LABEL FILTERING-BASED SELF-LEARNING FOR SOUND EVENT DETECTION USING FREQUENCY DYNAMIC CONVOLUTION WITH LARGE KERNEL ATTENTION

Ji Won Kim[1], Sang Won Son[1], Yoonah Song[1], Hong Kook Kim[1,2,]

Il Hoon Song[3], Jeong Eun Lim[3]

[1] AI Graduate School, [2] School of EECS
Gwangju Institute of Science and Technology
Gwangju 61005, Korea
{{jiwon.kim, ssw970519, yyaass0531}@gm., hongkook@}gist.ac.kr

[3]AI Lab.
Hanwha Vision
Seongnam-si, Gyeonggi-do 13488, Korea
{ilhoon, je04.lim}@hanwha.com

ABSTRACT

This paper proposes a convolutional recurrent neural network (CRNN)-based sound event detection (SED) model. The proposed model utilizes frequency dynamic convolution (FDY) with a large kernel attention (LKA) for convolution operations within the CRNN. This is designed to effectively capture time-frequency patterns and long-term dependencies for non-stationary audio events. In addition, we concatenate a pre-trained bidirectional encoder representation from audio transformers (BEATs) embedding with the output of FDY–LKA. This provides the FDY-based feature maps with semantic information. Given the limited labeled data condition of the DCASE Challenge dataset, we first employ the mean-teacher-based semi-supervised learning. Then, we propose label filtering-based self-learning for audio event data selection, when their pseudo labels predicted from the mean-teacher model are strong correlated with given weakly labels. This strategy applies weakly labeled and unlabeled data, and then extends to the AudioSet. We evaluate its performance of the proposed SED model on DCASE 2023 Challenge Task 4A, measuring the F1-score and polyphonic sound detection scores, namely PSDS1 and PSDS2. The results indicate that the proposed CRNN-based model with FDY–LKA improves the F1-score, PSDS1, and PSDS2 in comparison to the baseline for DCASE 2023 Challenge Task 4A. When we apply the BEATs embedding via average pooling to both the baseline and the proposed model, we find that the performance of the proposed model significantly outperforms the baseline, with an F1-score of 6.2%, a PSDS1 score of 0.055, and a PSDS2 score of 0.021. Consequently, our model is ranked first in the DCASE 2023 Challenge Task 4A evaluation for a single model track, and second for an ensemble model.

Index Terms—Sound event detection, semi-supervised learning, label filtering-based self-learning, frequency dynamic convolution, large kernel attention, BEATs embedding

1. INTRODUCTION

The objective of sound event detection (SED) is to recognize and classify individual sound events originating from acoustic signals, along with their corresponding time stamps. The potential applications of the SED model have been attracted from audio captioning [1] to various domains, such as wildlife tracking [2], equipment monitoring [3], and medical monitoring [4]. In recent years, SED has been extensively researched using deep learning models [5]. However, a significant challenge in using deep learning for SED is the requirement of strong labels, which are expensive and time-consuming. This problem has led to develop weakly supervised or semi-supervised learning techniques to mitigate such label requirement.

To address this problem, we apply a self-learning strategy based on label filtering to train the proposed SED model when the quantity of labeled training data is limited. The proposed model is based on a convolutional recurrent neural network (CRNN), where the convolution is realized with frequency dynamic convolution (FDY) [6] with large kernel attention (LKA) [7].

As a remedy for limited resources, we use select data from the AudioSet [8] as additional training material. In this context, the audio class of each data item from AudioSet is mapped into that of the DCASE Challenge Task 4A and data belonging to the DCASE audio class are selected. However, even though this approach of using additional AudioSet data improves SED performance [9], it leads to a data imbalance issue. Furthermore, this method tends to include audio data whose characteristics differ from those in the DCASE training set. Thus, we propose an alternative in the form of a label filtering-based self-learning method to select appropriate data from AudioSet by examining the inference probability during model training.

