EXPLORING MULTI-TASK LEARNING WITH WEIGHTED SOFT LABEL LOSS FOR SOUND EVENT DETECTION WITH SOFT LABELS

Tanmay Khandelwal and Rohan Kumar Das

Fortemedia Singapore, Singapore

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

ABSTRACT

The learning of sound events often depends on data that is manually labeled by human annotators. In this study, we explore the use of soft labels for sound event detection (SED), which takes into account the uncertainty and variability in human annotations. To address the challenges posed by uncertain or noisy labels, we propose a weighted soft label (WSL) loss function. This loss function effectively emphasizes reliable annotations while mitigating the influence of less confident or noisy labels. Additionally, we introduce auxiliary tasks into a multi-task learning (MTL) framework, which helps to leverage the shared information between the tasks and improves the overall performance of the model. Furthermore, we explore the usage of pretrained models and various front-end feature extraction methods. Experimental results on the MAESTRO-Real dataset introduced in the DCASE 2023 Task 4B demonstrate a significant improvement of 14.9% in the macro-average F1 score with optimum threshold per class compared to the challenge baseline model on the validation set, highlighting the effectiveness of our proposed system.

***Index Terms*—** sound event detection, soft labels, multi-task learning, acoustic scenes, weighted loss

1. INTRODUCTION

The primary aim of sound event detection (SED) is to autonomously identify and extract significant information from audio recordings, enabling the detection of specific events or activities. SED holds immense potential to augment diverse domains, leading to enhanced safety, convenience, and efficiency. It already plays a critical role in a wide range of applications, including surveillance systems [1, 2], acoustic monitoring [3], smart-homes [4–6], and human-computer interaction.

The lack of labeled training data presents a notable challenge in SED. The process of collecting and annotating extensive audio datasets with labeled sound events is time-consuming and demanding. The scarcity of annotated data impedes the training of accurate models and limits their performance. As a result, researchers are exploring alternative techniques such as employing soft-labeling training methods and transfer learning to mitigate this issue. Soft labels provide a representation of the degree of presence or confidence for specific sound events in each audio segment or frame, in contrast to hard labels that assign binary labels (e.g., 0 or 1). By incorporating confidence scores, soft labels effectively capture the uncertainty and variability associated with sound events, facilitating more nuanced analysis and decision-making processes.

In addition to soft-label generation, researchers are also investigating transfer learning as a means to enhance SED. Transfer learning enables the utilization of knowledge acquired from pretrained models on different but related tasks. Instead of training a model from scratch on a specific SED task, transfer learning allows the model to benefit from the learned representations and features of a pretrained model. Previous works have shown the effectiveness of using the features from pretrained models like pretrained audio neural networks (PANNs) [7–9], audio spectrogram transformers (ASTs) [10], and bidirectional encoder representation from audio transformers (BEATs) [11], trained on a large dataset. The models are fine-tuned using their learned features, customized to the specific SED task at hand, leading to improved performance.

Sound events occurring in nature are typically intricately linked with acoustic scenes. An acoustic scene encompasses the auditory environment in which sound events occur, reflecting the distinctive combination of various sound sources, background noise, and spatial characteristics. Understanding and analyzing sound events within their corresponding acoustic scenes play a pivotal role in SED and related applications. For instance, in the acoustic scene "cafe" the sound events "coffee machine" and "cutlery and dishes" are likely to occur, whereas the sound events "bird singing" and "wind blowing" occur infrequently. On the basis of these previous methods, [12] has proposed methods of SED that take into account acoustic scene information in an unsupervised manner. [13,14] have proposed scene classification methods considering sound events using Bayesian generative models. Similarly, the methods proposed in [15, 16] focus on the joint analysis of acoustic scenes and sound events using neural network models based on multi-task learning (MTL). Such MTL-based methods leverage the existing knowledge and reduce the need for manual labeling, thus effectively addressing the challenge of data scarcity.

