CNN-TRANSFORMER WITH SELF-ATTENTION NETWORK FOR SOUND EVENT DETECTION

Keigo Wakayama, Shoichiro Saito

NTT Corporation, Japan

ABSTRACT

In sound event detection (SED), the representation ability of deep neural network (DNN) models must be increased to significantly improve the accuracy or increase the number of classifiable classes. When building large-scale DNN models, a highly parameter-efficient DNN architecture should preferably be adopted. In image recognition, there has been a proposal to replace a convolutional neural network (CNN) extracting high-level features with a highly parameter-efficient DNN architecture, i.e., a self-attention network (SAN). The high-level features are essential information that contributes to prediction. In SED, we find that a model that exceeds the prediction accuracy of CNN-Transformer is difficult to build simply by replacing CNN with SAN, in the process of our experiments. To construct a model with high prediction accuracy while capturing the properties of acoustic signals well, we propose an architecture called a CNN-SAN-Transformer, which retains CNN in the blocks close to the input and uses SAN in all remaining blocks. Experimental results suggest that the proposed method has the same or higher prediction accuracy with a smaller number of parameters than the CNN-Transformer and higher prediction accuracy with a similar number of parameters to the CNN-Transformer and that the proposed method may be a parameter-efficient architecture.

***Index Terms*—** Sound event detection, Weakly-supervised SED, DNN architecture, Self-attention Network, Vector attention

1. INTRODUCTION

Environmental recognition technology is necessary to realize smart cities, smart homes, and autonomous cars, and one such technology is Sound Event Detection (SED), which estimates the correct answer class and the interval where the correct answer class exists from new data.

SED can be achieved by using a Deep Neural Network (DNN) model. To learn the parameters of the DNN model, we have to prepare training data assigned the correct classes and their existence intervals. However, assigning the existence intervals of the correct classes to the data requires a great deal of effort. In recent years, weakly-supervised SED, in which the parameters of the DNN model are learned only from the data with correct answer classes (called weakly-labeled data), has been studied [1–4].

In SED with weakly-labeled data, CNN-Transformer is commonly used to calculate the probability of correct answer classes and their existence intervals. CNN-Transformer combines Convolutional Neural Network (CNN) extracting high-level features from acoustic data and Transformer modeling temporal dependencies. This outperforms the 1st rank system of DCASE 2017 Task 4 [3]. Moreover, the following methods have been proposed to improve the prediction accuracy of SED. First is semi-supervised learning, which uses unlabeled data in addition to weakly-labeled data [5, 6]. Second is semi-supervised learning, which uses artificially generated strongly-labeled data (data with correct classes and their intervals) in addition to these data [7, 8]. Third is a method using source separation as a preprocessing for SED [9]. In these methods, however, the estimation error is large and the number of classifiable classes is limited.

The representation ability of the DNN model must be improved to significantly improve the accuracy of SED or increase the number of classifiable classes. One way to increase the ability of CNN is to increase the kernel size. Since feature aggregation by filters cannot adapt to local contents in the time-frequency direction, there is a limit to the improvement of the ability by increasing the kernel size. Increasing the number of layers and channels can also increase the ability of CNN, but the number of model parameters will be larger than necessary, and the computational complexity of inference will increase. In image recognition, there has been a proposal to replace CNN extracting high-level features with Self-attention Network (SAN) [10–15]. SAN can adapt to the content of the spatial direction in the image, i.e., the local content of the image, and by building SAN with vector attention, it is possible to adapt to the channel direction as well as CNN, thus building a model with high parameter efficiency.

In SED, a DNN architecture with high parameter efficiency should preferably be used when building the models, because large-scale models will be required to significantly improve the prediction accuracy. In this paper, we propose to incorporate SAN into the DNN architecture for SED and evaluate the effectiveness of the proposed method through experiments with datasets.

