Dynamic Thresholding on FixMatch with Weak and Strong Data Augmentations for Sound Event Detection

Tanmay Khandelwal and Rohan Kumar Das

Fortemedia Singapore, Singapore

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

Abstract

Recent state-of-the-art (SOTA) semi-supervised learning methods have shown great promise in improving sound event detection (SED) performance when the labeled data is scarce. Using a combination of consistency regularization, pseudo-labeling, and data augmentation techniques, the model predictions are constrained to be noise invariant. The recently proposed FixMatch achieved SOTA results on SED tasks. However, it uses a pre-defined constant threshold throughout the training process to generate the pseudo-labels, thus failing to account for the learning difficulties for each class and the model learning stage. To address this issue, we propose a dynamic thresholding method as an extension to FixMatch for generating pseudo-labels based on the model's predictions on weakly augmented features. This method retains the generated pseudo-labels based on the dynamic threshold value. The model is then trained to predict the generated pseudo-label when fed with a strongly augmented version of the same feature. On DCASE 2022 Task 4 2022 dataset, our method helped us in improving the SED system performance by 34.22% compared to the baseline in terms of polyphonic sound event detection score.

Index Terms: sound event detection, dynamic thresholding, data augmentation, consistency regularization

1. Introduction

The human auditory system is highly capable of detecting and segregating sound events to perceive changes in the surroundings. The sound event detection (SED) task automates the human auditory system to recognize sound events and mark their corresponding occurrences. Sound events in the real-world tend to have considerable overlap with each other, and the process of recognizing the overlapping events is referred to as polyphonic SED. It has a wide range of applications in real-world scenarios like smart-home devices [1], audio surveillance [2, 3] and monitoring biodiversity [4, 5].

Automatic SED systems are hindered by various challenges; some are dependent on the nature of sounds, while others are related to the data collection and annotation process. The recent advances in deep learning techniques have improved the performance of SED systems. Deep learning methods typically achieve their high performance by requiring a large amount of labeled data, which can be easily obtained for text and image applications compared to audio applications. The labeled data for audio applications is associated with higher annotation costs and is dependent on the subjective judgement of the annotator. As an alternative, labeled data can be generated synthetically from foreground and background samples, but it is still difficult to obtain an ample amount of foreground samples.

To address the scarcity of labeled data, systems commonly employ data augmentation (DA) techniques, consistency regularization (CR) [6, 7, 8], and pseudo-labeling [9, 10]. The DA methods artificially increase the amount as well as the diversity of data and improve the system's robustness by adding acoustic variability. The CR methods in contrast train the model to give consistent outputs for input and the perturbed variant. Whereas, pseudo-labeling makes predictions on the unlabeled samples to use the highly confident labels as training targets. The CR methods have a potential risk of confirmation bias when the loss is heavily weighed in training [11], as the consistency loss outweighs the classification loss, preventing the learning of new information. To reduce the risk, the mean-teacher (MT) [11] applies a consistency constraint in the model parameter space, as the teacher model uses the exponential moving average (EMA) weights of the student model. The recently proposed FixMatch [12] achieves a significant performance boost by combining weak as well as strong DA and applying the CR criterion. However, the FixMatch and other similar algorithms such as pseudo-labeling, and unsupervised domain adaptation [13] rely on a fixed constant threshold to compute the consistency loss. The use of a fixed threshold may lead to the selection of samples with wrong pseudo-labels and may not consider a few classes with learning difficulties.

To improve on the existing algorithm, we take inspiration from Dash [14] and incorporate a dynamic threshold in the MT model to select the data during the training process and then apply CR to the weakly and strongly augmented data. The threshold is gradually decreased over the number of iterations, adjusting to the learning stage of the model and the learning difficulty of some classes. The model initially learns using samples that are easy to learn and progresses towards hard samples, introducing a natural curriculum. It adds up another loss function to MT loss, to constrain the student model to give consistent predictions for the weakly and strongly augmented samples. We note that this technique requires no additional parameters or gradient computation and can be applied to any CR algorithm.

