proportion of partial labels reaches around

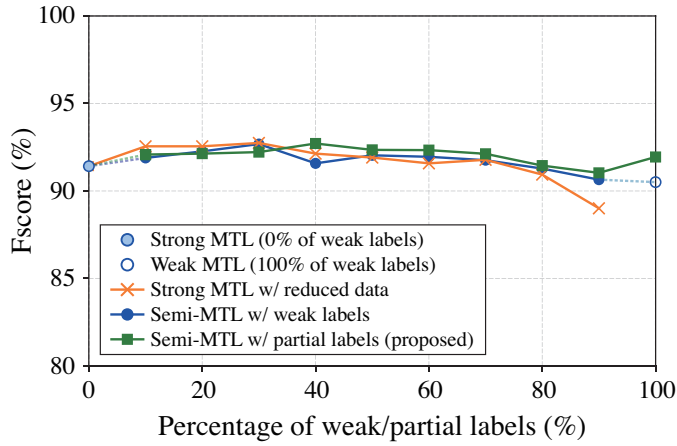

Figure 5: ASC performance for various ratios of weakly/partially labeled data of sound events in terms of micro-Fscore.

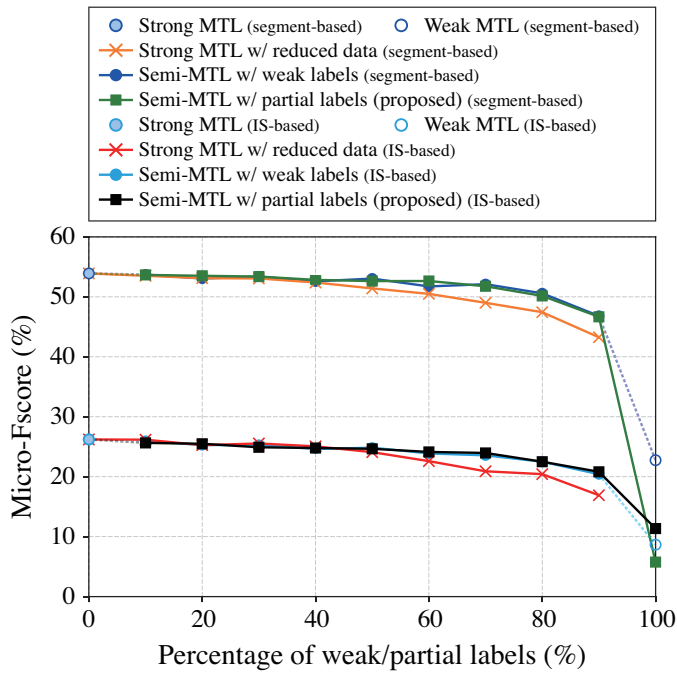

Figure 6: SED performance for various ratios of weakly/partially labeled data of sound events in terms of micro-Fscore.

