TECHNICAL REPORT ON LEE SUBMISSION: SOUND EVENT DETECTION USING CONFORMER AND ATST FRAMEWORK FOR DCASE CHALLENGE 2024 TASK 4

Technical Report

Yuna Lee, JaeHoon Jung

KT Corporation, Republic of Korea

ABSTRACT

Sound Event Detection (SED) has shown promising performance in detecting and classifying meaningful events on the given audio signal input. Since the real-world scenario does not provide well-labeled data, there had been an urge to extend the research to a rather “coarse” labeled dataset. In this report, we propose a novel model to perform robustly on the well-labeled datasets and potentially missing labeled datasets using large pre-trained audio transformers throughout the training process. Our method can improve the performance to 0.52 in $PSDS1$ and 0.77 in $pAUC_M$.

Index Terms— Sound Event Detection (SED), BEATs, Audio Teacher-Student Transformer (ATST), Conformer, Convolutional recurrent neural network (CRNN)

1. INTRODUCTION

Sound event detection (SED) is a field of research that detects the timestamp of a specific event and classifies its event in an audio signal. While traditional audio classification and detection tasks, such as audio tagging, recognize what types of sound events are present in an audio stream, SED aims to pinpoint precise onset and offset of distinct sound events. This specialty of SED makes it crucial for applications requiring real-time or post-hoc analysis of acoustic environments like security or home Internet of Things (IoT) services[1].

There have been several strategies in deep learning that significantly enhanced the capabilities of SED systems. For example, self-supervised audio transformers such as BEATs [2] and Audio Teacher-Student Transformer (ATST) [3] showed that obtaining better audio representation embeddings in training can boost the performance. Likewise, several methods designed new layers to reflect temporal and spectral characteristics of audio signals, leading to remarkable proficiency [4, 5, 6]. Researchers proposed research that applies novel frameworks such as knowledge distillation, semi-supervised learning, and contrastive learning to improve performance [7, 8, 9]. New metrics were suggested to evaluate SED systems’ performance more rigorously [10, 11], and a few methods suggested novel strategies to deal with post-processing the output of SED system [12, 13]. Despite significant progress, few challenges persist in developing robust SED systems. First, audio signals in the real world consist of multiple overlapping classes, making SED model hard to identify a specific sound event. This leads to rather disappointing detection results when multiple sound events co-occur. Second, SED systems are still vulnerable to variations in the acoustic environment. While there have been efforts to implement robust SED systems by applying heavy augmentations on the training audio signals [14, 15], current SED systems are still vulnerable to the unpredictability in acoustic environments of the real world. The most important problem beyond these two problems is the scarcity of labeled data. Since the task of SED is to pinpoint the exact timestamp of particular audio events, it requires a dataset with the precisely labeled onset and offset for a large number of classes. Previous studies have used the Audioset [16], a large-scale audio dataset with manually annotated events up to 632. However, the fact that the Audioset is manually annotated still holds uncertainty. Synthetic datasets have been used to address these uncertainty issues, but the limitation still exists. To deal with the drawbacks, the