Next, one of the most successful components in detection models is the application of an attention mechanism, which emphasizes semantic knowledge in the feature map. Of late, there have been several types of attention mechanisms, like squeeze-and-excitation (SE) [10] and convolutional block attention module (CBAM) [11], which are designed to accommodate channel and/or spatial information for attention. These mechanisms alter or reshape an image to obtain attention weights, given that images are shift-invariant for classification or detection. However, the spectrogram image of an audio event signal is neither shift-invariant nor stationary, necessitating an attention mechanism with unaltered attention weights.

Inspired by image classification and detection [7], we incorporate LKA into the sound event detection model. Combining this enables us to maintain long-term dependency for the attention, even when the audio signals are non-stationary. To the best of our

* This work was supported in part by Hanhwa Vision Co. Ltd., and by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2022-0-00963).

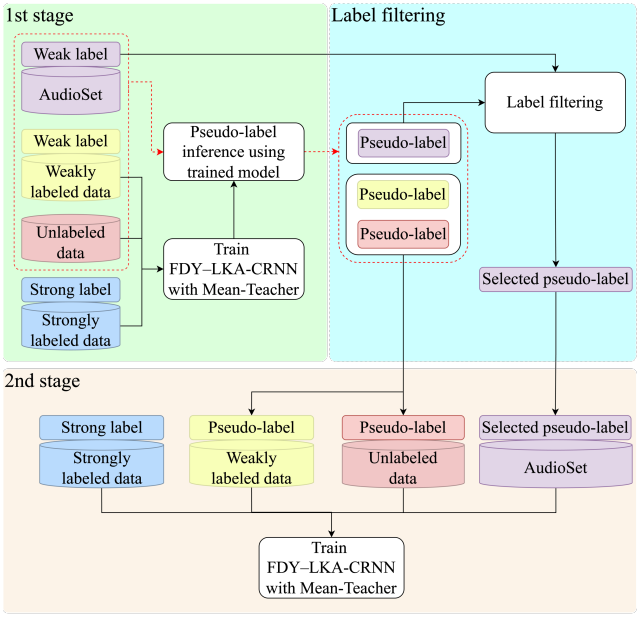

Figure 1: Illustration of the training procedure for the proposed FDY–LKA-CRNN-based SED model, including a label filtering-based self-learning strategy.

knowledge, we are the first to apply this to sound event detection tasks.

Our contributions can be summarized as follows:

- We have developed a label-filtering approach to supplement training data from weakly-labeled out-of-domain sources, such as AudioSet, within a self-learning-based model training framework. As a result, we have improved the detection accuracy of sound event detection.
- Additionally, we have integrated an attention mechanism, the large kernel attention (LKA), proposed for image classification and detection, into the sound event detection model. This is vital as audio signals are inherently non-stationary, necessitating the retention of long-term dependency for the attention.
- We have applied our proposed training strategy to designing SED models for the DCASE 2023 Challenge Task 4A and achieved the best performance in terms of F1-score and PSDSs [12] without an ensemble. Moreover, our ensemble model ranked second.

Following this introduction, Section 2 describes the dataset and input features of the SED model. Section 3 proposes a label filtering-based self-learning strategy applied to the FDY–LKA-CRNN model. Next, Section 4 evaluates the performance of the proposed SED models on the validation dataset task of DCASE 2023 Task 4A. Finally, Section 5 concludes this paper.

2. DATASET

The DCASE 2023 Challenge Task 4A consists of four datasets: weakly labeled data, unlabeled in-domain training data, strongly labeled synthetic data, and strongly labeled real data. All the audio clip data span 10 seconds each. The strongly labeled synthetic dataset is unique in that it is generated by Scraper [13]. The weakly labeled dataset only has class labels and is annotated for 1,578 clips. The unlabeled in-domain training dataset includes 14,412 audio clips. Meanwhile, the real strongly labeled and synthetic datasets comprise 3,470 and 10,000 clips, respectively. In addition to the DCASE dataset, we utilize a subset of AudioSet that includes 18,000 clips with in-domain weak labels.