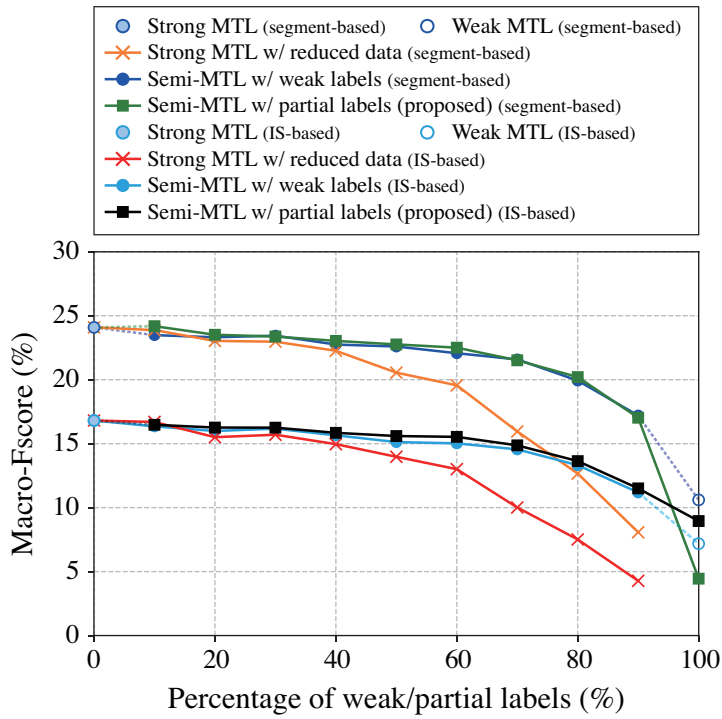

Figure 7: SED performance for various ratios of weakly/partially labeled data of sound events in terms of macro-Fscore.

60–70%. This result indicates that the proposed semi-MTL approach deliver reasonable performance even when only a small number of strongly labeled data are available alongside a large number of partially labeled data. Consequently, the proposed methods alleviate the challenges associated with annotating sound event labels. When comparing the method based on the semi-supervised MTL using weak labels with that using partial labels, we observed nearly equivalent performance in both these methods except when all the training data have weak or partial labels. This suggests that if part of audio data for the model training do not have strong labels, generating partial labels using LLMs instead of annotating weak labels would be a reasonable solution. On the other hand, when all training data consist of weak partial labels, the SED performance can degrade significantly. This result suggests that incorporating strong labels with partial labels and applying semi-supervised learning can substantially enhance the reliability of detection results.

Furthermore, since the results of the proposed method are comparable to those of the Semi-MTL w/ weak labels, it implies that the proposed method remains effective even when using partial labels of the quality shown in Table 1.

Table 6: ASC performance for each scene in terms of Fscore.

Method	city center	home	office	residential area
Strong MTL	91.38% ±2.38	94.23% ±3.47	96.54% ±1.85	84.57% ±6.28
Weak MTL	90.45% ±2.57	93.40% ±4.91	94.16% ±3.22	81.95% ±10.27
Strong MTL w/ reduced data	**92.27% ±2.23**	93.98% ±2.73	**96.63% ±1.69**	84.56% ±5.75
Semi-MTL w/ weak labels	90.71% ±2.39	95.77% ±3.09	95.63% ±2.25	86.20% ±5.82
Semi-MTL w/ partial labels (proposed)	90.40% ±4.13	**96.44% ±1.41**	96.52% ±1.88	**86.97% ±4.01**

Thus, the SED performance of the proposed method is comparable across reasonable variations in the size of the sound event label set between that of the actual weak label and the current partial label sets, suggesting that the proposed method is robust to the size of the partial label set.

4.2.3 Detailed Performance Evaluation for Each Acoustic Scene and Sound Event

Table 6 shows the detailed ASC performance for each acoustic scene. The result indicates that there are no significant differences in ASC performance among the strong and semi-MTL approaches. This also implies that the temporal information on sound events is not critical for each scene classification, and that clip-level sound event information is sufficient for ASC.

Table 7 presents the SED performance and sound duration for each sound event. These results indicate that the proposed semi-supervised MTL approach using partial labels achieves comparable performance to the method using weak labels in detecting each sound event. Furthermore, for sound events with longer durations, such as *bird singing*, *fan*, and *large vehicle*, the SED performance is comparable to that of strong MTL. However, for sound events with short duration, such as *cutlery* and *keyboard typing*, the performance of the proposed method slightly degrades compared with strong MTL. It is known that the SED model trained with strong labels tends to fail to detect short-duration events compared with that trained with weak labels [11, 10].

To further investigate this result, Table 8 shows the numbers of true positives (# TP), false positives (# FP), and false negatives (# FN) for each sound event. Although the proposed method achieves improvements in TP count, it also exhibits an increase in FP count. This indicates that the proposed method tends to be overconfident in the detection of sound events. We attribute the overconfidence to confirmation bias [1], which reinforces errors in pseudo labels through iterative label refining. In our proposed method, the partial labels are self-distillated and the MTL model is retrained using the

Table 7: SED performance for each sound event in terms of segment-based Fscore and sound duration. We conducted the experiments with 30% of the strongly labeled data and 70% of the weakly/partially labeled data under the semi-MTL condition.

