SELF TRAINING AND ENSEMBLING FREQUENCY DEPENDENT NETWORKS WITH COARSE PREDICTION POOLING AND SOUND EVENT BOUNDING BOXES

Technical Report

*Hyeonuk Nam, Deokki Min, Seungdeok Choi, Inhan Choi, Yong-Hwa Park**

Korea Advanced Institute of Science and Technology, South Korea,
{frednam, minducky, haroldchoi6, ds5amk, yhpark}@kaist.ac.kr

ABSTRACT

To tackle sound event detection (SED) task, we propose *frequency dependent networks (FreDNets)*, which heavily leverage frequency-dependent methods. We apply frequency warping and FilterAugment, which are frequency-dependent data augmentation methods. The model architecture consists of 3 branches: audio teacher-student transformer (ATST) branch, BEATs branch and CNN branch including either partial dilated frequency dynamic convolution (PDFD) or squeeze-and-Excitation (SE) with time-frame frequency-wise SE (tfwSE). To train MAESTRO labels with coarse temporal resolution, we apply max pooling on prediction for the MAESTRO dataset. Using best ensemble model, we apply self training to obtain pseudo label from DESED weak set, DESED unlabeled set and AudioSet. AudioSet labels are filtered to focus on high-confidence pseudo labels and AudioSet pseudo labels are used to train on DESED labels only. We used change-detection-based sound event bounding boxes (cSEBBs) as post processing for ensemble models on self training and submission models.

***Index Terms*—** frequency dynamic convolution, audio pretrained models, coarse prediction pooling, label filtering, sound event bounding boxes

1. INTRODUCTION

In this work, we address the problem of sound event detection (SED) with heterogeneous datasets, including Domestic Environment Sound Event Detection (DESED) and Multi-Annotator Estimated STROng labels (MAESTRO) [1, 2, 3]. Since SED is a very delicate task which requires time localization in addition to class information, the difference between two datasets must be carefully addressed. While DESED uses hard labels with fine temporal resolution (base unit of one millisecond) and includes ten target sound events those occur in domestic environment, MAESTRO uses soft labels representing confidence with coarse temporal resolution (base unit of one second) and includes seventeen target sound events those occur in outside environments. There are only few target sound events overlapping. Though the target sound events from one dataset might exist in the other dataset, we cannot know this since they are not labeled. This arouses the problem of potentially missing labels [1]. To tackle this problem, DCASE2024 Challenge Task 4 baseline is designed to train both datasets using single model architecture to output for twenty seven classes, while masking the classes from one dataset when training for data from the other dataset [1].

In this work, our primary approach is to build strong classifier that works on both datasets. To achieve this, we applied two frequency-dependent data augmentations: frequency warping and FilterAugment [4, 5]. Then, we applied advanced variants of frequency dynamic convolution (FDY conv) to CNN branch of the baseline [6, 7, 8]. We also used squeeze and excitation (SE) with time-frame frequency wise SE (tfwSE) to CNN branch [9]. In addition to BEATs branch, we added audio teacher student transformer (ATST) branch to form three-branched models consisting of ATST branch, BEATs branch and CNN branch [4, 10, 11]. Then, in order to match the granularity of strong prediction and MAESTRO strong labels, we pooled predictions to train with coarse MAESTRO label. Since frequency-dependent methods are heavily used. we call above network architecture as *Frequency Dependent Networks (FreDNets)*. We used change-detection-based sound event bounding boxes (cSEBBs) as post processing [12]. With ensemble of FreDNets post-processed by cSEBBs, we produced pseudo labels on AudioSets, and used them to train new FreDNets [13].

The main contributions of this paper are as follows:

1. Proposed *Frequency dependent networks (FreDNets)* heavily utilizes frequency-dependent methods to outperform the baseline by 15.1% without ensemble.
2. Proposed coarse prediction pooling harmonizes the temporal resolution difference between the model and label.
3. Partial dilated frequency dynamic convolution (PDFD conv) used in FreDNets are lighter than FDY conv or DFD conv and provides various variant models thus advantageous upon ensemble.

