SEMI-SUPERVISED SOUND EVENT DETECTION USING MULTISCALE CHANNEL ATTENTION AND MULTIPLE CONSISTENCY TRAINING

Yih-Wen Wang[1], *Chia-Ping Chen*[1], *Chung-Li Lu*[2], *Bo-Cheng Chan*[2]

[1]National Sun Yat-Sen University, Taiwan, m083040011@student.nsysu.edu.tw, cpchen@cse.nsysu.edu.tw
[2]Chunghwa Telecom Laboratories, Taiwan, {chungli,cbc}@cht.com.tw

ABSTRACT

We present a neural network-based sound event detection system that outputs sound events and their time boundaries in audio signals. The network can be trained efficiently with an amount of strongly labeled synthetic data and weakly labeled or unlabeled real data. Based on the mean-teacher framework of semi-supervised learning with RNNs and Transformer, the proposed system employs multi-scale CNNs with efficient channel attention, which can capture the various features and pay more attention to the important area of features. The model parameters are learned with multiple consistency criteria, including interpolation consistency, shift consistency, and clip-level consistency, to improve the generalization and representation power. For different evaluation scenarios, we explore different pooling functions and search for the best layer. To further improve the performance, we use data augmentation and posterior-level score fusion. We demonstrate the performance of our proposed method through experimental evaluation using the DCASE2021 Task4 dataset. On the validation set, our ensemble system achieves the PSDS-scenario1 of 40.72% and PSDS-scenario2 of 80.80%, significantly outperforming that of the baseline score of 34.2% and 52.7%, respectively. On the DCASE2021 challenge's evaluation set, our ensemble system is ranking 7 among the 28 teams and ranking 14 among the 80 submissions.

Index Terms— sound event detection, Transformer, channel attention, semi-supervised learning, consistency training

1. INTRODUCTION

Sound event detection (SED) is a useful technique for helping us what is happening in an environment by identifying sounds [1, 2, 3]. SED predicts not only the sound event types in an audio recording but also the corresponding onset and offset times. Recently, Detection and Classification of Acoustic Scenes and Events (DCASE) promotes researches on sound detection and classification by annual workshops and challenges. To learn less from human annotation and more from data, DCASE 2021 Task 4 [4] proposes semi-supervised learning to explore the possibility of learning SED with the data of strongly labeled, weakly labeled, and unlabeled. Furthermore, DCASE proposed two evaluation metrics: PSDS-scenario 1 (PSDS 1) requires that SED system needs to react fast upon an event detection; PSDS-scenario 2 (PSDS 2) requires that SED system must avoid confusion between classes but the reaction time is less crucial than in the previous scenario.

One well-known semi-supervised learning approach is to train CRNN [5] with the mean-teacher framework [6]. CRNN utilizes CNNs to extract the short-term and local information and RNNs to capture the long-term contextual information. The mean-teacher framework exploits consistency regularization to stabilize the classifier output for unlabeled data or weakly-labeled data. Besides, the transformer architecture [7] can extract global information while reducing the high computational cost of RNN and achieve state-of-the-art performance on multiple tasks, such as speech recognition [8], speaker recognition [9], speaker diarization [10], text-to-speech [11], audio tagging [12], and sound event detection [13].