Method	bird singing	brakes squeaking	car	cutlery	dishes	fan	glass jingling	keyboard typing	large vehicle	mouse clicking	people walking	washing dishes
Strong MTL	40.74% ±4.48	**51.64% ±5.31**	**51.80% ±2.04**	**12.22% ±7.90**	12.43% ±4.77	**97.03% ±1.20**	0.27% ±0.71	**56.20% ±3.52**	17.39% ±1.54	**71.30% ±1.86**	**19.35% ±3.01**	5.46% ±4.34
Weak MTL	28.46% ±17.61	21.97% ±19.38	30.73% ±16.26	10.28% ±9.55	8.38% ±6.27	50.47% ±45.46	**3.57% ±3.84**	10.76% ±10.29	9.75% ±6.21	1.69% ±1.83	11.34% ±2.77	10.86% ±10.48
Strong MTL w/ reduced data	37.08% ±8.15	8.59% ±11.03	49.35% ±4.19	0.08% ±0.36	3.52% ±4.84	95.17% ±2.03	0.00% ±0.00	46.23% ±9.11	18.80% ±3.11	27.68% ±22.40	14.01% ±4.97	12.49% ±12.95
Semi-MTL w/ weak labels	39.94% ±6.64	34.44% ±15.14	49.97% ±2.85	8.12% ±9.17	**13.77% ±7.45**	95.74% ±2.44	0.36% ±1.00	52.48% ±4.21	**19.42% ±4.24**	67.63% ±5.09	17.22% ±5.27	**12.70% ±11.43**
Semi-MTL w/ partial labels (proposed)	**42.08% ±7.85**	37.77% ±11.91	50.60% ±2.48	4.83% ±5.93	11.21% ±5.64	96.50% ±1.42	0.81% ±1.93	50.90% ±6.67	18.09% ±3.25	69.15% ±2.17	17.41% ±5.80	11.23% ±12.86
Average sound duration (s)	7.63 ±8.49	1.65 ±1.97	6.88 ±4.72	0.74 ±0.53	1.24 ±1.12	29.99 ±0.01	0.80 ±0.46	0.21 ±0.22	14.68 ±7.35	0.14 ±0.08	6.63 ±8.78	4.15 ±3.75

Table 8: Average numbers of true positive, false positive, and false negative samples for each sound event. We conducted the experiments with 30% of the strongly labeled data and 70% of the weakly/partially labeled data under the semi-MTL condition.