2. METHODS

2.1. Frequency-Dependent Data Augmentations

In addition to mixup applied in the baseline [1, 14], we added frequency warping and FilterAugment [4, 5]. The sequence of operation is as follows: mixup, frequency warping then FilterAugment. Frequency warping is random resize crop applied only along frequency dimension to zoom into frequency dimension with random proportion. As it also works as frequency shift, we did not apply additional frequency shift. Then, we applied linear type FilterAugment with dB range from -3 dB to +3 dB. This is narrower range

*This work was supported by the Institute of Civil Military Technology Cooperation funded by the Defense Acquisition Program Administration and Ministry of Trade, Industry and Energy of Korean government under grant No. UM22409RD4, and Korea Research Institute of Ships and Ocean engineering a grant from Endowment Project of "Development of Open Platform Technologies for Smart Maritime Safety and Industries" funded by Ministry of Oceans and Fisheries(PES5230).

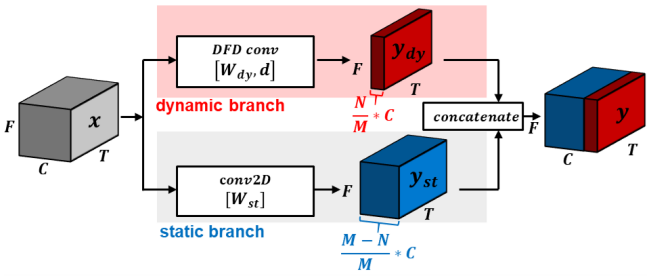

Figure 1: An illustration of partial dilated frequency dynamic convolution. It involves a dynamic DFD conv branches and a static 2D convolution branch.

compared to the setting in the original paper [5]. FilterAugment applies random weights over different frequency ranges to simulate different acoustic environments. Data augmentation is only applied to CNN branch as shown at the top of Fig. 2, because the other two branches are not trainable. Since frequency dependency is an important issue in SED, these two methods showed performance gain upon simultaneous application.

2.2. Frequency-Dependent CNN Methods

To further enhance the capacity of the network, CNN and RNN channels are doubled. Either variants of frequency dynamic convolution (FDY conv) or squeeze-and-excitation (SE) to make CNN modules to leverage frequency-dependent attention methods.

FDY conv applies frequency-adaptive convolution kernel to release translational equivariance along frequency axis of time-frequency features [6]. To lighten FDY conv, we applied partial frequency dynamic convolution (PFD conv) [8]. To diversify the basis kernels, we applied dilated frequency dynamic convolution (DFD conv) to PFD conv to obtain lighter and diverse version of FDY conv. We refer to this method as partial dilated frequency dynamic convolution (PDFD conv), which is illustrated in Fig. 1. Using different dilation sizes to PDFD conv resulted in various models which are advantageous on model ensemble [15]. While multi-dilated frequency dynamic convolution (MDFD conv) yields in the best performance, we used PDFD since it offers best cost-performance balance considering training time [8]. In addition to PDFD convs, we also used SE with time-frame frequency-wise SE (tfwSE) for model variety upon ensemble [9, 15].

2.3. Transformer-based Pre-trained Audio Models

In addition to CNN branch, two transformer-based pre-trained audio models are used: BEATs and ATST. Frame-wise feature of BEATs and ATST-frame are used to optimally enhance SED which needs to give frame-wise predictions. Embeddings extracted for both methods are pooled into same frame size as output by CNN module output, then concatenated with the output from CNN module along channel dimension, and then processed by fully connected layers along channel dimension. Then the output is fed to RNN module. Note that since transformer-based Audio models divide mel spectrogram into patches and then apply positional encoding to the patches, they implicitly apply frequency-dependent processing. Thus two audio models can be regarded frequency-dependent methods. Fine tuning of ATST is not used in this work as it negatively affects MPAUC on MAESTRO [4].

2.4. Coarse Prediction Pooling

In order to address the different temporal resolution of DESED and MAESTRO, we applied coarse prediction pooling on MAESTRO.