The detection and classification of acoustic scenes and events (DCASE) 2023 edition has recently introduced a new subtask, 4B [17], which aims to explore the potential benefits of incorporating soft labels in improving performance. In our study, we extend the idea of integrating soft labels into the training procedure of SED models. Our investigation specifically revolves around the utilization of soft labels using this newly released dataset in the DCASE 2023 Task 4B dataset [18]. This dataset was specifically designed for exploring the estimation of strong labels through crowdsourcing. It consists of 49 real-life audio files captured from 5 distinct acoustic scenes, accompanied by their corresponding annotation outcomes. To effectively leverage the soft-level probabilities provided in the dataset, we propose a novel weighted soft label (WSL) loss function that mitigates the impact of less confident or noisy labels. Moreover, we delve into the integration of two auxiliary tasks within an MTL framework to enhance the effectiveness of the SED model. To further improve the model's capabilities, we also explore the utilization of pretrained models and different front-end methods for feature extraction.

Table 1: Categorization of acoustic events into different acoustic scenes for the MAESTRO-Real dataset.

Acoustic scene \ Acoustic event	Bird singing	Car	People talking	Footsteps	Children voices	Wind blowing	Brakes squeaking	Large vehicle	Cutlery and dishes	Furniture and dragging	Coffee machine	Metro approaching	Metro leaving	Door opens/closes	Announcement	Shopping cart	Cash register beeping
Cafe/Restaurant				✓	✓			✓	✓	✓	✓						
City center		✓	✓	✓	✓		✓	✓									
Grocery store			✓	✓	✓										✓	✓	✓
Metro station			✓	✓	✓							✓	✓	✓			
Residential area	✓	✓	✓	✓	✓	✓											

2. PROPOSED METHODS

2.1. Multi-task learning (MTL) framework

Conventionally, acoustic scenes and sound events have been treated as separate entities in most methods. However, in reality, acoustic scenes play a crucial role in shaping the perception and interpretation of sound events by providing a contextual backdrop. Recognizing the significance of this relationship, we aim to leverage it to gain valuable insights that can enhance SED methods. In this study, we leverage the five acoustic scenes available in the DCASE 2023 Task 4B dataset, and provide a summary of the sound events that take place in these acoustic scenes, as shown in Table 1.

From the table, it is evident that certain sound events, such as "shopping cart" occur exclusively in a specific acoustic scene and are not present in any other acoustic scene. Similarly, the sound event "bird singing" is only observed in residential areas and not in any other acoustic scene. Additionally, we notice that some events, like "footsteps" and "children voices" are common across multiple acoustic scenes. As a result, we propose an additional task of classifying the acoustic environment associated with a sound event as either indoor (I) or outdoor (O). This classification helps to differentiate the surroundings in such sound events. We present two additional tasks related to acoustic scenes: (1) categorizing the acoustic scene for each frame where a sound event takes place, termed acoustic scene classification (ASC), and (2) determining whether each frame's sound event occurs indoors or outdoors, known as acoustic environment classification (AEC). As depicted in Table 2 the acoustic scenes associated with the sound events are separated into five different classes. Additionally, we determined whether the acoustic scenes were indoors (I) or outdoors (O) based on their respective environments. To enhance the performance of the SED model, we integrate the information from these two auxiliary tasks into the primary SED branch.

Table 2: Classification of the 5 acoustic scenes into different scene labels and environment labels.

Acoustic scene	Scene label	Environment	Environment label
Cafe/Restaurant	A	indoor	I
City center	B	outdoor	O
Grocery store	C	indoor	I
Metro station	D	indoor	I
Residential area	E	outdoor	O

In order to capture low-level features that can benefit all three tasks, we design the network to share certain common layers. These shared layers facilitate the extraction of features that are relevant to all tasks. Previous studies in [19] have demonstrated that leveraging knowledge from easier tasks can improve the performance of harder tasks. In our case, we consider SED as the most challenging task, followed by ASC, and finally AEC. Therefore, we anticipate that the two auxiliary tasks will contribute to improving the SED performance. To conduct the joint training with these two tasks, we use a combined loss function L_{MTL}, which is the weighted loss function. It can be expressed mathematically as

$$L_{MTL} = \alpha \times L_{SED} + \beta \times L_{ASC} + \gamma \times L_{AEC} \tag{1}$$

where α, β, and γ are the trade-off factors that regulate the weighted loss. By adopting an MTL framework with joint training, we benefit from the fact that once the MTL-based model is trained, the auxiliary branches can be removed from the model architecture. During inference, only the single SED branch is utilized, ensuring that the number of parameters remains the same as that of a single SED branch.