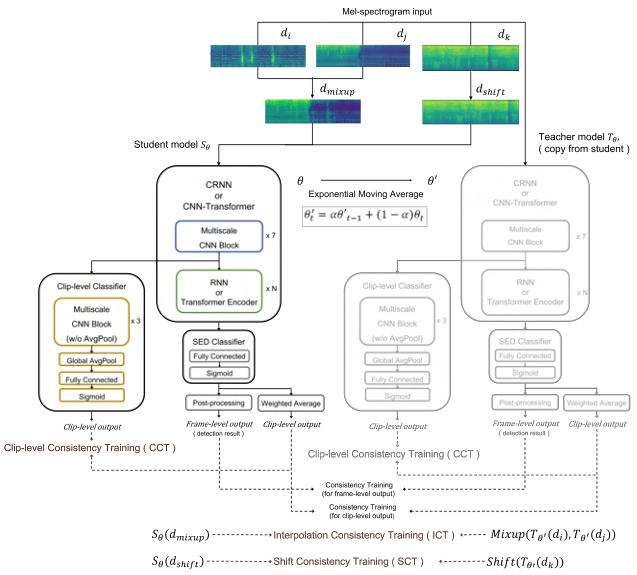

Figure 1: Overview of our proposed system. With the multi-scale CNNs and ECA-net based on RNNs/Transformer network, learning of the mean-teacher framework is enhanced with multiple objectives. ICT/SCT encourages the prediction of interpolated/time-shifted data to be consistent with the interpolated/time-shifted prediction. CCT encourages the origin output consistent with the clip-level classifier output. d_i, d_j, d_k: the original data points; $d_{\texttt{mixup}}$: the mixture of d_i and d_j; $d_{\texttt{shift}}$: time-shift of d_k; $S_\theta, T_{\theta'}$: the student and teacher model.

In this paper, we first explore the performance of RNNs-based and Transformer-based neural networks for two evaluation metrics, PSDS 1 and PSDS 2. Then, since the length of sound events is very different so that we apply the multi-scale CNNs [14] with efficient channel attention (ECA-Net) [15] to capture the more various and important features. Meanwhile, we extend the consistency criteria for model training in mean-teacher framework to include interpolation consistency (ICT) [16], shift consistency (SCT) [17], and clip-level consistency (CCT) [18]. In addition, we apply data augmen-

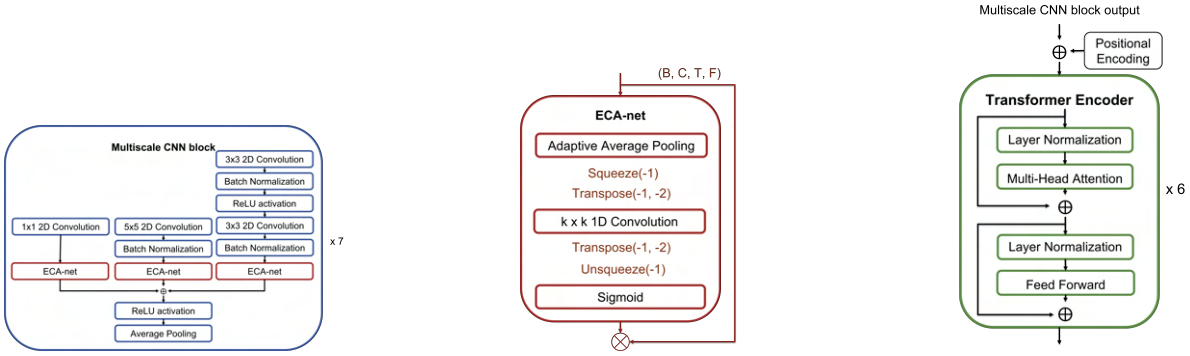

Figure 2: From left to right, the network architecture of multi-scale CNN block, efficient channel attention network (ECA-Net), and Transformer encoder block.

tation and posterior-level score fusion to further improve the model performance. Finally, on the validation set and public evaluation set of DCASE 2021 Task4, our proposed system both outperform the baseline system.

The rest of the paper is organized as follows. In Section 2, we introduce the proposed network architecture, multiple consistency schemes, data augmentation, and posterior-level score fusion to improve the SED system. Section 3 describes the dataset, audio preprocessing, and training setups. Section 4 presents the experimental results and analysis. Finally, we draw conclusion in Section 5.