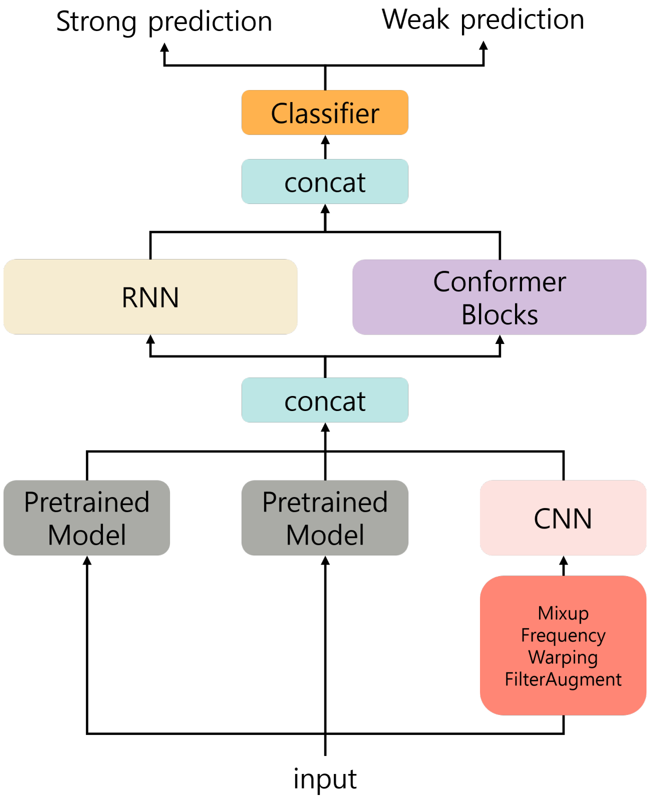

Figure 1: The figure of our proposed model CRNN-CON and overall framework. Instead of using extracted embeddings, we give raw audio signals as input to the pre-trained model. We used ATST and BEATs in our framework, but we believe that our method can be applied to any pre-trained model, hence we use the term pre-trained model in the place where ATST and BEATs exist.

following DCASE 2024 challenge [17] aims to take the SED system to another level by dealing with various datasets in training. The challenge provides a variety of datasets from different domains. In this paper, we propose a novel model architecture that integrates a typical Convolutional recurrent neural network (CRNN) with conformer [18]. We also optimize the training process for the different in-domain datasets so that the overall performance is guaranteed even when trained on the datasets from different domains. By doing so, our system will advance SED systems, thereby enhancing their utility and performance in real-world scenarios.

2. DATASET

The DCASE 2024 Challenge Task 4 development set consists of two datasets: DESED dataset and MAESTRO Real dataset [19]. DESED dataset [20], commonly used in the previous DCASE challenges, consists of 4 datasets: weakly labeled dataset, unlabeled in-domain dataset, synthetic dataset, and strongly labeled real dataset. The weakly labeled training set contains 1,578 clips of weak annotations, and the unlabeled training set consists of 14,412 clips. The synthetic set has 10,000 audio clips, while the real data set consists of 3,470 clips. On the DCASE 2023 challenge, a new soft-labeled training set, MAESTRO Real set [21], was introduced on Task4B. The MAESTRO set consists of real-world recordings lasting 3 minutes in several acoustic scenes. Since the audio was annotated with multiple annotators, it does not provide fine-label information. The DESED dataset aims to detect 10 classes of sound events such as “Alarm bell ringing”, “Speech”, and “Vaccum cleaner”. MAESTRO dataset contains 17 classes, and it shares overlapping labels such as “People talking”, “Cutlery and dishes”. MAESTRO dataset only provides information in 1-second segments, which means there is no fine-grained timestamp or labeling information inside the second segment. This makes them different from traditional SED datasets and raises questions about the applicability of conventional methods.

3. PROPOSED CRNN-CON FRAMEWORK

3.1. Model

Given that the dataset holds mass diversity, we intend to raise the models’ ability to extract necessary auditory features from signals even in label-deficient settings. In this submission, we used two types of models: an existing model and our novel model. For an existing model, we utilize the Frequency-dynamic convolution [4] model. We infer this model FDY-CRNN for easy explanation.

We use conformer encoder blocks with the CRNN model to extract all the essential features from the audio input. The architecture of our novel model is shown in the figure 1. The conformer block contains feed-forward layers, multi-head self-attention (MHSA), and convolution layers. This allows the conformer to get local information through CNN layers while extracting long-range global context from the self-attention layer. We presumed that implementing the following characteristic of conformer allows our systems to distinguish between different classes even when multiple sound events overlap. The original conformer model consists of 16 conformer encoder blocks, and implementing the whole model might lead to overfitting. Thus, we downsized the size of the conformer and rearranged the parameters through experiments. In the same way, we downsize the number of layers in the CNN network is essential as maintaining the original 7 layers of the CNN network might lead to performance degradation. We infer our model as CRNN-CON for easy explanation. A similar approach in applying the conformer model in SED tasks[22] exists, our framework differs in the way that we use both RNN module and conformer encoder blocks to use the extracted information at its utmost.