2.2. Weighted soft label (WSL) loss

The DCASE 2023 Task 4B baseline uses mean-square error (MSE) loss, to teach the system to predict outputs as close as possible to the provided soft activity indicators instead of binary as described below:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}(y_{ij} - p_{ij})^2 \tag{2}$$

where N represents the total number of samples, C is the number of classes, y_{ij} is the ground truth soft label for sample i and class j, and p_{ij} is the predicted value for sample i and class j. We extend this loss function to incorporate weights derived from the probabilities assigned to the soft labels by the annotator. Our proposed weighted soft label (WSL) loss function assigns varying importance to each prediction based on its associated probability, as described below:

$$WSL = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} \cdot (y_{ij} - p_{ij})^2 \tag{3}$$

where y_{ij} also acts as the weight assigned to the soft label for sample i and class j. Higher weight is given to predictions with higher

probabilities, indicating a higher level of confidence in those predictions. This weighted approach allows the model to focus more on accurately predicting instances with higher probabilities while considering the uncertainty associated with softer labels. As a result, the model can learn to optimize its performance by prioritizing predictions based on their probability-weighted importance, leading to improved accuracy and robustness.

3. ARCHITECTURE

3.1. Baseline

The baseline system [17] for DCASE 2023 Task 4B adopts the convolutional recurrent neural network (CRNN) architecture with a linear output layer. The convolutional neural network (CNN) component of the model consists of three layers, each featuring 128 filters. A kernel size of 3×3 is applied to each convolutional layer, followed by the activation function rectified linear unit and batch normalization [20]. Frequency and temporal pooling are performed using a max pooling layer with sizes of [[1, 5], [1, 2], [1, 2]], respectively. To mitigate overfitting, a dropout rate of 0.2 is applied after each layer. This is followed by the recurrent neural network (RNN) block, consisting of a single layer of 32 bidirectional gated recurrent units (Bi-GRUs) [21].

3.2. Proposed architecture

In this study, we incorporate large-scale PANNs [7] into our approach due to resource limitations. The PANNs have been pretrained on the extensive Audioset dataset, which consists of 5000 hours of audio spanning 527 sound classes. By leveraging the pre-existing knowledge encoded in these pretrained models, we aim to replace the CNN component of the baseline model with PANNs, thereby benefiting from their learned representations and features. The PANNs architecture comprises 6 convolutional blocks, with each block consisting of 2 convolutional layers using a 3×3 kernel size. In our study, we investigate the extraction of embeddings after each convolutional block within the PANNs model. These embeddings are subsequently inputted into a single-layer Bi-GRU containing 256 hidden units. The complete CRNN model, encompassing both the PANNs and Bi-GRU components, is unfrozen and trained throughout the experimentation process. Additionally, to include the two supplementary tasks outlined in Section 2.1, we uphold the CNN component as the shared element across all tasks. Furthermore, we integrate distinct Bi-GRUs and output layers for each distinct task, as depicted in Figure 1, each sized at (200×17), (200×5), and (200×2) for the SED, ASC, and AEC branches, respectively.

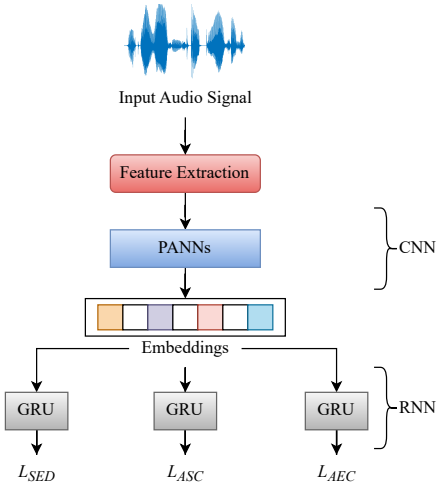

Figure 1: The proposed MTL framework with PANNs along with three parallel Bi-GRUs for different tasks.