Table 1: Components models of ensemble models. $1/8$ denotes that $1/8$ of PFD conv or PDFD conv output channel is from FDY conv or DFD conv. Sd, ds and st implies seed, dilation sizes and self training. For model names, CRNN is omitted for brevity.

ensemble	models
1	PFD(1/8), PFD(1/8, sd=2), PFD(1/8, sd=12), PFD(1/8, sd=16), PFD(1/8, sd=27), PFD(1/8, sd=34), PDFD(1/8, ds=1122), PDFD(1/8, ds=1133), PDFD(1/8, ds=2233), PDFD(1/8, ds=1123), PDFD(1/8, ds=1223), PDFD(1/8, ds=1233)
2	PFD(1/8), PFD(1/8, sd=16), PDFD(1/8, ds=1122), PDFD(1/8, ds=1133), PDFD(1/8, ds=1123), PDFD(1/8, ds=1223), PDFD(1/8, ds=1233), st-PFD(1/8), st-PFD(1/8, sd=2), st-PFD(1/8, sd=12), st-SE+tfwSE, st-PDFD(1/8, ds=1122), st-PDFD(1/8, ds=1123), st-PDFD(1/8, ds=1223), st-PDFD(1/8, ds=1233)
3	PFD(1/8), PFD(1/8, sd=16), SE+tfwSE, PDFD(1/8, ds=1122), PDFD(1/8, ds=1133), PDFD(1/8, ds=1123), PDFD(1/8, ds=1223), PDFD(1/8, ds=1233), st-PFD(1/8), st-PFD(1/8, sd=2), st-PFD(1/8, sd=12), st-PFD(1/8, sd=27), st-SE+tfwSE, st-PDFD(1/8, ds=1122), st-PDFD(1/8, ds=1123), st-PDFD(1/8, ds=1223), st-PDFD(1/8, ds=1233),

While FreDNets' predictions have temporal resolution of 64ms per frame (156 frames for 10 seconds), MAESTRO label has temporal resolution of 1s per frame (10 frames for 10 seconds). To make fine predictions into coarse predictions, we apply max pooling on FreDNets' MAESTRO prediction. To be more specific, we zero-padded 2 frames before and after the prediction and max pooled with filter size and stride of 16. Although this is not precise pooling, this choice was made to quickly and simply implement the idea.

2.5. Sound Event Bounding Boxes

Polyphonic sound detection score (PSDS) applies various thresholds to the SED prediction to obtain threshold-independent evaluation values [16, 17]. However, as threshold differs, onset and offset of sound events also varies. To make onset and offset of sound events independent of the thresholds, sound event bounding boxes (SEBBs) are proposed to combine confidence values with very fine onset and offset values into representative confidence, onset and offset values [12]. In this work, change-detection-based SEBBs (cSEBBs) are used.

2.6. Self Training using AudioSet

To obtain pseudo labels on DESED weak set, DESED unlabeled set and AudioSet, we used ensemble of FreDNet using PDFD-CNN modules with varying dilation size sets, SE+tfwSE-CNN and PFD-CNN with varying seeds and then applied cSEBBs [15, 18]. As DESED weak set is given with weak labels, pseudo label for weak set is masked with given weak labels as in [19]. Since AudioSet has inconsistent label quality, we applied self training on whole dataset to obtain confidence from our ensemble FreDNet. For AudioSet, we filtered data files having pseudo label values (confidence) above 0.7 on 27 target events to focus on labels with high confidence. we