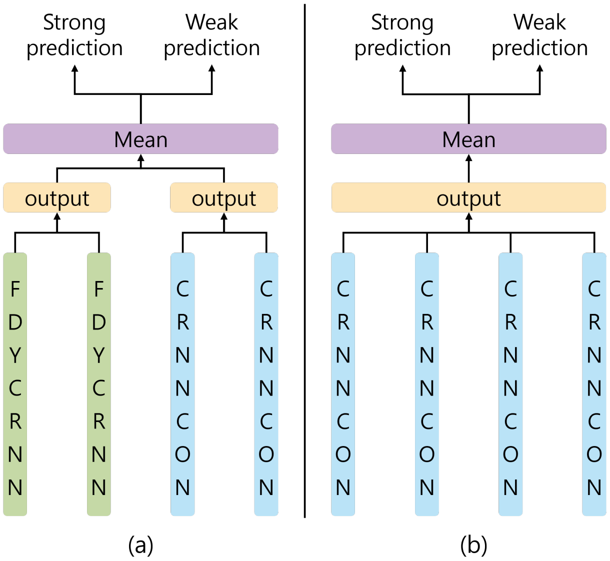

Figure 2: The figure depicts our ensemble. (a) shows our heterogeneous ensemble, which consists of two different models: FDY-CRNN and CRNN-CON. (b) describes the typical ensemble framework of the proposed model, CRNN-CON. Due to the capacity of GPU, we use 4 models for our ensemble.

3.2. Implementation Details

Figure 1 depicts our overall framework. In this challenge, we incorporate BEATs and frame-level ATST, combining baseline method [17] and state-of-the-art method [15]. We designed the framework based on the official implementation of ATST-Frame[1] and the baseline implementation[2] since the ATST method did not include MAESTRO dataset. ATST method trains the model through two stages, but we used stage 1 as the sole training process as the framework undergoes severe performance degradation after stage 2 training. The committee also provided a synthetic dataset for MAESTRO. However, we did not use the following dataset since it did not improve performance. We speculate that the amount of synthetic data in the DESED dataset is sufficient enough that additional MAESTRO datasets are unlikely to increase overall performance. For CRNN-CON, we stack only 4 layers of conformed blocks and rearrange the kernel size of MHSA layers and dimensions. We also prune the size of the CNN network from 7 layers to 4 layers. Similarly, we propose two types of ensemble framework: heterogeneous ensemble and typical ensemble. We configured FDY-CRNN and CRNN-CON in a 1:1 ratio in the heterogeneous ensemble. With this design, we speculated that the ensemble could perform robustly on both fine-grained and coarse-labeled datasets, thereby achieving high scores on both $PSDS1$ and $pAUC_M$ instead of excelling in

[1] https://github.com/Audio-WestlakeU/ATST-SED
[2] https://github.com/DCASE-REPO/DESED_task

methods	models	pre-trained model	postprocessing	augmentation	$PSDS1$	$pAUC_M$	total metric	submission
methods	baseline [17]	BEATs	median filter	-	0.485	0.643	1.128	-
	FDY-CRNN [4]	BEATs	median filter	-	0.484	0.677	1.161	-
	CRNN-CON	BEATs	median filter	-	0.473	0.686	1.159	-
	CRNN-CON	ATST + BEATs	median filter	True	0.481	**0.763**	1.244	-
	CRNN-CON	ATST + BEATs	SEBB	True	**0.502**	**0.760**	**1.262**	submission 1
	FDY-CRNN [4]	ATST + BEATs	median filter	True	**0.507**	**0.734**	**1.241**	submission 2
	FDY-CRNN [4]	ATST + BEATs	SEBB	True	0.484	0.711	1.195	-
ensemble	FDY-CRNN + CRNN-CON	ATST + BEATs	median filter	True	**0.517**	**0.760**	**1.277**	submission 3
	FDY-CRNN [4] + CRNN-CON	ATST + BEATs	SEBB	True	0.493	0.758	1.251	-
	CRNN-CON	ATST + BEATs	median filter	True	0.489	0.765	1.254	-
	CRNN-CON	ATST + BEATs	SEBB	True	**0.501**	**0.762**	**1.263**	submission 4
	FDY-CRNN [4] + CRNN-CON	BEATs	median filter	True	0.512	0.701	1.213	-