4. EXPERIMENTAL SETUP

4.1. Dataset

This study utilizes the multi-annotator estimated strong (MAESTRO)-Real dataset [18] released for the DCASE 2023 Task 4B. The dataset comprises 49 real-life audio files captured from 5 distinct acoustic scenes and includes corresponding annotation outcomes. The total duration of the dataset amounts to 189 minutes and 52 seconds. The audio files are a subset of the TUT Acoustic Scenes 2016 dataset and encompass five acoustic scenes: cafe/restaurant, city center, grocery store, metro station, and residential area. Each scene consists of 6 classes, with some classes being common across all scenes, resulting in a total of 17 classes as presented in Table 1. The dataset consists of the following components: (1) audio recordings comprising the 49 real-life recordings, each ranging from 3 to 5 minutes in length, and (2) soft labels representing estimated strong labels with a time resolution of 1s obtained through crowdsourced data, with values ranging between 0 and 1 indicating the certainty of the annotators. The soft labels follow a format that includes the start time, end time, textual label, and a corresponding value indicating the soft label for each event class within the given segment. For example: “2 3 car 0.9”, “2 3 footsteps 0.7”, and so on.

4.2. Feature extraction and training

For the baseline system [17], a batch size of 32 is employed, and the input features are mel-band energies extracted using a hop length of 200 ms and 64 mel filter banks. Additionally, we explored different front-end feature extraction techniques such as mel-frequency cepstral coefficient (MFCC), linear frequency cepstral coefficient (LFCC), and constant-Q transform (CQT) to replace the log-mel spectrogram. The DCASE 2023 Task 4B dataset is organized according to a 5-fold cross-validation setup, where around 70% of the data per class is allocated for training, and the remaining portion is dedicated to testing. To optimize the training process, we employ the Adam [22] optimizer, with an initial learning rate of 0.001. The training process is executed over a total of 150 epochs, utilizing the computational power of the Nvidia RTX A4000.

4.3. Evaluation

In this study, we utilize the macro-average segment-F1 score ($F1_{MO}$) under the optimum threshold [23] as our primary evaluation metric. It is calculated over 1s segments, following the same approach as the DCASE 2023 Task 4B challenge. The $F1_{MO}$ score considers the best F1 score per class achieved with a class-specific threshold. Additionally, we report the micro-average F1 score ($F1_m$), micro-average error rate (ER_m), and macro-average F1 score ($F1_M$) calculated over 1s segments using a decision threshold of 0.5 applied to the system output.

5. EXPERIMENTAL RESULTS

In this part, we present outcomes of the proposed methods, including ablation studies on the DCASE 2023 Task 4B validation set.

5.1. Architecture with feature extraction

We first present the outcome in Table 3 obtained for the baseline as reported by the organizers of the DCASE 2023 Task 4B. The baseline incorporates the log-mel spectrogram with the configuration specified in Section 4.2. Subsequently, we substitute the baseline architecture with the proposed architecture described in Section 3.2, which utilizes PANNs. When using PANNs, we extract the embeddings after the 6th block. Our observations show that employing PANNs with log-mel spectrogram alone enhances the $F1_{MO}$ score from 42.8 to 45.4 as represented in Table 3. The following analysis compares various commonly employed feature extraction methods discussed in Section 4.2. Our findings reveal that the MFCC-based feature extraction method outperforms the log-mel spectrogram utilized in the baseline, as well as the LFCC and CQT front-ends. It improves the $F1_{MO}$ score for the 6-blocks-based PANNs to 46.5 from 45.4. Having determined MFCC as the chosen feature extraction method, we proceed to explore the layer from which we extract the embeddings. We decrease it from the 6th Block to the 3rd Block and conduct experiments accordingly. Through this analysis, we discover that extracting embeddings after the 4th Block yields the most significant improvement in the $F1_{MO}$ score, increasing it from 46.5 to 48.2.

