PEER COLLABORATIVE LEARNING FOR POLYPHONIC SOUND EVENT DETECTION

Hayato Endo[1] *and Hiromitsu Nishizaki*[1]

[1] Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences,
University of Yamanashi, 4-3-11 Takeda, Kofu, Japan
endohayato@alps-lab.org, hnishi@yamanashi.ac.jp

ABSTRACT

This paper describes how semi-supervised learning, called peer collaborative learning (PCL), can be applied to the polyphonic sound event detection (PSED) task, which is one of the tasks in the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge. Many deep learning models have been studied to determine what kind of sound events occur where and for how long in a given audio clip. The characteristic of PCL used in this paper is the combination of ensemble-based knowledge distillation into sub-networks and student-teacher model-based knowledge distillation, which can train a robust PSED model from a small amount of strongly labeled data, weakly labeled data, and a large amount of unlabeled data. We evaluated the proposed PCL model using the DCASE 2019 Task 4 dataset, and achieved an F1-score improvement of about 8.2 points compared with the baseline model.

Index Terms— ensemble training, knowledge distillation, semi-supervised training, sound event detection, student-teacher model

1. INTRODUCTION

At present, speech recognition tasks have occupied a leading position in acoustic signal recognition, but it is unavoidable that various environmental sound processing, including speech, will become the acoustic information processing of the next era. Recently, the Detection and Classification of Acoustic Scenes and Events (DCASE) community [1] was established, where various research tasks (challenges) related to environmental sound processing have been proposed, and researchers around the world are studying them.

In this paper, we focus on the research of "Task 4: sound event detection in domestic environments" [2], proposed in DCASE 2019 [3] and DCASE 2020 [4], which is the task of identifying where a certain sound is occurring in an audio clip. When we focus on a certain moment in an audio clip, we have to consider all the sound events that may occur at the same time. This kind of task is called polyphonic sound event detection (PSED) [5]. Unlike monophonic sound event detection (MSED), which assumes that only one type of sound event occurs at a given moment, the PSED task is more difficult from various perspectives. For example, from a technical point of view, we have to estimate multiple sound types simultaneously, which is more difficult than MSED. The DCASE Task 4 is one of the PSED tasks.

To solve this PSED task using machine learning, such as deep learning, a large amount of supervised labeled data is required to train a machine learning model. In this case, the supervised label is the data with the exact sound event tag of what kind of sound event occurs in which section of a sound file. This kind of supervised data is called strongly labeled data. It would be easy to train the model if there was a large amount of such strongly labeled data. However, it takes a lot of time and human resources to prepare strongly labeled data. On the other hand, a labeling method assigns only a supervised label to an acoustic file with the information of what kind of sound event is contained in the file, instead of labeling the exact interval of the sound event occurrence. This type of supervised data is called weakly labeled data. Although weakly labeled data is easier to prepare than strongly labeled data, there is no doubt that labeling a large number of acoustic files is still expensive.

The DCASE 2019 and 2020 Task 4 provided a small amount of strongly labeled, weakly labeled, and a large amount of unlabeled data. Semi-supervised learning methods using these data are being actively studied. The baseline method provided by the Task 4 organizers also adopted semi-supervised learning, as shown in Fig. 1. The baseline model is trained in the framework of a mean-teacher method [2] based on a student-teacher model [6, 7]. The model parameters based on the loss value are updated only in the student model, and the parameters of the student model are also reflected in the teacher model. In other words, the teacher model will accumulate knowledge of the student model. By constraining the output of the teacher model to be the same as the output of the student model, the training of the student model is stabilized.

In previous work, for example, Lin et al. [8] proposed a guided learning method to train the student-teacher model more efficiently. Park et al. [9] suggested a tri-training approach, in which two different classifiers are used to acquire pseudo-labels from weakly labeled and unlabeled datasets, and to use them. Similarly, Ebbers et al. [10] proposed a method to train a classifier after assigning pseudo-labels to weakly labeled and unlabeled data sets. In addition, Kim et al. [11] improved the mean-teacher method and proposed a two-stage distillation method using a fine-tuning model based on semi-supervised loss, which achieved state-of-the-art performance on the DCASE 2020 Task 4 test set. Fukuda et al. [12] showed that the accuracy of knowledge distillation can be improved by ensemble fusion of multiple teacher networks in the Aurora 4 test set (speech recognition task under noise conditions). Thus, the methods for utilizing weakly labeled or unlabeled data can be divided into two categories: using knowledge distillation and pseudo-labeling with a tentative model. In this paper, we propose a method based on knowledge distillation. Our method differs from existing methods in that it performs knowledge distillation in the student's network, in addition to knowledge distillation between student-teacher networks. Therefore, our method can collaborate with other proposed methods, such as [11].