Method	Metric	bird singing	brakes squeaking	car	cutlery	dishes	fan	glass jingling	keyboard typing	large vehicle	mouse clicking	people walking	washing dishes
Strong MTL	# TP	6,745.1	1,767.0	20,341.2	67.4	262.4	37,386.1	0.4	1,131.7	2,822.1	515.6	1,833.7	230.1
	# FP	5,643.2	998.4	24,300.0	53.3	380.5	745.3	3.8	653.4	23,550.3	125.6	4,550.5	1,039.3
	# FN	13,743.0	2,270.1	13,539.8	877.6	3,203.6	1,559.9	269.6	1,092.3	3,340.9	289.4	10,683.3	6,310.9
Weak MTL	# TP	8,286.1	2,215.8	27,726.2	70.3	702.8	37,360.5	2.8	2,149.9	3,267.8	673.6	2,905.1	1,282.8
	# FP	10,596.8	12,257.7	53,787.9	2,205.9	5,680.4	2,207.6	485.4	23,670.9	32,417.8	24,349.7	24,488.0	4,777.8
	# FN	12,202.0	1,821.2	6,154.8	874.7	2,763.2	1,585.5	267.2	74.1	2,895.2	131.4	9,612.0	5,258.2
Strong MTL w/ reduced data	# TP	6,075.9	204.0	19,592.0	0.4	80.9	37,376.7	0.0	983.7	2,858.7	150.4	1,593.5	739.0
	# FP	5,296.1	36.6	25,261.8	0.2	221.4	2,256.4	0.0	951.6	22,369.6	11.8	6,594.2	2,191.1
	# FN	14,412.1	3,833.0	14,289.0	944.6	3,385.1	1,569.3	270.0	1,240.3	3,304.3	654.6	10,923.5	5,802.1
Semi-MTL w/ weak labels	# TP	6,679.1	1,001.0	19,050.3	45.8	332.0	37,339.6	0.5	1,141.3	2,430.1	470.9	1,641.5	678.4
	# FP	5,689.3	447.2	23,169.8	37.7	636.2	1,769.6	4.1	998.8	17,169.1	111.9	4,688.4	2,101.3
	# FN	13,809.0	3,036.1	14,830.7	899.2	3,134.1	1,606.4	269.5	1,082.7	3,732.9	334.1	10,875.5	5,862.6
Semi-MTL w/ partial labels (proposed)	# TP	8,012.2	1,407.2	24,614.3	57.7	434.6	37,397.0	1.3	1,369.2	3,081.1	552.6	2,051.9	974.4
	# FP	7,870.2	1,456.1	36,710.0	105.0	905.7	536.5	10.8	1,469.3	27,294.4	268.1	9,644.0	2,711.4
	# FN	12,475.8	2,629.8	9,266.7	887.3	3,031.4	1,549.1	268.7	854.8	3,082.0	252.4	10,465.1	5,566.6

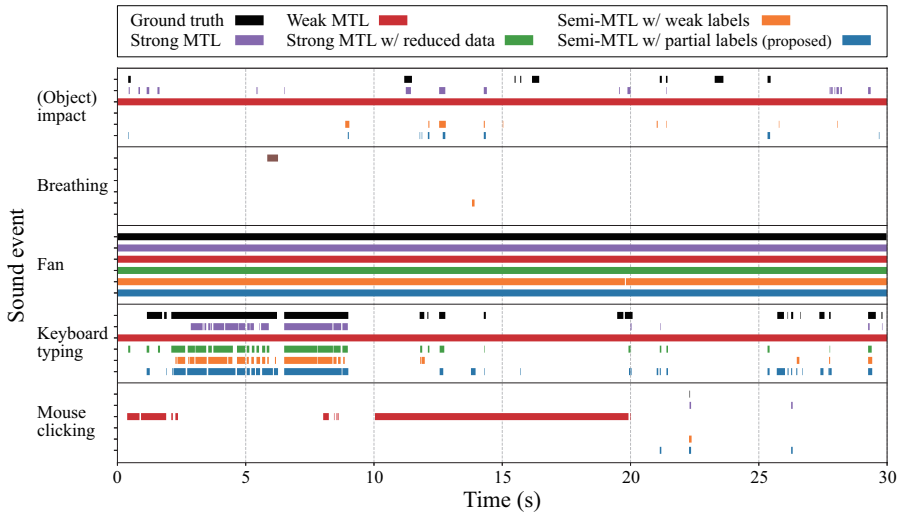

Figure 8: Sound event detection results for 276.wav recorded in an office scene from the TUT Acoustic Scenes 2016 dataset. Only sound events that include multiple ground truth labels or detected events are shown. We conducted the experiments with 30% of the strongly labeled data and 70% of the weakly/partially labeled data under the semi-MTL condition.

distillated labels. This procedure tends to amplify confidence of the distillated labels. In particular, prior work [1] pointed out that confirmation bias becomes more serious near detection boundaries. For SED, short duration events tend to contain many boundary frames relative to their total frames. As a result, the proposed method result in more TP and FP counts for short duration classes.