5.2. WSL loss function

Once we determine that the highest score is achieved by extracting embeddings after the 4th Block, we introduce the WSL loss function, as outlined in Section 5.2. The loss function prioritizes the learning of well-defined patterns while minimizing the influence of ambiguous or noisy instances. Consequently, this enhancement leads to an improvement in the $F1_{MO}$ score, increasing it from 48.2 to 48.9.

5.3. MTL framework

To enhance our system, we introduce the MTL framework comprising two auxiliary branches in addition to the primary SED branch. In an ablation study, we compare the performance of the proposed system (PANNs+WSL) by incorporating different MTL branches. Initially, we integrate only the ASC branch with the SED branch

Table 3: Comparison of performance, showing the impact of architectural changes and variations in feature extraction methods.

System	Blocks	Feature	ER_m	$F1_m$	$F1_M$	$F1_{MO}$
Baseline	-	Log-mel	0.487	70.34	35.83	42.8
PANNs	6 Blocks	Log-mel	0.442	72.64	36.97	45.4
PANNs	6 Blocks	CQT	0.493	67.53	31.84	42.0
PANNs	6 Blocks	LFCC	0.447	71.5	31.75	46.0
PANNs	6 Blocks	MFCC	0.415	74.18	34.33	**46.5**
PANNs	5 Blocks	MFCC	0.410	75.1	37.21	48.0
PANNs	4 Blocks	MFCC	0.408	76.74	39.42	**48.2**
PANNs	3 Blocks	MFCC	0.470	73.5	39.35	46.2

Table 4: Illustration of performance improvement following the implementation of the WSL loss function.

System	Feature	ER_m	$F1_m$	$F1_M$	$F1_{MO}$
Baseline	Log-mel	0.487	70.34	35.83	42.8
PANNs (4 Blocks) + WSL	MFCC	0.416	75.61	38.60	**48.9**

Table 5: Ablation study for analyzing the contribution of each branch.

System	MTL	ER_m	$F1_m$	$F1_M$	$F1_{MO}$
Baseline	-	0.487	70.34	35.83	42.8
PANNs (4 Blocks) + WSL	SED + ASC	0.416	76.33	39.65	49.2
PANNs (4 Blocks) + WSL	SED + AEC	0.412	76.29	40.85	49.0
PANNs (4 Blocks) + WSL	SED + ASC + AEC	0.406	76.61	39.87	**49.3**

with (α=0.85, β=0.15, and γ=0). After tuning the weights in the loss function, this configuration achieves the highest $F1_{MO}$ score of 49.2. Next, we replace the ASC branch with the AEC branch (α=0.85, β=0, and γ=0.15), which results in a $F1_{MO}$ score of 49.0. Finally, we introduce all three branches, including the SED, ASC, and AEC branches, with tuned hyperparameters (α=0.85, β=0.1, and γ=0.05). This configuration yields the best overall score of 49.3, demonstrating the effectiveness of the MTL framework and the impact of each auxiliary branch.

5.4. System comparison

Our experiments come to a close as we present the results of comparing our system with other high-performing submissions for DCASE 2023 Task 4B. Table 6 displays the reported performances of the baseline system as well as other systems, sorted based on the $F1_{MO}$ score. We observe that our system achieves a performance comparable to other systems while demonstrating an improvement of 14.9% over the baseline system. Additionally, it is worth noting that our system outperforms the 3rd system [24] in all metrics besides the $F1_{MO}$ score.

Table 6: Performance comparison of our proposed system with other submissions in DCASE 2023 Task 4B.