2. PROPOSED METHODS

2.1. Network architecture

2.1.1. Multi-scale CRNN / CNN-Transformer

From strongly labeled training data, we estimate duration of each sound event as below. 0∼2s: `alarm/bell/ringing`, `cat`, `dishes`, `dog`, and `speech`. 4∼6s: `blender` and `running water`. 7∼10s: `electric shaver/toothbrush`, `frying`, and `vacuum cleaner`. The length of sound events is various and cause the model to work with inconsistent accuracy for the event of different scales. Thus, we refer to [14] to build a multi-scale CNN block to capture the richer features, which contains the kernel size of 1x1, 3x3, 5x5 and uses addition to integrate features of different scales, as shown in the left of Figure 2. In 7 layers of multi-scale CNN block, we also utilize batch normalization and ReLU activation to speed up and stabilize training, each of which attaches an average-pooling layer to calculate the average for each patch of the feature map and downsample feature dimensions along both the time axis and the frequency axis.

To obtain the long-term contextual information, we use the RNNs and transformer encoder to form CRNN [19, 5] and CNN-Transformer [20, 13]. RNNs are applied to two layers of bi-directional gated recurrent unit (GRU) like DCASE 2021 baseline. The network architecture of the transformer encoder is as shown in the right of Figure 2. Positional encoding is used to enhance the output features from the multi-scale CNN blocks with order information before the transformer blocks. A transformer encoder block has layer normalization, multi-head attention, and feed-forward layer. The multi-head attention estimates the similarity between query and key and extracts value as a weighted sum. The mechanism allows the model to jointly pay attention to the information from different positions. The fully connected feed-forward layer with ReLU activation is applied to each position identically. For regularization, we adopt pre-layer normalization (Pre-LN) [21] and residual connection. Finally, the SED classifier consists of a fully connected layer and sigmoid function to discriminate the sound event types.

2.1.2. Efficient Channel Attention

The effect of the acoustic feature extraction largely determines the model ability to predict different sound events and affects the final classification result. However, the attention mechanism can make the model pay more attention to areas which may be important features, and improve the model ability to distinguish features of sound events. We combine the efficient channel attention network (ECA-Net) [15] in multi-scale CNN blocks before adding features of different scales, as shown in the left of Figure 2. ECA-Net is composed of adaptive average pooling (A-AvgPool) layer, 1D convolutional (1D-CNN) layer, and sigmoid function, as shown in the middle of Figure 2. A-Avgpool is applied along the time axis and 1D-CNN calculate the attention of each channel. The kernel size of 1D-CNN is defined by

$$k = \left| \frac{\log_2(C) + b}{\gamma} \right|_{odd} \quad (1)$$

where k and C denote kernel size and channel dimensional, γ and b are set to 2. Clearly, high-dimensional channels have longer range interaction, vice versa.

2.1.3. Pooling Function

Wang et al. [22] compared five different types of pooling functions in the multiple instance learning (MIL) framework for SED, namely attention pooling, max pooling, average pooling, linear softmax, and exponential softmax. The formula of each pooling function is presented in Table 2. The attention pooling estimates the weights for each frame are learned with a dense layer in the network. The max pooling simply take the large probability in all frames. The average pooling assigns an equal weight for all frames. The linear softmax assigns weights equal to the frame-level probability, while the exponential softmax assigns a weight of exponential to the frame-level probability. DCASE 2021 Task4 baseline [5] uses attention pooling to transform frame-level into clip-level. However, with different evaluation scenarios, there should be a relatively appropriate pooling function to replace.

2.2. Semi-Supervised Learning

We employ the mean-teacher framework for its fast convergence, instead of the Π model [23] or temporal ensembling [24], exploiting consistency regularization to stabilize the classifier output for unlabeled data or weakly-labeled data. In this work, we use Mean Square Error (MSE) loss for the consistency cost:

$$\text{MSE}(y, \hat{y}) = (y - \hat{y})^2, \tag{2}$$

where y and $\hat{y}$ denote the target and the prediction, respectively. Next, we propose multiple consistency criteria to regularize how the SED system should learn from unlabeled or weakly-labeled data.