Table 1: The table shows $PSDS1$ and $pAUC_M$ metric of baseline and models on the validation set. We submitted 4 systems in the DCASE 2024 challenge. The performance of each submission is emphasized in bold.

one metric. For an ordinary ensemble, we tried both models on the sole and chose the better one. In building an ensemble network, we combined 5 FDY-CRNN and 5 CRNN-CON models to build the heterogeneous model with pre-rained BEATs. For the ensemble model with ATST and BEATs, we used only 2 FDY-CRNN and 2 CRNN-CON due to the capacity of the GPU. The following architecture is specified in Figure 2.

3.3. Training

For training, we extract the mel spectrogram with 2048 FFT size, hop length of 256, and 128 mel bins from the raw waveform. For pre-trained ATST, we extract the mel spectrogram with 1024 FFT size, hop length of 160, and 64 mel bins. The dropout method was applied to the model at a rate of 0.5. We trained for 200 epochs. One of our baseline [15] gave different learning rates in different layers of the network, but we unified the learning rate to 0.001. We define f as our SED model and θ as our model parameter. Training dataset $\mathcal{D}$ can be defined as below.

$$\mathcal{D} = \{(x^1, y^1), (x^2, y^2), \cdots, (x^N, y^N)\} \tag{1}$$

where x^i denotes the acoustic feature gained from i-th audio clip, and y^i corresponds to the ground truth label. Each y^i can also be explained as $y^i = \{y_1^i, \cdots, y_T^i\}$ where $y_1^i \in \{0,1\}^K$ for K classes and T time frames. By this notations, We can define $\hat{y}$, which is the prediction of our model f and have the probability of K classes on T time frames.

$$\hat{y}^i = f(x^i, \theta) \in \{0,1\}^{K \times T} \tag{2}$$

As we use the binary cross entropy (BCE) loss function, loss function $\mathcal{L}_{\mathcal{BCE}}$ can be described below.

$$\mathcal{L}_{bce} = -\sum_{t,k=1}^{T,K} y_{t,k} log(\hat{y}_{t,k}) + (1 - y_{t,k}) log(1 - \hat{y}_{t,k}) \tag{3}$$

Since we adopt a teacher-student scheme as our baseline, we use the mean squared error (MSE) loss function between student and teacher models. Given that we defined θ as our model parameter, We can define θ' as the parameter of the teacher model. For the batch size $\mathcal{B}$, loss function $\mathcal{L}_{\mathcal{MT}}$ can be described below.

$$\mathcal{L}_{\mathcal{MT}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} ||f(x^i, \theta') - f(x^i, \theta)|| \tag{4}$$

With equation 3 and equation 4, the total loss function $\mathcal{L}$ can be simplified as below.

$$\mathcal{L} = \mathcal{L}_{\mathcal{BCE}} + \mathcal{L}_{\mathcal{MT}} \tag{5}$$

For data augmentation, We applied filteraugment [23], mixup [14] and frequency warping [15] methods for data augmentation on 50% of probability. We did not use augmentation regarding the time-axis which might lead to critical information loss. We avoid applying data augmentations to pre-trained model input. We discuss the effect of data augmentation on the pre-trained model in depth in the following section 4 and section 5.

3.4. Postprocessing

We used median filter and Sound Event Bounding Boxes (SEBB) [24] to boost the performance of our baseline system afterward. While conducting the experiments, we realized that any framework with FDY-CRNN tends to show a performance degradation in $PSDS1$ up to 0.02 when SEBB is applied. On the other hand, it showed promising improvements in the CRNN-CON framework. Thus, we used a median filter to Submission 2 and Submission 3 as it have an FDY-CRNN network. We applied SEBB in Submission 1 and Submission 4, which contain only CRNN-CON models. We compare and analyze the effect of SEBB further in section 4 and section 5.