The detection and classification of acoustic scenes and events (DCASE) 2022 Task 4 focuses on semi-supervised learning (SSL) to utilize labeled and unlabeled data for developing SED systems. The SED systems are targeted to provide the event classes as well as the time localization of multiple events occurring together. We consider the two-stage system [15] that we developed for DCASE 2022 Task 4 challenge participation by incorporating the proposed method of dynamic thresholding with weak and strong augmentations in the second stage. We compare the performances of the proposed dynamic thresholding against FixMatch (constant threshold), single branches of strong and weak augmentations, and the baseline for DCASE 2022 Task 4 to show its impact.

The remainder of the paper is organized as follows: Section 2 introduces the baseline and the proposed extension to FixMatch for the SED system. In Section 3, the specifics of the experiments are described. The results and analysis are reported in Section 4. Finally, Section 5 concludes our work.

979-8-3503-9796-3/22/$31.00 ©2022 IEEE

2. Sound Event Detection System

2.1. Baseline

The DCASE 2022 Task 4 baseline [16] utilizes the MT [11] SSL method to effectively exploit large amounts of unlabeled data. In the MT model, we average the model weights across the training steps to produce a more accurate model than simply using the final weights. The teacher model does not participate in backpropagation; its weights are updated using the EMA of the student model. The MT loss (L_{MT}) can be divided into two parts: classification loss and consistency loss, given by Eq. (1), where L_{class} is the classification loss, L_{cons} is the consistency loss, and λ is a fixed scalar hyperparameter that represents the relative weight of the consistency loss. Again, the consistency loss is made up of two components: clip-wise consistency and frame-wise consistency, which compare the labels of both the student model and the teacher model across the entire dataset. We use the teacher's predictions during the testing stage because they are more likely to be correct.

$$L_{MT} = L_{class} + \lambda L_{cons} \tag{1}$$

2.2. Proposed

In this section, we define the individual components on which the FixMatch method is based before introducing the proposed dynamic thresholding method (DTM) as an extension to FixMatch.

2.2.1. Consistency Regularization (CR)

CR [6, 7] is a key component in the recent SSL algorithms including the MT model, it was first proposed in [7]. Consistency training methods regularize model predictions so that they are not affected by noise added to the input samples, making the model more robust to small changes in the input samples. It is achieved by minimizing the difference between the original input prediction and the prediction of the perturbed version of the same input. The methods differ in how and where the noise is introduced.

2.2.2. Pseudo-labeling

Pseudo-labeling [9, 10] uses the model itself to make predictions on the unlabeled samples and selects the samples where the prediction is confident (above a threshold) [10]. This is a type of entropy minimization, where the density of data points at the decision boundaries is reduced. The advantage of pseudo-labeling over CR is that no DA is required.

2.2.3. FixMatch

Recent advances in SSL have increased the complexity of learning algorithms, resulting in complicated loss terms and difficult-to-tune hyperparameters. FixMatch defies this trend by proposing an algorithm that can be easily integrated on top of other algorithms. It employs both CR and pseudo-labeling to generate artificial labels. The generated label is produced on the weakly augmented audio sample, which is then used as a target in a standard cross-entropy loss function when the same model is fed with a strongly augmented version of the same audio sample, introducing a form of CR as shown in Figure 1. Similar to the method of pseudo-labeling, the method assigns a label to the weakly augmented version if it crosses a fixed threshold. The

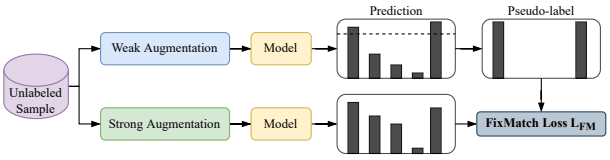

Figure 1: *Schematic diagram of FixMatch.*

FixMatch loss function L_{FM} is given as:

$$L_{FM} = \sum_{x \epsilon B} BCE(f(x_{strong}), f(x_{weak})) \tag{2}$$

where f indicates the student model, x_{strong} represents the strongly augmented sample, x_{weak} represents the pseudo-label generated from the weakly augmented sample and BCE represents the binary cross-entropy loss taken over a batch size of *B*. The model's predictions become more confident as the training progresses.