4.2.4 Qualitative Analysis of Sound Event Detection Results

To qualitatively assess the behavior of the proposed method, Figures 8 and 9 show the detection results on randomly selected sound clips. Each figure presents the ground truth of sound event labels, the detection outputs from the conventional and proposed methods.

In Figure 8, the proposed method shows a more accurate detection performance for the *keyboard typing* event than the conventional strong MTL and semi-MTL methods with weak labels. In addition, we observed false positives where events were detected at the correct time boundary but with incorrect labels; for example, *keyboard typing* and *mouse clicking* were detected instead of *(object) impact*. For these cross-triggering cases, incorporating a more refined mechanism for sound event classification may help mitigate such errors.

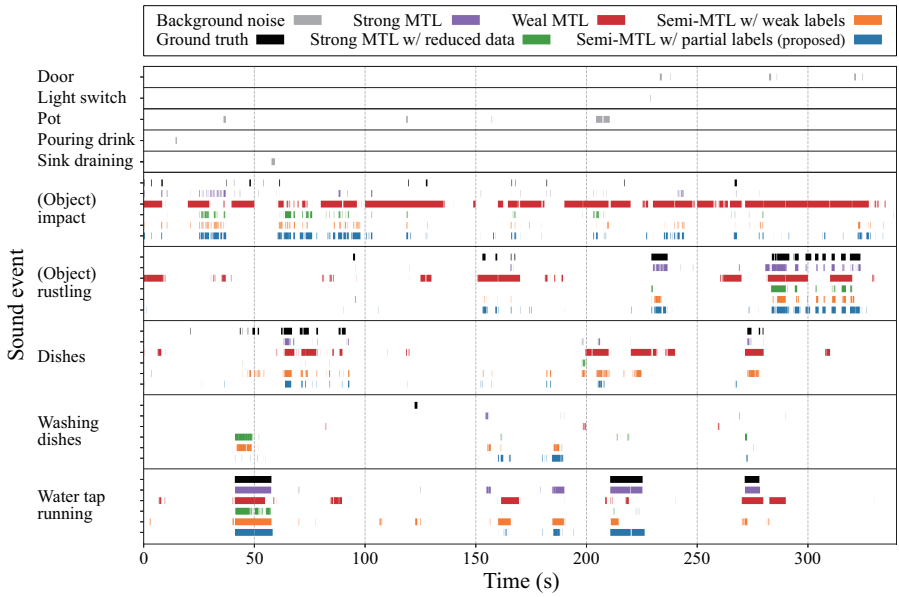

Figure 9: Sound event detection results for b043.wav recorded in a home scene from the TUT Sound Events 2016 dataset. The figure includes the additional visualization of background noise not annotated in the ground truth. Only sound events that include multiple ground truth labels or detected events are shown. We conducted the experiments with 30% of the strongly labeled data and 70% of the weakly/partially labeled data under the semi-MTL condition.

Figure 9 includes the additional visualization of background noise, which corresponds to sound events not annotated as ground truth labels. These visualizations enable us to assess the model robustness to background sounds. The results indicate that the proposed method is as robust as the conventional strong MTL and semi-MTL methods in ignoring irrelevant background noise, and it still can detect target sound events.

4.2.5 Model Complexity and Training Cost

Table 9 shows the numbers of model parameters and training costs for the proposed and conventional methods. Note that, in the proposed method, the parameters used in the distillation module can be reused within the main module, which eliminates the use of additional model parameters. As shown in the table, there are no significant differences in the number of model parameters and training time. This indicates that our proposed method can be implemented without considerably increasing additional computational cost or memory requirements.

Table 9: Comparison of model size and training cost among proposed and conventional methods. For the semi-MTL conditions, we conducted the experiments using 30% of the strongly labeled data and 70% of the weakly/partially labeled data.