System	ER_m	$F1_m$	$F1_M$	$F1_{MO}$
Xu-SJTU-task4b-3 [25]	0.246	86.13	57.91	69.85
Bai-JLESS-task4b-4 [26]	0.360	78.63	42.45	56.16
Liu-SRCN-task4b-2 [24]	0.430	72.90	28.80	49.70
PANNs (4 Blocks) + WSL + MTL (**Ours**)	0.406	76.61	39.87	**49.30**
Nhan-VNUHCMUS-task4b-1 [27]	0.450	72.43	37.32	46.71
Min-KAIST-task4b-1 [28]	0.445	72.78	36.12	45.81
Cai-NCUT-task4b-1 [29]	0.439	74.84	39.57	43.50
Baseline [17]	0.487	70.34	35.83	42.8

6. CONCLUSION

In this study, we present our methods for sound event detection using soft labels introduced in DCASE 2023 Task 4B. We propose several novel approaches and demonstrate their effectiveness through our findings. Firstly, we suggest using PANNs embeddings and modifying the feature extraction process. Secondly, we propose a weighted soft label (WSL) loss function. Lastly, we incorporate an MTL framework with auxiliary branches for ASC and AEC tasks, enhancing the performance of the primary SED task through joint training. In the future, we intend to explore making task weights adaptive rather than relying on hyperparameter tuning.

7. REFERENCES

[1] 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.

[2] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Computing Surveys (CSUR)*, pp. 1–46, 2016.

[3] 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.

[4] J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "SONYC: A system for monitoring, analyzing, and mitigating urban noise pollution," *Communications of the ACM*, pp. 68–77, 2019.

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

[6] 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.

[7] 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*, pp. 2880–2894, 2020.

[8] T. Khandelwal, R. K. Das, A. Koh, and E. S. Chng, "Leveraging audio-tagging assisted sound event detection using weakified strong labels and frequency dynamic convolutions," *IEEE Statistical Signal Processing Workshop*, 2023.

[9] ——, "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.

[10] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," *Interspeech*, pp. 571–575, 2021.

[11] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," 2022. [Online]. Available: https://arxiv.org/abs/2212.09058

[12] A. Mesaros, T. Heittola, and A. Klapuri, "Latent semantic analysis in sound event detection," *European Signal Processing Conference*, pp. 1307–1311, 2011.

[13] K. Imoto and S. Shimauchi, "Acoustic scene analysis based on hierarchical generative model of acoustic event sequence," *IEICE Transactions on Information and Systems*, pp. 2539–2549, 2016.

[14] K. Imoto and N. Ono, "Acoustic topic model for scene analysis with intermittently missing observations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 367–382, 2019.

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

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

[17] I. Martín-Morató, M. Harju, P. Ahokas, and A. Mesaros, "Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[18] I. Martín-Morató and A. Mesaros, "Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 902–914, 2023.

[19] T. Khandelwal and R. K. Das, "A multi-task learning framework for sound event detection using high-level acoustic characteristics of sounds," *Interspeech*, 2023.

[20] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating deep network training by reducing internal covariate shift," *International Conference on Machine Learning (ICML)*, pp. 448–456, 2015.

[21] 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.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015.

[23] J. Ebbers, R. Haeb-Umbach, and R. Serizel, "Threshold-independent evaluation of sound event detection scores," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1021–1025, 2022.

[24] Y. Jin, M. Chen, J. Shao, Y. Liu, B. Peng, and J. Chen, "DCASE 2023 challenge Task 4 technical report," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2023.

[25] X. Xuenan, M. Ziyang, Y. Fei, Y. Guanrou, W. Mengyue, and C. Xie, "Sound event detection by aggregating pretrained embeddings from different layers," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2023.

[26] H. Yin, J. Bai, S. Huang, and J. Chen, "How information on soft labels and hard labels mutually benefits sound event detection tasks," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2023.

[27] T.-D. Nhan, B. Param, and Y. Zhang, "Sound event detection with soft labels using self-attention mechanisms for global scene feature extraction," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2023.

[28] D. Min, H. Nam, and P. Yong-Hwa, "Application of spectro-temporal receptive field for DCASE 2023 challenge Task 4B," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2023.

[29] H. Zhang, L. Zuo, J. Chen, X. Cai, and M. Wu, "Sound event detection based on soft label," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2023.