2.2.4. Data Augmentation (DA)

DA is a common strategy to increase the amount of training data. Such techniques are useful when building models with limited training datasets. We extend FixMatch to SED and employ two types of augmentations, weak and strong, as in [17]. In this work, using a similar set of augmentation as in [18], we further employed frame shift [18], mixup [19] and time masking [20] common to both strong and weak augmentations. Table 1 shows the weak and strong augmentations utilized in our experiments.

Table 1: *The weak and strong augmentation used in our experiments for SED.*

Weak Augmentation	Strong Augmentation
Filter augmentation	Filter augmentation
	Frequency masking
	Gaussian noise addition

Mixup randomly combines selected samples with a mixing parameter, assisting in linear interpolation to improve the robustness of the model. The features and labels are shifted along the time axis by frame shifting. Time masking masks consecutive time steps chosen from a uniform distribution, whereas filter augmentation [18] applies random weights on random frequency regions. Again, frequency masking [20] randomly masks 16 of 128 mel bins and Gaussian noise addition helps to generalize well to noisy data as well.

2.2.5. Dynamic Thresholding Method (DTM)

Inspired by Dash [14], we extend the idea of FixMatch and propose to integrate a dynamic threshold for the selection of samples into the MT model, as illustrated in Figure 2. The threshold is decreased with the number of epochs. During the early stages of learning, the model may blindly predict samples into certain classes depending on the parameter initialization. To address this issue, the threshold value is reduced after a fixed number of warm-up epochs *(w)*, allowing the model to learn the representations first, and ensuring that only highly confident labels are selected in the early learning stage. Mathematically, we set the dynamic threshold Th as a decreasing function of *t* (number of epochs), given by:

$$Th(\rho) = Ce^{-\alpha \times phase} \tag{3}$$

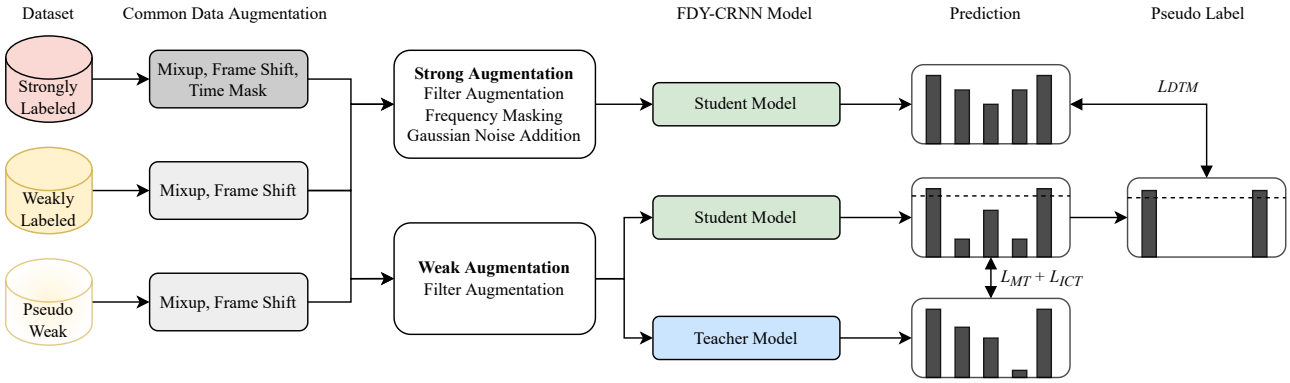

Figure 2: *Schematic diagram of the proposed dynamic thresholding on FixMatch for MT model.*

$$phase = \frac{t}{T} \quad \text{for} \quad t > w \tag{4}$$

where $C > 1$ such that ρ is in between [0,1]. Here, the phase is given by the ratio of t and the total number of epochs *(T)*. The decreasing threshold helps the model to learn progressively from easy to difficult data for training. This addresses the differences in learning between sound events. We use this threshold value to select the samples with predictions above the threshold from the weakly augmented samples and use them as the ground truth against the strongly augmented samples, achieving CR. The theoretical analysis in Dash also shows the convergence guarantee of the proposed dynamic thresholding. In the case of image classification, the model only had to classify into one of the target images, but in sound event detection due to its polyphonic behavior, multiple sound events can be present. We modify FixMatch to select more than one sound event at the same time, instead of taking the maximum across all the events. The updated total loss function L_{Total} comprises:

$$L_{Total} = L_{MT} + \gamma L_{DTM} \tag{5}$$

where L_{MT} is the mean-teacher loss as described in Section 2.1, L_{DTM} is the substitute for the updated FixMatch loss (L_{FM}) described in Section 2.2.3 and γ is the weighing parameter. This method helps in stabilizing the training procedure, as the self-consistent predictions on strong and weak augmentation hold regardless of the correctness of the predictions.

3. Experimental Setup

The following subsections describe the experimental setup for our studies.

3.1. Dataset

The DCASE 2022 Task 4 dataset utilized in this work is composed of 10 seconds audio clips either taken from AudioSet [21] or synthesized using Scaper to simulate a domestic environment. Table 2 shows the development dataset distribution.

3.2. Pre-processing

We resampled the audio clips at 16 kHz to a mono channel using librosa. They are then divided into segments with a window size of 2048 samples and a hop length of 256 samples for each succeeding frame. The short-time Fourier transform is applied to the segmented waveforms to extract their spectrograms. Then, log-mel spectrograms are produced by applying mel-filters in the frequency domain spanning from 0 to 8 kHz, followed by a logarithmic operation. Silence padding is used for clips that are under 10 seconds.

Table 2: *DCASE 2022 Task 4 dataset split for development set.*

Clips	Description
10,000	Synthetic strongly labeled data
3,470	Real strongly labeled data
1,578	Real weakly labeled data
14,412	In-domain unlabeled data
1,168	Real strongly labeled validation data

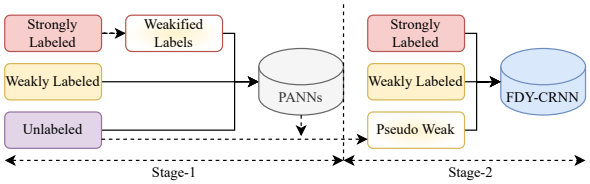

Figure 3: *Proposed two-stage learning setup, with Stage-1 focusing on AT and Stage-2 focusing on SED.*

3.3. Two-stage System (TSS) for SED

This section describes the details of our two-stage system (TSS) shown in Figure 3, which we considered for the studies. Stage-1 focuses on audio-tagging (AT) and Stage-2 uses the reliable pseudo-labels generated by Stage-1 to improve SED. Furthermore, each stage makes use of MT, to exploit the unlabeled training data. In both stages of the TSS, we use another SSL method called interpolation consistency training (ICT) [6] in addition to MT. The ICT replaces all input samples with interpolated samples, assisting the model's generalization ability. We further extend Stage-2 of this TSS to utilize the proposed DTM described in Section 2.2.5 for CR and apply the weak and strong DA described in Section 2.2.4. We provide a detailed description of the models used in each stage in the following.

3.3.1. Stage-1

As the feature extractor, we used convolutional neural network (CNN)-14-based pre-trained audio neural networks (PANNs) [22] to extract the embeddings. The embedding features are fed into the bi-directional gated recurrent unit (Bi-GRU) [23]. The PANNs-based embedding parameters are unfrozen and trained. There are two layers of Bi-GRU with 1024

hidden units following the feature extractor. Stage-1 is trained utilizing a strongly labeled set converted into weak predictions referred to as a weakified set, a weakly labeled set, and an unlabeled set to improve the AT performance as demonstrated in Figure 3. The Bi-GRU output is followed by a dense layer with sigmoid activation to produce frame-level predictions, and the aforementioned linear layer is multiplied by a dense layer with a softmax activation function to produce clip-level predictions.