2.2.1. Interpolation Consistency Training

The interpolation consistency training (ICT) [16] has been proposed for semi-supervised learning. ICT encourages the prediction at an interpolation of unlabeled data points to be consistent with the interpolation of the prediction at these data points. Learning from interpolation samples can help the model discriminate ambiguous samples to improve the generalization ability. We define the ICT loss function by

$$\begin{aligned} L_{ICT} = \text{MSE}(&S_\theta(\lambda d_i + (1-\lambda) d_j), \\ &\lambda T_{\theta'}(d_i) + (1-\lambda) T_{\theta'}(d_j)), \end{aligned} \tag{3}$$

where S_θ and $T_{\theta'}$ denote a student model and a teacher model, d_i and d_j denote data points, and λ is randomly sampled from a Beta distribution.

2.2.2. Shift Consistency Training

Inspired by ICT, we consider time-shift as another way to enhance consistency which is similar to proposed by [17], called shift consistency training (SCT). We define the SCT loss function by

$$L_{SCT} = \text{MSE}(S_\theta(\text{shift}(d_k)), \text{shift}(T_{\theta'}(d_k))). \tag{4}$$

SCT encourages the prediction of time-shift input to be consistent with time-shift prediction. In theory, it allows the model to learn shift-invariance and temporal localization of sound events.

2.2.3. Clip-level Consistency Training

In addition to ICT and SCT, we also apply clip-level consistency training (CCT) [18] to enhance the ability to extract the features. We define the CCT loss function by

$$L_{CCT} = \text{MSE}(\text{NN}(d_x), \text{ClipLevel}(f_x)), \tag{5}$$

where $\text{NN}(d_x)$ is the weighted average pooling of the multi-scale CRNN or CNN-Transformer frame-level network output of data d_x, and $\text{ClipLevel}(f_x)$ is obtained by feeding the feature map f_x of the final multi-scale CNN block to a clip-level classifier. As shown in Figure 1, the clip-level classifier consists of 3 extra multi-scale CNN blocks, a global average pooling, and a fully connected layer.

2.2.4. Overall Consistency Training

In summary, the overall loss is

$$L = L_0 + L_{ICT} + L_{SCT} + L_{CCT}, \tag{6}$$

where L_0 denotes the loss without the proposed consistency, namely mean square error for original consistency cost and binary cross-entropy for the supervised cost.

2.3. Data Augmentation

- Mixup [25]. It mixes two randomly selected samples from the original training data and uses λ sampled from Beta distribution to control the strength of interpolation between two samples. The linear interpolation technique can enhance the data diversity and robustness of the network.
- Shift [26]. It shifts a feature sequence on the time axis, and overrun frames are concatenated with the opposite side of the sequence. The usage helps the network learn temporal localization information of the sound event.
- Masks [26]. It creates artificial data by masking a block of consecutive time steps or frequency channels on the mel-spectrogram instead of the raw audio. It can help the network learn the beneficial features to be robust to the partial loss of spectral information or speech segments.

2.4. Posterior-level Score Fusion

To improve generalization performance, we perform score fusion as a model ensemble technique. We utilize different data augmentation methods to build several single systems based on multi-scale CRNN and CNN-Transformer models with different schemes. Then, we average the raw posterior outputs $p(X)$ for inputs X of the multiple models:

$$p_{\text{fusion}}(X) = \frac{1}{N} \sum_{n=1}^{N} p_n(X), \tag{7}$$

where N means the total number of models for our fusion.