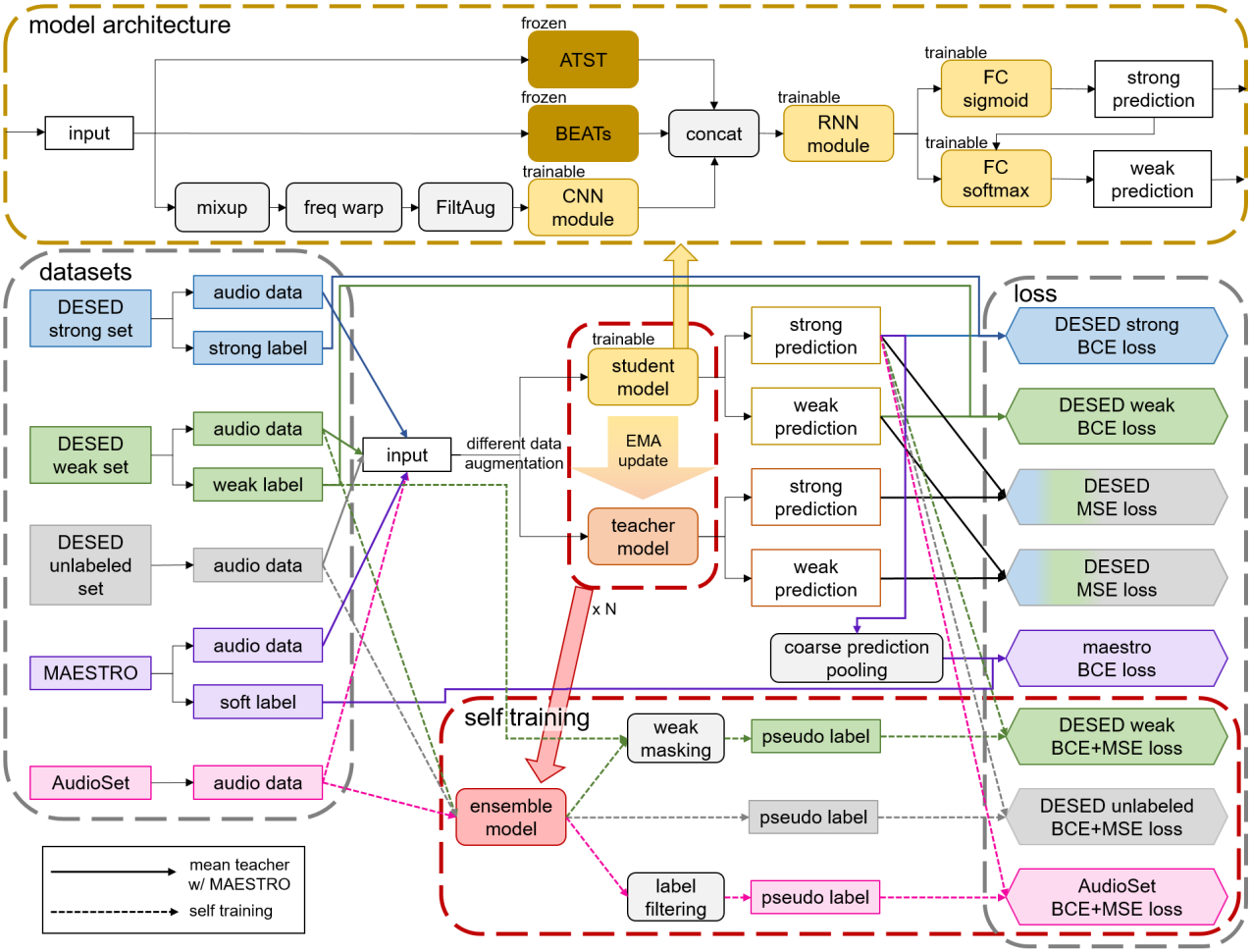

Figure 2: An illustration of framework for training and self training FreDNets.

discarded event labels with confidence value below 0.01 to reduce pseudo label metadata size, and removed the files of which events above 0.7 are only composed of subset of (speech, people talking, children voices) to reduce the data imbalance toward speech. The count of filtered AudioSet files is 153,977.

Upon use of pseudo label, both hard label obtained by threshold of 0.5 and soft label are used to train SED model on AudioSet data. For hard label, binary cross entropy (BCE) loss is used and for soft label, mean square error (MSE) loss is used as shown in red dashed line box in Fig. 2. Only 10 target sound events for DESED are trained using filtered AudioSet as it degraded MPAUC when trained on MAESTRO target sound events, although it was meant to train on 17 target sound events in MAESTRO as well.

2.7. Ensemble

Ensemble model averaged predictions from various models. To maximize the effect of ensemble, we used different models including PFD-CRNN, PDFD-CRNN with different dilation size sets, and SE+tfwSE-CRNN, and PFD-CRNN with different seeds. For each model setting, the student and teacher models with the best sum score (PSDS1+MPAUC) are used for ensemble. The model combinations used for each ensemble setting is shown in Table 1. Ensemble 1 is used to extract pseudo labels from AudioSet. Ensemble 2 and 3 are used for DCASE Challenge submission. While PFD-CRNNs with different seeds are generally worse than models with seed of 42, models with different seeds do help enhancing ensemble performance.