In the remaining pages, Section 2 describes the weakly-supervised SED, Section 3 describes the SAN, Section 4 describes the proposed method, Section 5 describes the experiments, and the final section summarizes the results.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

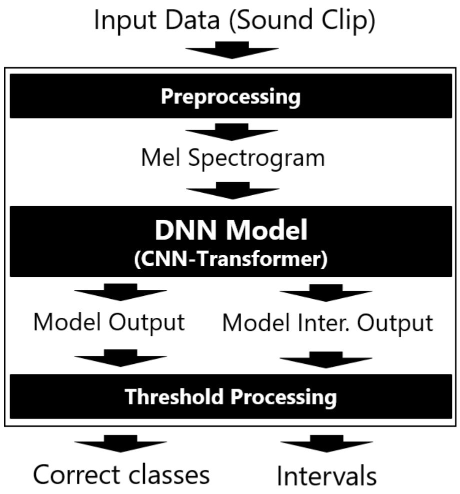

Fig. 1. Sound Event Detection

2. WEAKLY-SUPERVISED SOUND EVENT DETECTION

In weakly-supervised SED, the parameters of the DNN model are trained using weakly-labeled data, and the correct answer classes and their existence intervals are estimated from the new data using the trained DNN model and threshold processing, as shown in Fig.1 [3]. When only the correct answer class is estimated from the new data, it is called Audio Tagging (AT).

Specifically, using a time-frequency representation such as a log-mel spectrogram obtained from the input data, high-level features are extracted by CNN, and the existence probability of each class is predicted for each frame by the encoder in Transformer [16]. In addition, the prediction results for each frame are aggregated to calculate the existence probability of each class for each clip. To train the CNN-Transformer described above, we calculate the Binary Cross Entropy loss function using the predicted results per clip and the correct class of the label.

For SED, as in the case of AT, an arbitrary threshold is applied to the probability of existence of each class predicted for each clip, and if the predicted value is greater than the threshold, each class is estimated to exist in the clip. If each class exists in the clip, it applies two high and low threshold values to the existence probability of each class predicted for each frame and estimates the correct class and their existence interval (start time and end time) in the target clip.

In the process of extracting high-level features, the CNN filter cannot capture the local content of the time-frequency representation, and the number of parameters in the CNN model becomes redundant. To significantly improve the accuracy or to increase the number of classifiable classes, a large-scale DNN model with high representation ability is required, and considering the computational complexity of inference, a DNN architecture with high parameter efficiency should be devised.

3. SELF-ATTENTION NETWORK [12]

Self-attention, which is adaptable to the local content of the data, is expressed by the following equation:

$$\boldsymbol{y}_i = \sum_{j \in \mathcal{R}(i)} \alpha(\boldsymbol{x}_i, \boldsymbol{x}_j) \odot \beta(\boldsymbol{x}_j).$$

Here, $\odot$ is the Hadamard product and i is the subscript of the feature vector $\boldsymbol{x}_i$. Note that the aggregation footprint $\mathcal{R}(i)$ is the set of subscripts that specify which feature vectors are aggregated to construct a new feature $\boldsymbol{y}_i$. In addition, β is a function that generates the feature vector $\beta(\boldsymbol{x}_j)$ that is aggregated by the adaptive weight vector $\alpha(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and α is a function that calculates $\alpha(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for combining $\beta(\boldsymbol{x}_j)$. It is decomposed as follows.

$$\alpha(\boldsymbol{x}_i, \boldsymbol{x}_j) = \gamma(\delta(\boldsymbol{x}_i, \boldsymbol{x}_j)).$$

Here, the relational function δ outputs a vector that represents the relation between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and γ is a function that maps this vector to a vector that can be combined with $\beta(\boldsymbol{x}_j)$. Because there is the function γ, we can use δ to generate vectors of various dimensions that do not need to match the dimensions of $\beta(\boldsymbol{x}_j)$. If δ is a subtraction such as the following, it is a vector attention.

$$\delta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \varphi(\boldsymbol{x}_i) - \psi(\boldsymbol{x}_j).$$

Here, φ and ψ are trainable transformation functions, and the dimensions of their outputs are identical. In the above case, the dimension of $\delta(\boldsymbol{x}_i, \boldsymbol{y}_j)$ is equal to the dimension of the transformation function. Note that if δ is an inner product, it becomes a scalar attention.