3.3.2. Stage-2

In this work, we used the AT system (Stage-1) to make predictions on the unlabeled set to use them as pseudo-weak labels in Stage-2 training, as shown in Figure 3. In Stage-2, we used frequency dynamic convolution recurrent neural network (FDY-CRNN) [24] based frequency-dependent architecture to replace the convolution recurrent neural network (CRNN) [16] in the baseline. It is trained on a pseudo-weakly labeled set, in addition to the strongly labeled set and the weakly labeled set. We apply the strong and weak augmentations described in Section 2.2.4. Using the weakly augmented features with high confidence (above the threshold), we establish the ground truth against the strongly augmented features. We reduce the threshold exponentially with the number of epochs, as described in Section 2.2.5. Thus, the total loss function (L_{Total}) demonstrated in Figure 2 for this stage with $\gamma = 1$ for the L_{DTM} and with addition of loss for ICT (L_{ICT}) defined in Section 3.3 is:

$$L_{Total} = L_{MT} + L_{ICT} + L_{DTM} \tag{6}$$

3.4. Training Process

For all experiments, the batch size was set as 48 (1/4 strong set, 1/4 weak set, 1/2 unlabeled set). We used Adam optimizer [25] with a learning rate of 0.001 and an exponential warm-up for the first 50 epochs, increasing the weighing parameter λ defined in Section 2.1 linearly in steps from 0 to 2, and then holding constant at 2. A 90% training set and a 10% cross-validation set were created from the weakly labeled set. Cross-validation is performed on the 10% held-out weak subset and the additional synthetic validation data. The system was built with PyTorch Lightning and trained on NVIDIA Quadro RTX 5000 GPU. In a total of 200 epochs, we used 50 as the warm-up epochs to keep the threshold fixed at 0.9 and help the model learn using highly confident predictions. We then reduced the threshold in accordance with the method described in Section 2.2.5 with values *C=1.094796* and α*=0.783716* for the parameters, generated by fitting the exponential curve between the empirically chosen values 0.9 and 0.5 with the required decrease in 150 epochs. After the warm-up, the threshold is dynamically decreased from 0.9 to 0.5 in the remaining epochs.

3.5. Evaluation Metric

We used the polyphonic sound event detection score (PSDS) [26] as a performance metric in our studies. The PSDS is more robust to the labeling subjectivity, leaving sufficient room for interpretation of the temporal structure of both the ground truth and the detections. It computes a single PSDS using polyphonic receiver operating characteristic curves, allowing the comparison to be independent of the operating point. It can be tailored to various applications, ensuring that the desired user experience is met. As a result, it overcomes the limitations of traditional collar-based event F-scores. The DCASE 2022 Task 4 employs two distinct scenarios that highlight different system properties. The first scenario (PSDS1) requires the system to respond quickly to event detection, focusing on the temporal localization of the sound event. On the other hand, the second scenario (PSDS2) focuses on preventing class confusion rather than on reaction time.

Table 3: *Performance comparison showing the importance of the proposed method on the DCASE 2022 Task 4 validation set.*

System	PSDS1	PSDS2	PSDS1+PSDS2
Baseline: CRNN	0.351	0.552	0.903
Two-stage system (TSS)	0.472	0.721	1.193
TSS + Weak-DA	0.437	0.678	1.115
TSS + Strong-DA	0.420	0.647	1.067
TSS + FixMatch (ρ=0.5)	0.485	0.717	1.202
TSS + FixMatch (ρ=0.9)	0.480	0.723	1.203
Proposed: TSS + DTM-FixMatch	**0.489**	**0.723**	**1.212**

4. Results and Analysis

We are first interested in comparing the performance of the baseline with the TSS described in Section 3.3, followed by the studies involving FixMatch with empirically chosen fixed thresholds (0.5 and 0.9) and our proposed DTM on FixMatch. For ablation studies, we also consider the single weak and strong DA branches employed in Stage-2 of TSS without CR and pseudo-labeling demonstrated in Figure 2. We also employed some existing post-processing techniques to further assess the efficacy of the proposed method and conduct fair comparisons with the current state-of-the-art (SOTA). We used adaptive post-processing [27] in both stages of the TSS, using a different median filter window size for each event category. During inference, we set the temperature parameter [28] in the sigmoid function to 2.1.