3. EXPERIMENTAL SETTINGS

3.1. Implementation Details

DESED and MAESTRO processed to be 10 seconds clip with 16kHz sampling rate are used in this work [1, 3, 20]. The network is composed of three-branched ATST-BEATs-CNN modules which are then fed to RNN module and Fully Connected layers as shown in Fig. 2. The Mean Teacher method is employed to train FreDNets using the DESED unlabeled set [20, 21]. Binary cross entropy (BCE) loss is used to train strong prediction for DESED strong set and its strong label, weak prediction for DESED weak set and its weak label, and strong prediction of MAESTRO and its soft label. Note that strong prediction goes through coarse label prediction before the loss function to match the granularity of prediction and label. For consistency loss for strong and weak predictions of DESED sets, mean square error (MSE) loss is used. For pseudo labels for DESED weakly labeled set, unlabeled set and AudioSet, both BCE and MSE losses are used. Default seed is set to 42. GPU used for training is NVIDIA RTX A6000. For post-processing, we use either cSEBBs or a median filter as reported in Table 2. The median filter

Table 2: Performance of FreDNets.

models	pre-trained models	post-processing	self training	PSDS1	MPAUC	sum	# submission
Baseline [1]	BEATs	median filter	-	0.520	0.637	1.157	-
PFD-CRNN(1/8)	ATST + BEATs	median filter	-	0.516	0.775	1.293	-
PFD-CRNN(1/8, sd=2)	ATST + BEATs	median filter	-	0.502	0.766	1.268	-
PFD-CRNN(1/8, sd=12)	ATST + BEATs	median filter	-	0.514	0.765	1.279	-
PFD-CRNN(1/8, sd=16)	ATST + BEATs	median filter	-	0.514	0.772	1.286	-
PFD-CRNN(1/8, sd=27)	ATST + BEATs	median filter	-	0.514	0.763	1.277	-
PFD-CRNN(1/8, sd=34)	ATST + BEATs	median filter	-	0.508	0.769	1.276	-
PDFD-CRNN(1/8, 1122)	ATST + BEATs	median filter	-	0.519	0.773	1.292	-
PDFD-CRNN(1/8, 1133)	ATST + BEATs	median filter	-	0.523	0.767	1.290	-
PDFD-CRNN(1/8, 2233)	ATST + BEATs	median filter	-	0.515	0.772	1.287	-
PDFD-CRNN(1/8, 1123)	ATST + BEATs	median filter	-	0.518	**0.776**	1.294	-
PDFD-CRNN(1/8, 1223)	ATST + BEATs	median filter	-	**0.526**	0.772	**1.298**	-
PDFD-CRNN(1/8, 1233)	ATST + BEATs	median filter	-	0.518	0.774	1.292	-
SE+tfwSE-CRNN	ATST + BEATs	median filter	-	0.507	0.773	1.280	-
Ensemble 1	ATST + BEATs	median filter	-	**0.527**	**0.790**	**1.317**	-
Ensemble 1	ATST + BEATs	cSEBBs	-	**0.577**	**0.790**	**1.367**	-
PFD-CRNN(1/8)	ATST + BEATs	median filter	True	**0.539**	0.773	**1.312**	-
PFD-CRNN(1/8, sd=2)	ATST + BEATs	median filter	True	0.534	0.766	1.300	-
PFD-CRNN(1/8, sd=12)	ATST + BEATs	median filter	True	0.534	0.753	1.287	-
PFD-CRNN(1/8, sd=27)	ATST + BEATs	median filter	True	0.531	0.750	1.287	-
PDFD-CRNN(1/8, 1122)	ATST + BEATs	median filter	True	0.530	0.774	1.304	-
PDFD-CRNN(1/8, 1133)	ATST + BEATs	median filter	True	0.535	0.761	1.296	-
PDFD-CRNN(1/8, 1123)	ATST + BEATs	median filter	True	0.537	**0.775**	**1.312**	-
PDFD-CRNN(1/8, 1223)	ATST + BEATs	median filter	True	0.533	0.772	1.305	-
PDFD-CRNN(1/8, 1233)	ATST + BEATs	median filter	True	0.532	0.772	1.304	-
SE+tfwSE-CRNN	ATST + BEATs	median filter	True	0.525	0.767	1.292	-
PFD-CRNN(1/8)	ATST + BEATs	cSEBBs	True	0.551	0.773	1.324	1
PDFD-CRNN(1/8, 1123)	ATST + BEATs	cSEBBs	True	**0.557**	**0.775**	**1.332**	2
Ensemble 2	ATST + BEATs	median filter	True	**0.537**	0.788	**1.325**	-
Ensemble 3	ATST + BEATs	median filter	True	0.536	**0.789**	**1.325**	-
Ensemble 2	ATST + BEATs	cSEBBs	True	**0.575**	0.788	**1.363**	3
Ensemble 3	ATST + BEATs	cSEBBs	True	0.574	**0.789**	**1.363**	4

refers to class-independent 7-frames-sized median filter.