4. PROPOSED METHOD

While CNN is adaptable in the channel direction and scalar attention used in the Transformer is adaptable in the spatial direction, SAN based on vector attention, is adaptable in both the channel and spatial directions, enabling us to build a model with high parameter efficiency. However, in the process of conducting experiments, we find that it is difficult to build a model that exceeds the prediction accuracy of CNN-Transformer by simply replacing the CNN of CNN-Transformer used in SED with SAN (details will be discussed in the next section). Therefore, we consider how to build a model with high prediction accuracy while capturing the properties of acoustic signals well. In particular, by leaving CNN only in blocks close to the input, we obtain an intermediate representation independent of the nature of the data from low-level features specific to acoustic signals. Moreover, we replace all the remaining blocks, which account for the majority of the parameter count, with parameter-efficient SAN to efficiently extract high-level features from intermediate representations that are more abstract than the data. As

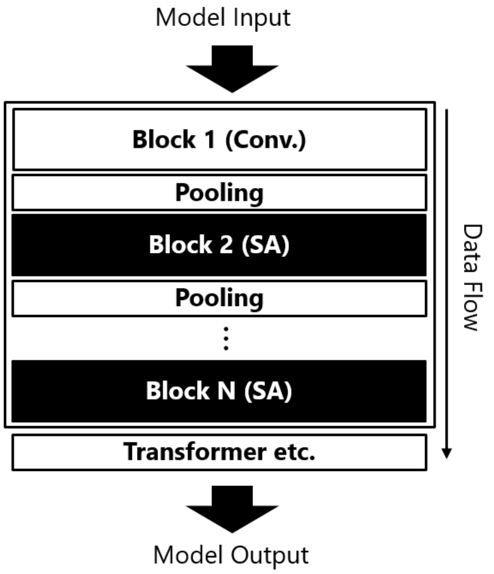

Fig. 2. Proposed CNN-SAN-Transformer

mentioned above, we propose a new DNN architecture, i.e., CNN-SAN-Transformer, as shown in Fig.2.

5. EXPERIMENT

The experiments conducted to determine whether the proposed method can build parameter-efficient DNN models are described below. Specifically, Section 5.1 describes the dataset used in the experiment, Section 5.2 describes the details of feature extraction and training by DNN model as the experimental conditions, Section 5.3 describes the details of evaluation metrics and experimental results, and Section 5.4 describes the experiment in semi-supervised SED.

5.1. Dataset

As benchmark data for weakly-supervised SED, we use the dataset from DCASE 2017 Task 4 "Large-scale weakly supervised sound event detection for smart cars" [17]. We use 51,172 training data and 1,103 evaluation data in the dataset. One clip is 10 seconds and the number of classes is 17. To evaluate AT and SED, not only the correct class but also the interval information (start time and end time) of the correct class is given in the evaluation data.

5.2. Experimental conditions

As the input feature of the DNN model used in the training and estimation of SED, we use the log-mel spectrogram, which is a time-frequency representation of the acoustic signal, as in the experiment of Q. Kong et al. [3]. To calculate the log-mel spectrogram, we first convert the audio clip to a mono signal and resample it to 32 kHz. Next, a short-time Fourier transform with a Hanning window of 1,024 samples and a shift of 320 samples is used to extract the spectrogram. Then, 64 mel filter bank and logarithmic transformation are applied to the spectrogram.