4. EXPERIMENTS

Our experiments can be divided into single-model systems and ensemble systems. First of all, we conducted experiments on the original baseline, FDY-CRNN, and CRNN-CON models under the same conditions to compare the performance between models. Then we compare FDY-CRNN and CRNN-CON with different pre-trained models: BEATs and BEATs+ATST to compare the effect of ATST on our framework. For ensemble systems, we compare the heterogeneous ensemble and typical ensemble to compare the efficacy of mixing different models. For the ablation study, we conducted experiments to see how applying augmentations to a pre-trained model embedding affects the performance of the SED framework.

As the challenge deals with so-called “fine-grained” labeled datasets and “coarse” labeled datasets, different metrics for each dataset are required for an objective evaluation. Therefore, we use two metrics: polyphonic sound detection scores on scenario 1 ($PSDS1$) and $pAUC_M$. $PSDS1$ were applied to the DESED

models	filteraugment [23]	frequency warping [15]	$PSDS1$	$pAUC_M$
CRNN-CON	✓		0.470	0.672
		✓	0.384	0.654
	✓	✓	0.381	0.673
FDY-CRNN	✓		0.449	0.683
		✓	0.430	0.671
	✓	✓	0.421	0.639

Table 2: Performance comparison between models with different augmentation on the pre-trained model. Each model already has the two given augmentations applied to it. All model uses BEATs as pretrained model.

dataset, and $pAUC_M$ to the MAESTRO dataset [25]. This way, we can see objective results without the different characteristics of the datasets being influenced by each other.

5. RESULTS

The performance of our proposed methods and ensembles are shown in Table 1. The first three rows of the table compare the performance of the baseline model with FDY-CRNN and CRNN-CON under the same conditions. It shows that two models can outperform the CRNN model by a small margin. The results show that the combination of BEATs and ATST plays a huge role in the model's performance. We speculated that this result comes from the differentiation between BEATs and ATST-Frame. Both BEATs and ATST-Frame have in common that they are models trained with self-supervised learning. However, ATST-Frame is better than BEATs at extracting frame-wise auditory representations since it is trained on a frame-by-frame basis [15]. BEATs can obtain information about the global representation [2]. This has the advantage that the long-term representation between frames can be retained. Therefore, we can assume that these advantages create synergy, increasing the performance to 1.262.

When we compare the performance between single models, we find the possibility of a heterogeneous ensemble. we can see that the $pAUC_M$ shows drastic change while the $PSDS1$ metric is stable up to 0.50 with FDY-CRNN as a baseline model. By these results, we can infer that Frequency dynamic convolution may be robust in DESED dataset. In the case of CRNN-CON, CRNN-CON can achieve high performance on MAESTRO dataset as $pAUC_M$ of CRNN-CON baseline shows up to 0.763. With these results, we can infer that the conformer may show robust performance even in the "coarse" labeled dataset. Based on these speculations, Wwe thought that if we ensemble these two models in the right proportions, we could get good performance for both $pAUC_M$ and $PSDS1$. The performance of heterogeneous ensembles with BEATs achieves up to 0.501 in $PSDS1$ and 0.701 in $pAUC_M$, almost equivalent to training a single model with two large pre-trained models. These results demonstrate ensembles can deliver performance that single models cannot. For the ensemble using BEATs and ATST together, the increase in performance is slightly less than for BEATs, but the fact that $PSDS1$ and $pAUC_M$ are close to the best performance of each model is notable.

In table 5, We also conducted experiments to verify the effect of data augmentations on the pre-trained model input as an ablation study. When we compare the performance of CRNN-CON in table 1, $PSDS1$ and $pAUC_M$ showed performance degradation. The dramatic performance drop in $PSDS1$ indicates that applying data augmentation on pre-trained models negatively impacts the CRNN-CON framework. Similarly, applying augmentation on the FDY-CRNN framework also results in performance degradation on both metrics. Considering that the performance drop is smaller than CRNN-CON, we can infer that the role of frequency dynamic convolution and the depth of convolution layers in CNN may play a part in the robust performance in SED.