3.2. Evaluation Metrics

True PSDS1 was used to evaluate SED performance on DESED [16, 17]. While previous DCASE challenge task 4 used two types of PSDS (PSDS1 favoring time localization and PSDS2 favoring accurate classification), only PSDS1 is used in this year as PSDS2 is rather an audio tagging metric [12, 19]. For MAESTRO performance evaluation, MPAUC is used [1]. We optimized the model based on average score of PSDS1 + MPAUC on 4 independent training runs. The scores reported in the table are from the models with best sum scores among 4 independent training runs within each model setting.

4. RESULTS

The results are summarized in Table 2, highlighting the performance improvements achieved by our proposed methods. As shown in the results, PFD-CRNN and PDFD-CRNNs do not significantly vary in their performance. However, as their roles differ from each other, ensembling differently dilated PDFD-CRNNs results in decent performance. Likewise, although slightly worse than PDFD-CRNNs, SE-tfwSE-CRNN and PFD-CRNNs with different seeds do help for ensemble. Final best score without ensemble model outperforms the baseline by 15.1% and best score with ensemble outperforms the baseline by 18.2%. While ensemble 1 model slightly outperforms ensemble 2 and 3 those outperformed the baseline by 17.8%, submission was made with latter two as they contain self-trained models thus are expected to retain better generalization capability.

5. CONCLUSION

In this study, we presented Frequency Dependent Networks (FreDNet) for SED. Our method leverages frequency-dependent data augmentation techniques, such as frequency warping and FilterAugment, and incorporates advanced convolutional and transformer-based pre-trained models. Our experiments show that the proposed FreDNet architecture, when combined with techniques like partial dilated frequency dynamic convolution (PDFD), squeeze-and-excitation (SE), and coarse prediction pooling, significantly improves SED performance. The use of change-detection-based Sound Event Bounding Boxes (cSEBBs) further enhances performance by refining onset and offset predictions. The ensemble models, integrating various FreDNet settings, achieved substantial performance gains over the baseline, with the best ensemble model outperforming the baseline by 18.2%. Our approach shows promise for robust SED in diverse environments, highlighting the effectiveness of frequency-dependent methods and the importance of ensemble strategies in improving model performance.

6. REFERENCES

[1] 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," *arXiv preprint arXiv:2406.08056*, 2024.

[2] 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*, 2019.

[3] 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*, 2023.

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

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

[6] 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.

[7] H. Nam, S.-H. Kim, D. Min, J. Lee, and Y.-H. Park, "Diversifying and expanding frequency-adaptive convolution kernels for sound event detection," *arXiv preprint arXiv:2406.05341*, 2024.

[8] H. Nam and Y.-H. Park, "Pushing the limit of sound event detection with multi-dilated frequency dynamic convolution," *arXiv preprint arXiv:2406.13312*, 2024.

[9] H. Nam, S.-H. Kim, D. Min, and Y.-H. Park, "Frequency & channel attention for computationally efficient sound event detection," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2023.

[10] 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 *International Conference on Machine Learning*, 2023.

[11] X. LI and X. Li, "Atst: Audio representation learning with teacher-student transformer," in *Proc. Interspeech*, 2022.

[12] J. Ebbers, F. G. Germain, G. Wichern, and J. L. Roux, "Sound event bounding boxes," *arXiv preprint arXiv:2406.04212*, 2024.

[13] J. F. Gemmeke, D. P. W. 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)*, 2017.

[14] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[15] H. Nam, S.-H. Kim, D. Min, B.-Y. Ko, S.-D. Choi, and Y.-H. Park, "Frequency dependent sound event detection for dcase 2022 challenge task 4," DCASE2022 Challenge, Tech. Rep., 2022.

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

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

[18] J. W. Kim, S. W. Son, Y. Song, H. K. Kim, I. H. Song, and J. E. Lim, "Semi-supervised learning-based sound event detection using frequency dynamic convolution with large kernel attention for DCASE challenge 2023 task 4," DCASE2023 Challenge, Tech. Rep., 2023.

[19] 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," DCASE2021 Challenge, Tech. Rep., 2021.

[20] N. Turpault. Dcase2021 task4 baseline. GitHub. Available: https://github.com/DCASE-REPO/DESED_task. [Online]. Available: https://github.com/DCASE-REPO/DESED_task

[21] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.