Recently, an online knowledge distillation method [13, 14, 15], an ensemble-based model for knowledge distillation into sub-networks, was proposed, and it has achieved high accuracy in image recognition tasks. This paper adopts a knowledge distillation model based on semi-supervised learning called "peer collaborative learning" (PCL) [16]. Originally, PCL was applied to image recognition tasks, and its effectiveness has been confirmed. In this study, we applied it to the PSED task for the first time. The main feature of this PCL model is that it combines the advantages of both the

online knowledge distillation method and the mean-teacher method. In other words, in the framework of the student-teacher model, the stronger teacher model that accumulates knowledge of the student model stabilizes the training of the student model, and the student model itself is equipped with a more accurate sound event detection function by knowledge distillation, using ensemble learning with multiple branching sub-networks inside the student model. This paper also proposes an original method to design the subnetworks inside the student model, depending on data augmentation methods for input sound data.

In the evaluation experiments, we used the test set used in the DCASE 2019 Task 4. As a baseline method, we used the mean-teacher method provided by the Task 4 organizer citeTurpault2019. The online knowledge distillation method [13] was used as a comparison method with the proposed method. As a result of the experiment, we obtained a significant improvement in the PCL approach over the baseline, with 8.2 points improvement in the F1-score, which is a public evaluation index for a PSED task. The PCL also showed an accuracy improvement of about 1 point compared with the online knowledge distillation method. We also found that the suitable design of the sub-networks, based on the data augmentation methods within the student model, could be improved.

The contributions of this paper are summarized as follows:

- We first demonstrated the usefulness of PCL in the PSED task. PCL is not always possible to apply a method that has been successful in the image recognition task to the acoustic signal classification task. In this study, we applied PCL with modifications to fit into the framework of acoustic event detection and confirmed that the results were better than the baseline scores.
- We also showed that the accuracy of the model can be improved by designing the internal sub-networks of the student model based on data augmentation methods.

2. SOUND EVENT DETECTION MODELS

2.1. Mean-Teacher Model

The mean-teacher method [2] was adopted as the baseline method for the DCASE 2019 Task 4. Figure 1 provides an overview of the mean-teacher method, and Fig. 2 shows the specific network structure of the student and teacher models.

This method uses two models with the same structure, and the two models are trained in a consistent way to maintain the consistency of their outputs. First, normal input sound data (sound signals) are input into the student model. On the other hand, the data with noise added to the sound signal are input to the teacher model, and each output result is obtained. Then, for the student model, the model parameters are updated using the classification loss (cross-entropy) for the assigned labels and the consistency loss (minimum square error) with the output of the teacher model. Finally, the parameters of the teacher model were updated by an exponential moving average of the parameter values of the student model. This exponential moving average is a parameter-copying method that gives weight to the most recently trained weight parameters of the model and considers model parameters that have been trained in the past. This makes the teacher model a temporal ensemble model of the student model (i.e., the teacher model stores the past learning states of the student model).

As this process is repeated, the teacher model is gradually trained into a model that reflects the learning process of the student model and finally becomes a role model for the student model, and can guide the training of the student model through the calculation

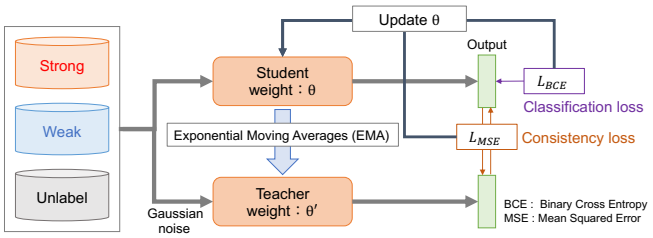


Fig. 1. A framework for a mean-teacher model.

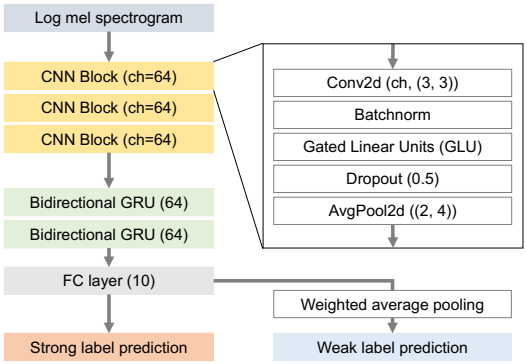


Fig. 2. The model architecture of the student and teacher models.

of consistency loss. The loss for parameter update using back propagation is calculated based on the total sum of L_{BCE} and L_{MSE}, shown in Fig. 1.