The CNN in the DNN model (CNN-SAN-Transformer) uses one convolution block consisting of two convolution layers with a kernel size of 3×3. Inside the convolution block, a feature transformation with an output channel of 64 is realized. In addition, 2×2 average pooling is applied after the convolution block. The SAN consists of three SA blocks, and each SA block consists of two SA layers with a footprint size of 7×7. Note that 2×2 average pooling is applied after the first two SA blocks, and the transition layer before the SA blocks realizes the feature transformation whose output channels are Y, Z, and 512. The CNN-SAN described above is used to extract high-level features. Then, the values are averaged along the frequency axis of the output to generate a time series signal with 512 channels. Next, the probability of the presence of acoustic events in each frame is predicted by applying the Transformer and other methods to the time series signal. Then, the arithmetic mean is applied as an aggregation function to obtain the prediction results for each clip.

As an optimization algorithm, we use Adam, which has a learning rate of 0.001, as in the experiment by Q. Kong et al. [3]. The learning process stops after 60,000 iterations. After 50,000 iterations, a learning rate of 0.0001 is used. The batch size is 32, and Mixup is used as data augmentation.

5.3. Evaluation metrics and experimental results

To compare the performance of the models, mAP [3] is used as an evaluation metric for AT and SED. A high mAP indicates high performance. To calculate the true positives, false positives, and false negatives required for calculating F1, a threshold value needs to be selected and applied to the output of the model. The AP, on the other hand, is defined as the lower region of the precision-recall curve calculated for multiple thresholds. Therefore, mAP is independent of a threshold value.

As experimental results (a) and (e) in Table 1 show, when CNN is replaced by SAN, i.e., in the case of SAN-Transformer, the mAPs of AT and SED are lower than that of the conventional method (i.e., CNN-Transformer), although the number of parameters is 29 %. In addition, as shown in results (c) and (f) in Table 1, the mAP of AT is lower than and the mAP of SED in SAN-Transformer is almost the same as that in the CNN-Transformer with a similar number of parameters. However, as results (a) and (g) in Table 1 show, the proposed method (i.e., CNN-SAN-Transformer) has 70 % lower parameters than the CNN-Transformer, and the mAP of AT is almost the same and the mAP of SED is higher. As shown by results (b) and (g) in Table 1, the mAPs of AT and SED in the CNN-SAN-Transformer are much higher than that in CNN-Transformer with a similar number of parameters. In addition, as shown by results (c) and (h) in Table 1,

Table 1. Experiment results (weakly-supervised SED)

		CNN-Transformer				SAN-Transformer		CNN-SAN-Transformer	
		(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
Block [Layer]@ Channel		Conv.[2]@ 64 Conv.[2]@128 Conv.[2]@256 Conv.[2]@512	Conv.[2]@ 64 Conv.[1]@ 96 Conv.[1]@ 96 Conv.[1]@512	Conv.[2]@ 64 Conv.[1]@128 Conv.[1]@256 Conv.[1]@512	Conv.[2]@ 64 Conv.[2]@512 Conv.[2]@512 Conv.[2]@512	SA[2]@ 64 SA[2]@128 SA[2]@256 SA[2]@512	SA[2]@ 64 SA[2]@512 SA[2]@512 SA[2]@512	Conv.[2]@64 SA[2]@ 128 SA[2]@ 256 SA[2]@ 512	Conv.[2]@64 SA[2]@ 512 SA[2]@ 512 SA[2]@ 512
Params		5.75M (100%)	**1.68M (29%)**	**2.65M (46%)**	13.19M (229%)	**1.69M (29%)**	**2.64M (46%)**	**1.71M (30%)**	**2.66M (46%)**
mAP	AT	0.658	0.635	0.653	**0.659**	0.604	0.604	0.657	**0.661**
	SED	0.452	0.436	0.435	**0.462**	0.422	0.438	**0.461**	**0.476**

the mAP of AT in CNN-SAN-Transformer is higher than that in CNN-Transformer with a similar number of parameters, and the mAP of SED is much higher. As shown by results (d) and (h) in Table 1, the mAP of AT is almost the same and the mAP of SED is higher in CNN-SAN-Transformer than in CNN-Transformer, although the number of parameters is significantly low. Those mAPs are the best scores in DCASE 2017 Task 4 as far as we know.