The following preprocessing steps are employed to prepare the data for input to the model: First, the mono-channel signals are resampled from 44.1 to 16 kHz. Subsequently, the audio signals are divided into frames of 2,048 samples each, with a hop length of 160 samples. Each frame first undergoes a 2,048-point fast Fourier transform (FFT), followed by a 128-dimensional mel-filterbank analysis. This results in input feature dimensions of (1001x128). The extracted mel-spectrogram features are then normalized using the mean and standard deviation of all the training audio samples.

3. PROPOSED FDY–LKA-CRNN-BASED SED MODEL

Fig. 1 illustrates the training procedure of the proposed FDY–LKA-CRNN-based SED model, which employs a label filtering-based self-learning strategy. As depicted in the upper-left arm of the figure, an SED model is initially trained using the mean-teacher approach, where the entire DCASE Challenge Task 4A dataset is utilized. For a detailed procedure of this first-stage training, please refer to the training description in [14]. Subsequently, label filtering is carried out to select audio event data from AudioSet for the training of the second-stage SED model. This selection process is designed to choose audio event data for which the pseudo-labels, predicted from the first-stage SED model, strongly correlate with the weak labels provided by the AudioSet data descriptors. Finally, the second-stage SED model as shown in the lower arm of Fig. 1 is retrained using both the entire DCASE challenge data and the selected AudioSet data.

The following subsections explains the network architecture of the proposed LKA-CRNN-based SED model, LKA-based attention, and label filtering-based self-learning.

3.1. Network architecture

Table 1 displays the network architecture of this proposed model. The model comprises one stem block, six FDY–LKA blocks, one optional fusion block, and one RNN block. Initially, all input features for each audio clip are grouped to form a spectral image of dimensions (1001×128×1), which serves as the input to the stem block. In detail, the stem block consists of one convolutional block with 32 kernels of size (3×3) and a stride of (1×1), which is further processed by batch normalization (BN), gated linear unit (GLU) activation, and a 2×2 average pooling layer. Note that (x×y×z) and (x×y) indicate (frame×frequency×channel) and (frame×channel), respectively.

Next, the output from the stem block is processed by the first FDY–LKA block. This block is made up of FDY, LKA, BN, GLU, and an average pooling layer, as indicated in the table. The output of each FDY–LKA block is then passed to the next FDY–LKA block. Consequently, the output from the last FDY–LKA block, which is also the output of FDY–LKA-CNN, becomes a feature map with a dimension of (250×1×256).

In the fusion block, we optionally use the bidirectional encoder representation from audio transformers (BEATs) encoder [15] which is pretrained with AudioSet. The BEATs encoder ex-

Table 1. Network architecture of the proposed FDY–LKA-CRNN-based SED model, where the Fusion Block is optionally performed when BEATs embedding is applied.

Name	Layers	Output shape
Input Layer	Input: log-mel spectrogram	1001×128×1
Stem Block	3x3, Conv2D, @32 GLU, BN 2x2 average pooling layer	500×64×32
FDY–LKA Blocks	(FDY(K=4), @64, GLU, BN LKA 2x2 average pooling layer)	250×32×64
	(FDY(K=4), @128, GLU, BN LKA 1x2 average pooling layer)	250×16×128
	(FDY(K=4), @256, GLU, BN LKA 1x2 average pooling layer)	250×8×256
	(FDY(K=4), @256, GLU, BN LKA 1x2 average pooling layer)	250×4×256
	(FDY(K=4), @256, GLU, BN LKA 1x2 average pooling layer)	250×2×256
	(FDY(K=4), @256, GLU, BN LKA 1x2 average pooling layer)	250×1×256
Fusion Block (optional)	Average pooling or interpolation on BEATs embedding	250×768
	Channel-wise Concatenation (Output of FDY–LKA blocks (250×256) BEATs embedding (250×768))	250×1024
	Fully connected layer (1024×256)	250×256
RNN Block	(256 Bi-GRU cells) x 2	250×512