2.2. Online Knowledge Distillation

Figure 3 shows the training framework for the online knowledge distillation method and displays an example of the whole network with two sub-networks. The number of sub-networks can be increased, and the number of sub-networks was set to five in this paper. The detailed layer structure is the same as that of the PCL method in Fig. 4, and the student model of the PCL method and the network structure of the online knowledge distillation model in Fig. 3 are consistent.

This method uses parallel sub-networks inside the student model and an ensemble net that combines the feature representations extracted by each sub-network. The ensemble net is a very powerful feature extraction model because it integrates the output of each sub-network. By distilling the output of this ensemble net into each sub-network, they can extract complementary features for sound event classification. This means that the feature extraction performance of each sub-network is improved, and the generalization performance of the entire model is improved. Knowledge distillation using ensemble nets is very effective.

The number of sub-networks will be explained in the next section. For simplicity, we adopted a method that determines the sub-networks according to what kind of data augmentation process has been adopted on the input acoustic data. Note that the loss used to train the knowledge distillation model is the sum of the consistency losses and classification losses of the outputs of the ensemble net and sub-networks.

2.3. Peer Collaborative Learning

This section describes the PCL method, which combines the advantages of both the online knowledge distillation method and the average teacher method. The entire network of the model is shown in Fig. 4.

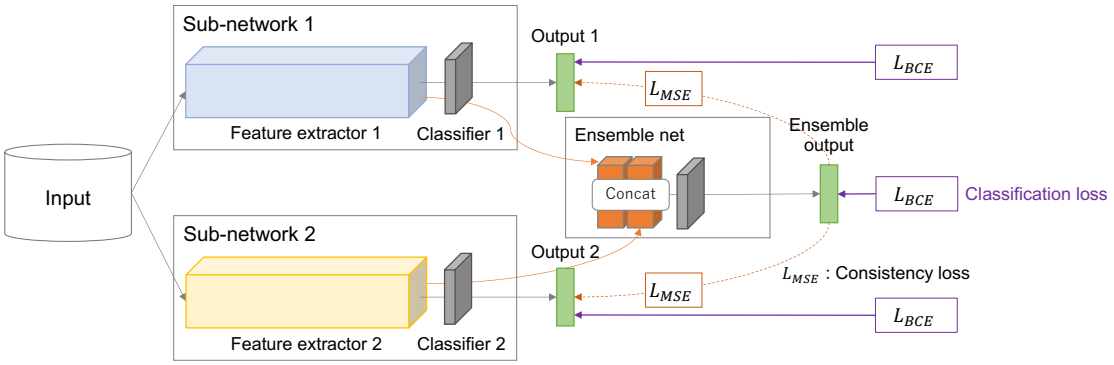
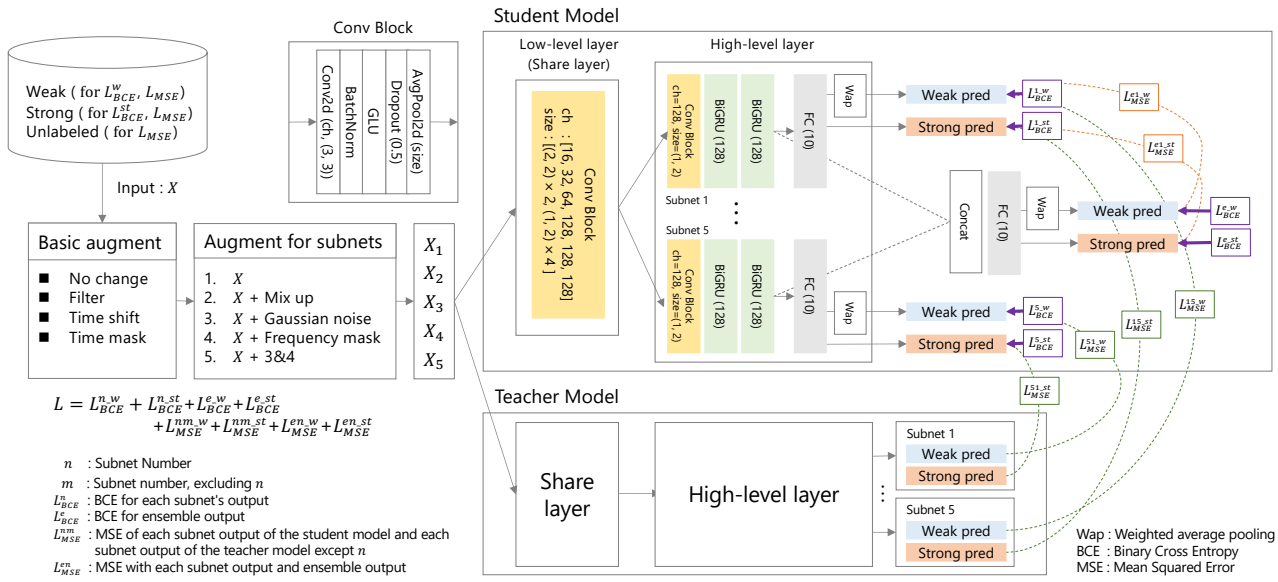