Considering the above results, the proposed CNN-SAN-Transformer has the same or higher prediction accuracy with a smaller number of parameters than the CNN-Transformer, and the proposed method has higher prediction accuracy than the CNN-Transformer with a similar number of parameters. Those experimental results suggest that the proposed method may be a highly parameter-efficient architecture.

5.4. Experiment in semi-supervised SED

Experiments were also conducted to evaluate the performance of the proposed method in semi-supervised SED. As benchmark data for semi-supervised SED, we use the dataset from DCASE 2021 Task 4 "Sound Event Detection and Separation in Domestic Environments". We use 1,578 weakly-labeled data and 14,412 unlabeled data and 10,000 strongly-labeled data as training data and 1,168 strongly-labeled data as evaluation data in the dataset. The number of classes is 10. We use Cross-Referencing Self-Training (CRST) [18, 19] which is a state-of-the-art semi-supervised SED method, and evaluate SED performance with PSDS1 and PSDS2, which are the evaluation metrics in the DCASE 2021 Task 4. Those metrics are independent of a threshold value because of being calculated with multiple thresholds, like mAP [20].

As shown by experimental results (a) and (d) in Table 2, the CNN-SAN-Transformer has higher PSDS1 and PSDS2 with a smaller number of parameters than the CNN-Transformer. As shown by experimental results (b) and (e) in Table 2, the CNN-SAN-Transformer has almost the same PSDS1 and higher PSDS2 with the same number of parameters as the CNN-Transformer. As shown by results (c) and (f) in Table 2, the CNN-SAN-Transformer has almost the same PSDS1 and PSDS2 with a smaller number of parameters than the CNN-Transformer.

Table 2. Experiment results (semi-supervised SED)

		CNN-Transformer		
		(a)	(b)	(c)
Block [Layer]@ Channel		Conv.[1]@ 16 Conv.[1]@ 32 Conv.[1]@ 32 Conv.[1]@128	Conv.[1]@ 16 Conv.[1]@ 32 Conv.[1]@ 32 Conv.[1]@ 32 Conv.[1]@ 64 Conv.[1]@ 64 Conv.[1]@128	Conv.[1]@ 16 Conv.[1]@ 32 Conv.[1]@ 32 Conv.[1]@ 32 Conv.[1]@ 90 Conv.[1]@128 Conv.[1]@128
Params		2.0M	2.3M	3.0M
PSDS1	Student	0.285	0.283	0.278
	Teacher	0.296	0.290	0.277
PSDS2	Student	0.490	0.517	0.546
	Teacher	0.502	0.524	0.561
		CNN-SAN-Transformer		
		(d)	(e)	(f)
Block [Layer]@ Channel		Conv.[1]@16 Conv.[1]@32 SA[2]@ 64 SA[2]@ 128 SA[2]@ 128 SA[2]@ 128 SA[2]@ 128	Conv.[1]@16 Conv.[1]@32 SA[2]@ 64 SA[2]@ 128 SA[2]@ 256 SA[2]@ 256 SA[2]@ 128	Conv.[1]@16 Conv.[1]@32 SA[2]@ 64 SA[2]@ 128 SA[2]@ 384 SA[2]@ 384 SA[2]@ 128
Params		1.8M	2.3M	2.9M
PSDS1	Student	0.300	0.292	0.273
	Teacher	0.309	0.286	0.288
PSDS2	Student	0.500	0.550	0.542
	Teacher	0.520	0.552	0.550

6. CONCLUSION

In this paper, we proposed DNN architecture for SED that effectively incorporates a parameter-efficient Self-attention Network (SAN), namely CNN-SAN-Transformer, which is a combination of CNN that obtains intermediate representations from low-level time-frequency features, SAN that can capture local contents in the time-frequency direction from intermediate representations, and a Transformer that can capture global dependencies in the time-direction. We also demonstrated the effectiveness of the proposed method through experiments using the datasets of weakly-supervised SED, etc. Future work includes further improving the prediction accuracy by improving the DNN architecture.