tracts the embedding corresponding to high-level semantic information. To align the dimensions between the output of the FDY–LKA-CNN and the BEATs embedding, we employ either average pooling or nearest neighbor interpolation. This results in four distinct models, constructed by applying one of these methods at the first or second stage for model diversity. The aligned BEATs embedding is then concatenated with the output of the FDY–LKA-CNN, followed by a fully connected (FC) layer to produce a feature map with a dimension of (250×256).

Finally, this feature map is processed by the RNN block, which comprises two bidirectional gated recurrent units (Bi-GRUs) designed to learn temporal context information. To perform SED, the output from the RNN block is processed by an FC layer and then a sigmoid function, generating an output with a dimension of (250×10), where 10 indicates the number of sound events to be detected.

The following subsection provides detailed explanations of our contributions, such as the LKA-based attention and label-filtering-based self-learning strategy, which are two key factors in achieving state-of-the-art SED performance.

3.2. LKA-based attention

The FDY, in each FDY–LKA block, is designed to capture the specific frequency characteristics associated with each event class category in the DCASE challenge. However, it is not enough to only use FDY; we also need to represent the long-term dependency of audio signals. Audio signals are inherently non-stationary, which means that we need to apply LKA-based attention, as illustrated in the FDY–LKA block in Table 1. Originally, LKA was proposed for image classification and detection tasks [7] to assign attention to a pixel by considering its adjacent pixels. In this paper, we interpret the spectrogram of an audio event sound as an image. Therefore, the attention for a specific time-frequency bin should be assigned by taking into account its adjacent time-frequency bins or bands.

The LKA attention mechanism comprises three distinct convolution layers: a depth-wise convolution layer, a depth-wise dilation convolution layer, and a (1x1) convolutional layer. The depth-wise convolution layer utilizes the local time-frequency information derived from the feature map procured by FDY. Following this, the depth-wise dilation convolution layer extracts essential long-range time-frequency band information. The final convolutional layer focuses on a channel that represents audio events as the functionality of the attention mechanism.

3.3. Label filtering-based self-learning

We propose a label filtering method to address the scarcity of strongly labeled data provided by the DCASE challenge. First, we prepare the data for label filtering, which includes 1) all the weakly labeled and unlabeled data from the DCASE dataset, and 2) a segment of AudioSet data that corresponds to one of the DCASE audio classes. We then use the first-stage SED model to infer these data and obtain the class prediction probabilities.

Next, we generate a strong pseudo-label, l_c^F, of the c-th class at the F-th frame for a given audio data using the following equation:

$$l_c^F = \begin{cases} 1, \text{ if } (p_c^F > \alpha) \text{ and } (p_C > \beta) \\ 0, \text{ otherwise,} \end{cases} \quad \text{for all } c \qquad (1)$$

where p_c^F represents the probability of the c-th class at the F-th frame of the audio signal for the strong pseudo-label, and p_c represents the probability of the c-th class for the weak pseudo-label. If p_c^F and p_c exceed the given thresholds, α and β, respectively, then the strong pseudo-label is assigned as class c. If (1) is not met, the audio data is discarded. Note that we set α and β to 0.5 and 0.7, respectively, from the exhaustive search.

After completing the label filtering process, all audio data with strong pseudo-labels are utilized as the second-stage training data. Here, the strongly labeled data from the DCASE dataset is also incorporated in the second stage.