Table 3 shows a comparison of all the systems stated above along with the ablation studies. We observe that the TSS system used in this work is performing much better than the baseline, proving it as one of the SOTA systems. We further find that only introducing either weak or strong augmentations on TSS does not help to improve the results; but considering them together in FixMatch improves the performance due to consideration of CR with pseudo-labeling, which is more evident when a fixed threshold of 0.9 is considered. Further, our proposed DTM that dynamically varies the threshold with the number of epochs on FixMatch outperforms the systems with fixed thresholds. Compared to the baseline CRNN, the proposed DTM-FixMatch achieves a 34.22% noticeable improvement in terms of the total PSDS (PSDS1+PSDS2), which depicts the effectiveness of the proposed system.

5. Conclusion

In this work, we proposed a DTM as an extension to FixMatch for the SED system. The method employs a combination of pseudo-labeling and CR, as well as two types of data augmentations: strong and weak. It generates an artificial label using the model's predictions that are confident (above a threshold) on a weakly augmented version of the sample. The generated pseudo-labels are used to enforce cross-entropy loss against the model's prediction for the strongly augmented version. The threshold is exponentially decreased with the number of epochs after a few warm-up epochs. The studies on the DCASE 2022 Task 4 dataset reveal that the proposed dynamic thresholding on FixMatch improved the performance of FixMatch with a fixed threshold, as well as outperformed the DCASE 2022 Task 4 baseline by a large margin.

6. References

[1] 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*, vol. 62, no. 2, pp. 68–77, 2019.

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

[3] 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.

[4] S. Chu, S. Narayanan, and C. J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1142–1158, 2009.

[5] B. Furnas and R. Callas, "Using automated recorders and occupancy models to monitor common forest birds across a large geographic region," *The Journal of Wildlife Management*, pp. 325–337, 2014.

[6] V. Verma, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3635–3641, 2019.

[7] P. Bachman, O. Alsharif, and D. Precup, "Learning with pseudo-ensembles," *Advances in Neural Information Processing Systems*, pp. 3365–3373, 2014.

[8] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," *International Conference on Neural Information Processing Systems*, pp. 1171–1179, 2016.

[9] J. Ebbers and R. Haeb-Umbach, "Self-trained audio tagging and sound event detection in domestic environments," *Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, pp. 226–230, 2021.

[10] D.-H. Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 2013.

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

[12] K. Sohn, D. Berthelot, C. L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," *International Conference on Neural Information Processing Systems*, 2020.

[13] H. Dinkel, X. Cai, and Z. Yan, "The smallrice submission to the DCASE 2021 Task 4 challenge: A lightweight approach for semi-supervised sound event detection with unsupervised data augmentation," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2021.

[14] Y. Xu, L. Shang, J. Ye, Q. Qian, Y.-F. Li, B. Sun, H. Li, and R. Jin, "Dash: Semi-supervised learning with dynamic thresholding," *International Conference on Machine Learning (ICML)*, 2021.

[15] T. Khandelwal, R. K. Das, A. Koh, and E. S. Chng, "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.

[16] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for DCASE 2019 Task 4 technical report," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2019.

[17] C. Kim and S. Yang, "Sound event detection system using FixMatch for DCASE 2022 challenge Task 4," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2022.

[18] H. Nam, B.-Y. Ko, G.-T. Lee, S.-H. Kim, W.-H. Jung, S.-M. Choi, and Y.-H. Park, "Heavily augmented sound event detection utilizing weak predictions," 2021. [Online]. Available: https://arxiv.org/abs/2107.03649

[19] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017. [Online]. Available: https://arxiv.org/abs/1710.09412

[20] 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," *Interspeech*, pp. 2613–2617, 2019.

[21] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "AudioSet: An ontology and human-labeled dataset for audio events," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017.

[22] 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*, vol. 28, pp. 2880–2894, 2020.

[23] 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.

[24] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, "Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection," 2022. [Online]. Available: https://arxiv.org/abs/2203.15296

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

[26] C. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulovic, "A framework for the robust evaluation of sound event detection," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 61–65, 2020.

[27] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Convolution-augmented transformer for semi-supervised sound event detection," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2020.

[28] X. Zheng, H. Chen, and Y. Song, "Zheng USTC team's submission for DCASE 2021 Task 4 - semi-supervised sound event detection," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2021.