7. REFERENCES

[1] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, “Weakly-Supervised Sound Event Detection with Self-Attention,” Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, May, 2020, pp. 66–70.

[2] X. Xia, R. Togneri, F. Sohel, Y. Zhao, and D. D. Huang, “Sound Event Detection Using Multiple Optimized Kernels,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, 2020, pp. 1745–1754.

[3] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, “Sound event detection of weakly labelled data with CNN-transformer and automatic threshold optimization,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, 2020, pp. 2450–2460.

[4] H. Dinkel, M. Wu, and K. Yu, “Towards Duration Robust Weakly Supervised Sound Event Detection,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, 2021, pp. 887–900.

[5] L. JiaKai, “Mean teacher convolution system for DCASE 2018 Task 4,” Detection and Classification of Acoustic Scenes and Events Challenge, Sep., 2018.

[6] L. Lin, X. Wang, H. Liu, and Y. Qian, “Guided Learning for Weakly-Labeled Semi-Supervised Sound Event Detection,” Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Spain, May, 2020, pp. 626–630.

[7] N. Turpault, R. Serizel, A. Shah, and J. Salamon, “Sound event detection in domestic environments with weakly labeled data and soundscape synthesis,” Proc. Workshop on Detection and Classification of Acoustic Scenes and Events, USA, Oct., 2019, pp. 253–257.

[8] L. Lin, X. Wang, H. Liu, and Y. Qian, “Guided learning convolution system for DCASE 2019 Task 4,” Proc. Workshop on Detection and Classification of Acoustic Scenes and Events, USA, Oct., 2019, pp. 134–138.

[9] N. Turpault, S. Wisdom, H. Erdogan, J. R. Hershey, R. Serizel, E. Fonseca, P. Seetharaman, and J. Salamon, “Improving sound event detection in domestic environments using sound separation,” Proc. Workshop on Detection and Classification of Acoustic Scenes and Events, Japan, Nov., 2020, pp. 205–209.

[10] H. Hu, Z. Zhang, Z. Xie, and S. Lin, “Local Relation Networks for Image Recognition,” Proc. IEEE/CVF International Conference on Computer Vision, USA, Korea, Nov., 2019, pp. 3464-3473.

[11] P. Ramachandran, N. Plarmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, “Stand-Alone Self-Attention in Vision Models,” Proc. 33rd Conference on Neural Information Processing System, Canada, Dec., 2019.

[12] H. Zhao, J. Jia, and V. Koltun, “Exploring Self-attention for Image Recognition,” Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, Jun., 2020, pp. 10076-10085.

[13] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L. C. Chen, “Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation,” Proc. 16th European Conference on Computer Vision, Online, Aug., 2020.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Proc. International Conference on Learning Representation, Austria, May, 2021.

[15] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and Li Zhang, “Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers,” Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, Jun., 2021, pp. 6881-6890.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” Proc. Thirty-first Conference on Neural Information Processing Systems, USA, Dec., 2017.

[17] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, “DCASE 2017 challenge setup: Tasks, datasets and baseline system,” Proc. Workshop on Detection and Classification of Acoustic Scenes and Events, Germany, Nov., 2017, pp. 85–92.

[18] S. Park, A. Bellur, D. K. Han, and M. Elhilali, “Self-Training for Sound Event Detection in Audio Mixtures,” Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, May, 2020, pp. 341–345.

[19] S. Park, D. K. Han, and M. Elhilali, “Cross-Referencing Self-Training Network for Sound Event Detection in Audio Mixtures,” arXiv, May, 2021, pp. 1–13.

[20] S. Park, W. Choi, and M. Elhilali, “Sound Event Detection with Cross-Referencing Self-Training,” Detection and Classification of Acoustic Scenes and Events Challenge, Jun., 2021.