Fig. 3. A framework for training an online knowledge distillation model using ensemble net.



$$L = L_{BCE}^{n\_w} + L_{BCE}^{n\_st} + L_{BCE}^{e\_w} + L_{BCE}^{e\_st} + L_{MSE}^{nm\_w} + L_{MSE}^{nm\_st} + L_{MSE}^{en\_w} + L_{MSE}^{en\_st}$$

n : Subnet Number
m : Subnet number, excluding n
L_{BCE}^{n} : BCE for each subnet's output
L_{BCE}^{e} : BCE for ensemble output
L_{MSE}^{nm} : MSE of each subnet output of the student model and each subnet output of the teacher model except n
L_{MSE}^{en} : MSE with each subnet output and ensemble output

Fig. 4. The framework of peer collaborative learning, the layers that compose each model, and the calculation of losses.

First, the input acoustic data is subjected to a data augmentation process. In this study, we prepared a mixup, Gaussian noise addition, and frequency mask as data augmentation methods. In practice, the following five processes are employed: 1. no data augmentation, 2. mixup [17], 3. Gaussian noise addition, 4. frequency mask [18], and 5. Gaussian noise and frequency mask.

In the student model, we adopted a method called the peer ensemble model. The characteristic of this method is that the network is branched depending on the data augmentation process. When inputting a sound into the model, each augmented sound after data augmentation techniques passes through the lower layers shared by subnetworks. The features of the sound extracted in the lower shared layers are then input to the respective sub-network corresponding to the training of that sound. After each sub-network extracts the features of the input sound, it passes through the classifier layer (fully connected layer) and outputs the sound event prediction results. Here, as in Section 2.2 for online knowledge distillation, ensemble feature extraction is achieved by combining the features finally extracted by each sub-network. The classifier using these ensemble features is treated as an ensemble net and outputs the sound event prediction results. This ensemble output is then distilled into the output of each sub-network. In addition, the parameters of the teacher model are updated using the method described in Section 2.1. Note that ensemble nets are not used in the teacher model.

The aim of this method is to increase the generalizability of the acoustic feature representation extracted by the sub-networks according to the data augmentation methods. In addition, the proposed method incorporates the advantages of traditional online knowledge distillation methods, in that the sub-networks can extract complementary acoustic feature representations from each other because the ensemble net, which consolidates the knowledge of each subnetwork, distills the knowledge for the sub-networks inside the student model. Furthermore, knowledge distillation is performed between sub-networks to make the training between sub-networks more consistent. The knowledge distilled here is not the knowledge of the sub-networks in the student model but the knowledge of the more powerful teacher model built by the mean-teacher method, which allows for more stabilized training.

3. EXPERIMENTS

3.1. Experimental Setup

3.1.1. Dataset and evaluation metrics

The dataset consisted of audio clips with a maximum length of 10 seconds that were recorded in a home environment or synthesized to assume a home environment. The number of classes of sound events

was 10. The training dataset consisted of three types of datasets: a weakly labeled set, an unlabeled set, and a strongly labeled set. The weakly labeled set and the unlabeled set were sounds obtained from Audio Set [19], while the strongly labeled set was a synthetic sound generated by Scaper [20]. The weakly labeled, unlabeled, and strongly labeled sets contained 1578, 14412, and 2045 audio clips, respectively. In this paper, two types of datasets were used to evaluate the model: a strongly labeled validation set and a strongly labeled test set. The validation set was a combination of the validation set and the evaluation set from the DCASE 2018 Task 4 [21], and the test set was part of the Audio Set. The validation and test datasets contained 1168 clips and 692 clips, respectively.