Method	**# parameters (k)**	**Training time (s)**	**Training time ratio (Strong MTL = 1.0)**
Strong MTL	1,323	314.8 ± 2.2	1.00
Weak MTL	1,323	311.9 ± 1.3	0.99
Strong MTL w/ reduced data	1,323	118.8 ± 1.0	0.37
Semi-MTL w/ weak labels	1,331	327.1 ± 1.9	1.04
Semi-MTL w/ partial labels (proposed)	1,331	340.2 ± 2.5	1.08

5 Conclusions

We proposed the method for the joint analysis of acoustic scenes and sound events based on the semi-supervised SED strategy using partial labels of sound events. We further introduced the LLM-based label creation and self-distillation-based label refining methods for the proposed partial label learning in SED. The results of experiments using our constructed dataset show that the semi-supervised approach using partial labels achieve reasonable performance even with a small number of strongly labeled data and a large number of partially labeled data. Future work should focus on exploring more effective approaches to refining partial labels of sound events. Also, the application of partial label learning to single-task SED settings where acoustic scene labels are not available should be addressed. This will require new strategies for generating candidate event label sets without scene context, which poses a more challenging and general problem.

Appendix: Prompts Used to Generate Partial Labels of Sound Events

To generate partial labels of sound events, we utilized the ChatGPT o3-mini-high on February 02, 2025. The input prompts used to generate partial labels are shown in Table 10, which includes the possible sound events, the supplemental explanation of a sound event class, the instruction to consider the partial labels of sound events for each scene, and the output format. We obtain partial labels and the reasons for including the sound events in the list. The lists of partial labels and reasons are available.[5]

[5] https://github.com/KeisukeImoto/SED_ASC_partial_label.git.

Table 10: Prompts for generating partial labels of sound events input into ChatGPT o3-mini-high.

Here is the list of 25 possible sound events: object banging, object impact, object rustling, object snapping, object squeaking, bird singing, brakes squeaking, breathing, car, children, cupboard, cutlery, dishes, drawer, fan, glass jingling, keyboard typing, large vehicle, mouse clicking, mouse wheeling, people talking, people walking, washing dishes, water tap running, wind blowing. Here, "object" refers to an unknown sound source, although we can understand how the sound is produced. We can include these ambiguous object sounds in the list. If we are in a <scene name> scene, which sound events are likely to be heard? Please list all the sound events one by one (without merging) in CSV format, and provide your reasoning process in a two-column CSV format.

References

[1] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning", *arXiv, arXiv:1908.02983*, 2019.

[2] H. L. Bear, I. Nolasco, and E. Benetos, "Towards Joint Sound Scene and Polyphonic Sound Event Recognition", *Proc. INTERSPEECH*, 2019, 4594–8.

[3] C. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulovic, "A Framework for the Robust Evaluation of Sound Event Detection", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2020, 61–5.

[4] E. Çakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection", *IEEE/ACM Trans. Audio Speech Lang. Process.*, 25(6), 2017, 1291–303.

[5] T. Cour, B. Sapp, and B. Taskar, "Learning from Partial Labels", *Journal of Machine Learning Research*, 12(42), 2011, 1501–36.

[6] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the Multiple Instance Problem with Axis-parallel Rectangles", *Artificial Intelligence*, 89(1), 1997, 31–71.

[7] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, X. Favory, J. Jordi, and X. Serra, "General-purpose Tagging of Freesound Audio with AudioSet Labels: Task Description, Dataset, and Baseline", *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), 2018, 69–73.

[8] L. Ford, H. Tang, F. Grondin, and J. Glass, "A Deep Residual Network for Large-Scale Acoustic Scene Analysis", *Proc. INTERSPEECH*, 2019, 2568–72.

[9] Y. Hou, S. Song, C. Yu, W. Wang, and D. Botteldooren, "Audio Event-Relational Graph Representation Learning for Acoustic Scene Classification", *IEEE Signal Processing Letters*, 2023, 1–5.