4. PERFORMANCE EVALUATION

4.1. Model training

In the first training stage, the FDY–LKA-CRNN-based SED model parameters were initialized using the Xavier initialization [16]. The Adam optimization technique [17] was employed with a dropout rate [18] of 0.5. The learning rate was determined according to the ramp-up strategy [19], with the maximum learning rate reaching 0.001 after 50 epochs. Various augmentation techniques were applied to the training data, including time-frequency shift [20], time mask [21], mix-up [22], and filter augmentation

Table 2: Performance comparison of the baseline and different versions of the proposed SED models on the validation and evaluation dataset of the DCASE 2023 Challenge Task 4A.

Model	AudioSet	BEATs embedding	Ensemble	Validation dataset			Evaluation dataset		
				F1-score (%)	PSDS1	PSDS2	F1-score (%)	PSDS1	PSDS2
Baseline [25]	–	–	–	40.7	0.359	0.562	37.7	0.327	0.538
	√	Average pooling	–	57.6	0.491	0.787	56.7	0.510	0.798
Wenxin-TJU [26]	√	√	–	–	0.512	0.808	58.2	0.546	0.831
FDY–LKA-CRNN (Stage 1)	–	–	–	58.3	0.471	0.715	54.5	0.459	0.701
	–	Interpolation	–	63.3	0.527	0.782	–	–	–
	–	Average pooling	–	62.9	0.525	0.776	61.2	0.576	0.809
FDY–LKA-CRNN (Stage 2)	√	Interpolation	–	63.4	0.543	0.806	63.8	0.581	0.835
	√	Average pooling	–	63.8	0.546	0.808	64.6	0.591	0.831
FDY–LKA-CRNN (Stages 1 & 2)	√	Both	√	65.6	0.567	0.815	65.5	0.611	0.846

[23]. In the second stage, all training hyperparameters were set identically to those in the first stage.

4.2. Experimental results

The performance of the proposed SED model was evaluated using the measures defined in the DCASE 2023 Challenge Task 4A [24]: an event-based F1-score and PSDSs. Table 2 compares the performance between the baseline and various versions of the proposed SED models on the validation and evaluation datasets of the DCASE 2023 Challenge Task 4A. The performance on the validation dataset was drawn from the results released by DCASE 2023 Challenge Task 4A [25]. Note that there are blanks in the performance on the evaluation dataset for the first-stage SED model with interpolation since we did not submit this version to the DCASE challenge. Additionally, all the numbers in the table were averaged over three evaluations for each model, according to the DCASE challenge guideline.

We first compared the performance of our proposed model trained in the first stage with the baseline; both models were trained with the DCASE 2023 Challenge dataset without BEATs embeddings. As shown in the first and fourth rows of the table, the proposed FDY–LKA-CRNN-based SED model achieved a higher F1-score, PSDS1, and PSDS2 by 17.6%, 0.112, and 0.153, respectively, than the baseline. Upon applying BEATs embedding in the form of either interpolation or average pooling to the first-stage SED model, we observed increased F1-score, PSDS1, and PSDS2, compared to the first-stage model without BEATs embedding. The superior performance of the first-stage SED model over the baseline can be attributed to the contribution of FDY–LKA to the representation learning for this sound event detection task.

Second, we examined the effectiveness of expanding the training data from AudioSet on the SED performance. From the second and eighth rows in the table, it is clear that the addition of AudioSet data via the proposed label filtering significantly improved the SED performance. Specifically, the second-stage SED model with average pooling provided higher F1-score, PSDS1, and PSDS2 by 6.2%, 0.055, and 0.021, respectively, than the baseline with average pooling. Moreover, the second-stage SED model outperformed the first-stage SED model, indicating that label filtering is an efficient method for expanding training data.

Next, we constructed an ensemble model by combining 24 different models from each of the first- and second-stage SED models, which were taken according to different training epochs. This ensemble outperformed the baseline and individual stage models, due to inherent benefits of ensemble modeling such as reducing overfitting and improving model robustness.

Lastly, we compared our results with those of the Wenxin-TJU system [26] that was ranked the third place in the single model system track of DCASE 2023 Challenge Task 4A. As shown in the third and eighth rows of the table, the second stage of the proposed FDY–LKA-CRNN model provided higher PSDS1 for both the validation and evaluation dataset than Wenxin-TJU system, while two models had similar PSDS2.