The event-based F1-score [5] was used as the evaluation measure. This is a measure of how accurately the acoustic event intervals were estimated for the audio clips in the validation and test sets. The evaluation program used was provided by the Task 4 organizer[1].

3.1.2. Models and training condition

In this paper, we conducted the following six experiments to evaluate the effectiveness of the proposed method. First, as a baseline method, we used the mean-teacher method used in the DCASE 2019 Task 4, which was described in Section 2.1. In the DCASE 2020 Task 4, a slightly improved baseline model was proposed [22], denoted as baseline-advanced (adv). Next, as a comparison method, we used the online knowledge distillation method (Oneline KD) described in Section 2.2, which is a knowledge distillation method using sub-networks and ensemble nets. The number of sub-networks was set to five to accommodate the data augmentation process. For the PCL method, three models were prepared: five sub-networks with data augmentation (PCL w/ DA), five sub-networks without data augmentation (PCL w/o DA), and PCL w/ DA without ensemble net (PCL w/o ensemble). In all the models, basic data augmentation processes, such as frequency filters, time shifts, and time masks, were applied. Note that Online KD, PCL w/ DA model, and PCL w/o ensemble apply mixup, etc., in addition to these basic data augmentation processes for branching the sub-networks in the student model. Although the output of all the sub-networks and the ensemble net can be used for inference, the output of the ensemble nets is used as the final result. During inference, no data expansion is performed, and the acoustic features of the input data are fed directly into the model. The same output from the shared layer is input to all the sub-networks.

The baseline model structure is shown in Fig. 2. As input acoustic features, a 64-dimensional log Mel spectrogram was extracted from an audio clip recorded at 44.1 kHz using a window function with a width of 2048 points and 511 point hops. The optimization function used to train the model was Adam, and the learning rate was set to 0.001.

The baseline-adv. model is an improved version of the baseline model (see [22] for details of the model), and the input features differ from the baseline. The model structures used in the Oneline KD and PCL methods are shown in Figs. 3 and 4, respectively. The input acoustic features for these models, including the baseline-adv., are 128-dimensional log Mel spectrograms, computed from audio clips downsampled to 16 kHz, and cut with a width of 2048 points and 255 point hops window function. The optimization function was Adam, and the learning rate followed a ramp-up strategy [23], where the maximum learning rate was set to reach 0.001 after 50 epochs. Note that we used the dataset of DCASE 2019 Task 4, so the experimental conditions of DCASE 2019 were as close as possible to those of DCASE 2020 to emulate the experimental environment.

[1] https://github.com/TUT-ARG/sed_eval

Table 1. Sound event detection performance of each model (F1-score [%]). The numbers in parentheses are the published values on the DCASE website

Model	Validation set	Test set
Baseline	25.9 (23.7)	31.1 (29.0)
Baseline-adv.	34.7 (34.8)	36.2
Online KD	43.1	43.4
PCL w/ DA (*proposed*)	**43.8**	**44.2**
PCL w/o DA	41.7	42.4
PCL w/o ensemble	41.9	41.0

3.2. Results

Table 1 shows the experimental results for the two test sets.

First, it can be seen that the proposed method, PCL w/ DA, has the highest performance. This demonstrates the usefulness of the proposed method. Second, we can also see that both the Online KD and PCL w/o ensemble models show significant improvement compared with the two baselines. This can be attributed to the significant effect of knowledge distillation by the ensemble net and knowledge distillation from the teacher model to the student model in the mean-teacher framework. When comparing Online KD, PCL w/o ensemble, and PCL w/ DA, the F1-score had improved by 1 to 2 points compared with the case where each distillation method was separately applied. Therefore, it is clear that the fusion of each knowledge distillation method was more effective in improving the model than treating each method independently.

Furthermore, it is evident from the comparison between PCL w/ DA and PCL w/o DA that the model was improved by building sub-networks that depended on the data augmentation process.

These experimental results indicated two things: the PCL method was an effective method for the PSED task, and the accuracy of the model could be improved by designing the internal network of the student model based on the data augmentation method for acoustic data.