6. CONCLUSION

In this report, we proposed our novel CRNN-Con and framework using two pretrained model BEATs and ATST-Frame. Our novel model and framework shows that we can preserve the $PSDS1$ performance of the CRNN model while boosting the $pAUC_M$ metric on the new soft-labeled dataset. Further, we conducted an experiment to verify the effect of applying data augmentation to pertrained model. We tend to expand our framework for the future research.

7. REFERENCES

[1] A. H. Yuh and S. J. Kang, "Real-time sound event classification for human activity of daily living using deep neural network," in *2021 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing Communications (GreenCom) and IEEE Cyber, Physical Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*, 2021, pp. 83–88.

[2] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 5178–5193. [Online]. Available: https://proceedings.mlr.press/v202/chen23ag.html

[3] X. Li, N. Shao, and X. Li, "Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 32, p. 1336–1351, jan 2024. [Online]. Available: https://doi.org/10.1109/TASLP.2024.3352248

[4] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, "Frequency Dynamic Convolution: Frequency-Adaptive Pattern Recognition for Sound Event Detection," in *Proc. Interspeech 2022*, 2022, pp. 2763–2767.

[5] H. Nam, S.-H. Kim, D. Min, J. Lee, and Y.-H. Park, "Diversifying and expanding frequency-adaptive convolution kernels for sound event detection," 2024.

[6] D. Min, H. Nam, and Y.-H. Park, "Auditory neural response inspired sound event detection based on spectro-temporal receptive field," *arXiv preprint arXiv:2306.11427*, 2023.

[7] Y. Xiao and R. K. Das, "Dual knowledge distillation for efficient sound event detection," *arXiv preprint arXiv:2402.02781*, 2024.

[8] 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.

[9] Y. Guan, J. Han, H. Song, W. Song, G. Zheng, T. Zheng, and Y. He, "Contrastive loss based frame-wise feature disentanglement for polyphonic sound event detection," *arXiv preprint arXiv:2401.05850*, 2024.

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

[11] J. "Ebbers, R. Haeb-Umbach, and R. Serizel, "Post-processing independent evaluation of sound event detection systems," in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, Tampere, Finland, September 2023, pp. 36–40.

[12] L. Cances, T. Pellegrini, and P. Guyot, "Multi task learning and post processing optimization for sound event detection," *IRIT, Universit de Toulouse, CNRS, Toulouse, France, Tech. Rep*, 2019.

[13] P. Giannakopoulos, A. Pikrakis, and Y. Cotronis, "Improving post-processing of audio event detectors using reinforcement learning," *IEEE Access*, vol. 10, pp. 84 398–84 404, 2022.

[14] Y. N. D. D. L.-P. Hongyi Zhang, Moustapha Cisse, "mixup: Beyond empirical risk minimization," *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[15] N. Shao, X. Li, and X. Li, "Fine-tune the pretrained atst model for sound event detection," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 911–915.

[16] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[17] S. Cornell, J. Ebbers, C. Douwes, I. Martín-Morató, M. Harju, A. Mesaros, and R. Serizel, "Dcase 2024 task 4: Sound event detection with heterogeneous data and missing labels," 2024.

[18] A. Gulati, C.-C. Chiu, J. Qin, J. Yu, N. Parmar, R. Pang, S. Wang, W. Han, Y. Wu, Y. Zhang, and Z. Zhang, Eds., *Conformer: Convolution-augmented Transformer for Speech Recognition*, 2020.

[19] http://dcase.community/challenge2024/.

[20] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, New York City, United States, October 2019. [Online]. Available: https://hal.inria.fr/hal-02160855

[21] 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*, vol. 31, pp. 902–914, 2023.

[22] 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," *dim*, vol. 1, no. 4, 2020.

[23] H. Nam, S.-H. Kim, and Y.-H. Park, "Filteraugment: An acoustic environmental data augmentation method," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4308–4312.

[24] J. Ebbers, F. G. Germain, G. Wichern, and J. L. Roux, "Sound event bounding boxes," 2024.

[25] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.