5. CONCLUSION

We proposed an FDY–LKA-CRNN-based SED model with BEATs embedding for sound event detection. To achieve state-of-the-art performance in the DCASE 2023 Challenge Task 4A, we integrated the LKA-based attention to capture long-term dependency within the convolutional architecture. Additionally, we proposed a label filtering approach to select data from another public domain dataset—AudioSet. Accordingly, we developed a two-stage model training approach; the first-stage model was trained using DCASE 2023 Challenge data, while the second-stage model was trained using both DCASE 2023 Challenge data and selected AudioSet data. Finally, we constructed several versions of SED models based on the first- or second-stage training and their ensemble, which included models constructed by BEATs embedding using two different methods—interpolation and average pooling.

Various versions of the proposed FDY–LKA-CRNN-based SED models were evaluated on the validation dataset for DCASE 2023 Task 4A, and their performance was compared with the baseline. The results revealed that the proposed second-stage SED model, featuring LKA-based attention and label filtering-based data selection, significantly improved the SED performance compared to the baseline and the first-stage SED models. Moreover, an ensemble model consisting of the first- and second-stage models outperformed other versions of the proposed models.

6. REFERENCES

[1] K. Drossos, S. Adavanne, and T. Virtanen, "Automated audio captioning with recurrent neural networks," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA),* 2017, pp. 374–378.

[2] Z. Zhao, S.-H. Zhang, Z.-Y. Xu, K. Bellisario, N.-H. Dai, H. Omrani, and B. C. Pijanowski, "Automated bird acoustic event detection and robust species classification," *Ecological Informatics*, vol. 39, pp. 99–108, 2017.

[3] S. Grollmisch, J. Abeßer, J. Liebetrau, and H. Lukashevich, "Sounding industry: Challenges and datasets for industrial sound analysis," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[4] N. C. Phuong and T. D. Dat, "Sound classification for event detection: Application into medical telemonitoring," in *Proc. International Conference on Computing, Management and Telecommunications (ComManTel)*, 2013, pp. 330–333.

[5] T. K. Chan and C. S. Chin, "A comprehensive review of polyphonic sound event detection," *IEEE Access*, vol. 8, pp. 10333–103373, 2020.

[6] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, "Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection," *arXiv preprint*, arXiv:2203.15296, 2022.

[7] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," *arXiv preprint*, arXiv:2202.09741, 2022.

[8] 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 *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[9] S. Suh and D. Y. Lee, "Data engineering for noisy student model in sound event detection," *Tech. Rep. in DCASE 2022 Challenge,* 2022.

[10] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[11] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 3–9.

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

[13] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 344–348.

[14] N. K. Kim and H. K. Kim, "Polyphonic sound event detection based on residual convolutional recurrent neural network with semi-supervised loss function," *IEEE Access*, vol. 9, pp. 7564–7575, 2021.

[15] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," *arXiv preprint*, arXiv:2212.09058, 2022.

[16] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010, pp. 249–256.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint*, arXiv:1412.6980, 2014.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

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

[20] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for DCASE 2019 Task 4," *Tech. Rep. in DCASE 2019 Challenge*, 2019.

[21] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint*, arXiv:1904.08779, 2019.

[22] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," *arXiv preprint*, arXiv:1710.09412, 2017.

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

[24] https://dcase.community/challenge2023/task-sound-event-detection-with-weak-labels-and-synthetic-soundscapes.

[25] https://dcase.community/challenge2023/task-sound-event-detection-with-weak-labels-and-synthetic-soundscapes-results.

[26] W. Duo, X. Fang, and J. Li, "Semi-supervised sound event detection system for DCASE 2023 task4A," *Tech. Rep. in DCASE 2023 Challenge*, 2023.