[10] A. Igarashi, S. Tsubaki, D. Niizumi, D. Takeuchi, N. Harada, and K. Imoto, "Joint Analysis of Acoustic Scenes and Sound Events Based on Semi-Supervised Approach", *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (*APSIPA ASC*), 2023, 2050–6.

[11] K. Imoto, S. Mishima, Y. Arai, and R. Kondo, "Impact of Data Imbalance Caused by Inactive Frames and Difference in Sound Duration on Sound Event Detection Performance", *Applied Acoustics*, 196(108882), 2022.

[12] K. Imoto and N. Ono, "Acoustic Topic Model for Scene Analysis With Intermittently Missing Observations", *IEEE/ACM Trans. Audio Speech Lang. Process.*, 27(2), 2019, 367–82.

[13] J. Jung, H. J. Shim, J. H. Kim, and H. J. Yu, "DCASENET: An Integrated Pretrained Deep Neural Network for Detecting and Classifying Acoustic Scenes and Events", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2021, 621–5.

[14] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Sound Event Detection of Weakly Labelled Data with CNN-Transformer and Automatic Threshold Optimization", *IEEE/ACM Trans. Audio Speech Lang. Process.*, 28, 2020, 2450–60.

[15] 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 Trans. Audio Speech Lang. Process.*, 28, 2020, 2880–94.

[16] A. Kumar and B. Raj, "Audio Event Detection Using Weakly Labeled Data", *Proc. ACM International Conference on Multimedia* (*ACMMM*), 2016, 1038–47.

[17] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the Variance of the Adaptive Learning Rate and Beyond", *Proc. International Conference on Learning Representations* (*ICLR*), 2020, 1–13.

[18] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, B. Raj, and T. Virtanen, "DCASE 2017 Challenge Setup: Tasks, Datasets and Baseline System", *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), 2017, 85–92.

[19] A. Mesaros, T. Heittola, and A. Klapuri, "Latent Semantic Analysis in Sound Event Detection", *Proc. European Signal Processing Conference* (*EUSIPCO*), 2011, 1307–11.

[20] A. Mesaros, T. Heittola, and T. Virtanen, "TUT Database for Acoustic Scene Classification and Sound Event Detection", *Proc. European Signal Processing Conference* (*EUSIPCO*), 2016, 1128–32.

[21] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Conformer-Based Sound Event Detection with Semi-Supervised Learning and Data Augmentation", *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), 2020, 100–4.

[22] V. Morfi, R. F. Lachlan, and D. Stowell, "Deep Perceptual Embeddings for Unlabelled Animal Sound", *IEEE/ACM Trans. Audio Speech Lang. Process.*, 150(1), 2020, 2–11.

[23] T. Nishida, K. Dohi, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Anomalous Sound Detection Based on Machine Activity Detection", *Proc. European Signal Processing Conference* (*EUSIPCO*), 2022, 269–73.

[24] N. Tonami, K. Imoto, M. Niitsuma, R. Yamanishi, and Y. Yamashita, "Joint Analysis of Acoustic Events and Scenes Based on Multitask Learning", *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), 2019, 333–7.

[25] S. Tsubaki, K. Imoto, and N. Ono, "Joint Analysis of Acoustic Scenes and Sound Events with Weakly labeled Data", *Proc. International Workshop on Acoustic Signal Enhancement* (*IWAENC*), 2022, 1–5.

[26] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound Event Detection in Domestic Environments with Weakly Labeled Data and Soundscape Synthesis", *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), 2019, 253–7.

[27] M. Valenti, S. Squartini, A. Diment, G. Parascandolo, and T. Virtanen, "A Convolutional Neural Network Approach for Acoustic Scene Classification", *Proc. International Joint Conference on Neural Networks* (*IJCNN*), 2017, 1547–54.

[28] Y. Wang, J. Li, and F. Metze, "A Comparison of Five Multiple Instance Learning Pooling Functions for Sound Event Detection with Weak Labeling", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2019, 31–5.