4. CONCLUSIONS

In this paper, we proposed the knowledge distillation method, PCL, for the PSED task, which makes effective use of weakly labeled and unlabeled data. The PCL method is different from previous knowledge distillation methods for PSED tasks in that it uses multiple sub-networks and an ensemble net that combines them in the student model. As a result of experiments using the test set of DCASE 2019 Task 4, we confirmed the effectiveness of knowledge distillation based on an ensemble net. In addition, the PCL approach, incorporating the existing mean-teacher method, further improved the performance of PSED. We found that the model with better generalization performance could be trained by changing the input sub-network according to the data augmentation method of the input sound data.

In future work, we will implement and experiment with new knowledge distillation methods, such as collaboration with other knowledge distillation methods (e.g., [11]).

5. ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 21H00901. Part of this work was also supported by the Hoso Bunka Foundation.

6. REFERENCES

[1] "Detection and classification of acoustic scenes and events," http://dcase.community/, [Online; accessed 7-Oct-2021].

[2] Nicolas Turpault, Romain Serizel, Justin Salamon, and Ankit Parag Shah, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, October 2019, pp. 253–257.

[3] Michael Mandel, Justin Salamon, and Daniel P. W. Ellis, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, October 2019.

[4] Nobutaka Ono, Noboru Harada, Yohei Kawaguchi, Annamaria Mesaros, Keisuke Imoto, Yuma Koizumi, , and Tatsuya Komatsu, *Proceedings of the Fifth Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2020)*, Tokyo, Japan, November 2020.

[5] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "Metrics for Polyphonic Sound Event Detection," *Applied Sciences*, vol. 6, no. 6, pp. 162, 2016.

[6] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean, "Distilling the Knowledge in a Neural Network," in *reprint arXiv:1502.02531*, 2015.

[7] Vimal Manohar, Pegah Ghahremani, Daniel Povey, and Sanjeev Khudanpur, "A Teacher-Student Learning Approach for Unsupervised Domain Adaptation of Sequence-Trained ASR Models," in *Proc. of 2018 IEEE Spoken Language Technology Workshop*, 2018, pp. 250–257.

[8] Liwei Lin, Xiangdong Wang, Hong Liu, and Yueliang Qian, "Guided Learning Convolution System for DCASE 2019 Task 4," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, October 2019, pp. 134–138.

[9] Hyoungwoo Park, Sungrack Yun, Jungyun Eum, Janghoon Cho, and Kyuwoong Hwang, "Weakly Labeled Sound Event Detection using Tri-training and Adversarial Learning," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, October 2019, pp. 184–188.

[10] Janek Ebbers and Reinhold Haeb-Umbach, "Forward-Backward Convolutional Recurrent Neural Networks and Tag-Conditioned Convolutional Neural Networks for Weakly Labeled Semi-Supervised Sound Event Detection," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, November 2020, pp. 41–45.

[11] Nam Kyun Kim and Hong Kook 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.

[12] Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran, "Efficient Knowledge Distillation from an Ensemble of Teachers," in *Proc. Interspeech 2017*, 2017, pp. 3697–3701.

[13] Jangho Kim, Minsung Hyun, Inseop Chung, and Nojun Kwak, "Feature Fusion for Online Mutual Knowledge Distillation," in *Proc. of ICPR 2021*, 2021, pp. 4619–4625.

[14] Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo, "Online Knowledge Distillation via Collaborative Learning," in *Proc. of CVPR 2020*, June 2020, pp. 11020–11029.

[15] Umar Asif, Jianbin Tang, and Stefan Harrer, "Ensemble Knowledge Distillation for Learning Improved and Efficient Networks," in *Proc. of 24th European Conference on Artificial Intelligence*, 2020, pp. 1–8.

[16] Guile Wu and Shaogang Gong, "Peer collaborative learning for online knowledge distillation," *Proc. of AAAI 2021*, vol. 35, no. 12, pp. 10302–10310, May 2021.

[17] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," in *Proc. of ICLR 2018*, 2018, pp. 1–13.

[18] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. of INTERSPEECH 2019*, 2019, pp. 2613–2617.

[19] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. of ICASSP 2017*, 2017, pp. 776–780.

[20] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 344–348.

[21] Romain Serizel, Nicolas Turpault, Hamid Eghbal-Zadeh, and Ankit Parag Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 19–23.

[22] Nicolas Turpault and Romain Serizel, "Training sound event detection on a heterogeneous dataset," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, November 2020, pp. 200–204.

[23] Samuli Laine and Timo Aila, "Temporal Ensembling for Semi-Supervised Learning," in *Proc. of ICLR 2017*, 2017, pp. 1–13.