3. EXPERIMENTS

3.1. Dataset and Signal Preprocessing

The DESED dataset of DCASE 2021 Task 4 is comprised of 10-sec audio clips and 10 classes of sound events. The data are in two domains: real data (44.1kHz) extracted from AudioSet [27] and synthetic data (16kHz) generated by Scaper [28]. Each audio clip can be strongly labeled with the sound events and their time boundaries annotated, weakly labeled with only the sound events annotated, or unlabeled without any annotation. All dataset is divided into 5 subsets: weakly labeled (1,578 clips), unlabeled (14,412 clips), strongly labeled (10,000 clips), validation set (1,168 clips), public evaluation set (692 clips). Audio signals are resampled to 16kHz sampling rate at first by FFmpeg tool [29]. Then, 128-channel mel-spectrogram from them is extracted with a window size of 2048 and hop size of 256 by Librosa tool [30]. Consequently, the size of the input acoustic features to the deep neural network is 626×128.

3.2. Network Setups

The 7 layers of multi-scale CNN blocks have the number of filters:[16, 32, 64, 128, 128, 128, 128] and pooling size:[[2, 2], [2, 2], [1, 2], [1, 2], [1, 2], [1, 2], [1, 2]]. The 6 layers of transformer encoder blocks have multi-head attention with 256 units and 8 heads and a feed-forward layer with 2048 units. For ICT and mixup augmentation, the parameter λ is sampled from $\text{Beta}(\alpha, \alpha)$ and α from 0.1 to 0.9 in increments of 0.1. For SCT and shift augmentation, we choose the amount of time-shift by sampling from a normal distribution with a zero mean and a standard deviation of 90. For masks augmentation, the size of time-mask and frequency-mask are sampled from a uniform distribution from 0 to 30 and 40, respectively.

4. EVALUATION RESULTS

The evaluation of DCASE 2021 Task4 contains PSDS 1 (react fast) and PSDS 2 (avoid class confusion). From Table 1, we can find that the results of RNNs-based network is better than Transformer-based one, especially on PSDS 2. Then, whatever neural network is CRNN or CNN-Transformer, the incorporation of ICT, SCT, and CCT has significant achievement on two scenarios. The multiple consistency training schemes on CRNN improved PSDS 1 from 34.04% to 37.86%, PSDS 2 from 53.30% to 60.87%, and on CNN-Transformer, PSDS 1 from 33.46% to 37.33%, PSDS 2 from 48.77% to 55.87%. In addition, we observe that multi-scale CNN blocks and ECA-Net can help the model obtain various and important features of sound events so that CRNN can reach 65.54% and CNN-Transformer can reach 61.10% for PSDS 2. From Table 2, both types of neural networks are best when using attention pooling at PSDS 1 and using exponential softmax at PSDS 2. We consider that attention pooling learns weights from the network so that they have a time series relationship. Therefore, it has better performance under stricter evaluation standards with time requirements. Then, exponential softmax uses exponentials as weights to conform to monotonicity so that the higher the prediction probability of the time point, the higher the weight. Thus, it has better performance under the stricter evaluation criteria with category requirements.

We combine CRNN/CNN-Transformer with proposed schemes to build three single systems so that PSDS 1 and PSDS 2 can have the best performance:

- (i) CNN-Transformer + ICT, SCT, CCT, Multiscale
- (ii) CRNN + ICT, SCT, CCT, Multiscale
- (iii) CRNN + ICT, SCT, CCT, Multiscale, ECA, ExpSoftmax

Based on mixup data augmentation following the baseline, we find that (i) and (ii) improve the performance on PSDS 1, and (iii) reach significant achievement on PSDS 2. To ensemble the several systems, we apply several data augmentation methods to build each single system, which includes mixup, time-shift, and time-frequency masks, as shown in Table 3. From Table 4, our fusion systems can achieve 40.72% of PSDS 1 and 80.80% of PSDS 2 on the validation set, 37.42% of PSDS 1 and 69.73% of PSDS 2 on the public evaluation set.

Table 1: Results of different schemes, based on two networks with mixup data augmentation.

Scheme	Model	PSDS 1	PSDS 2
-	CRNN	34.04%	53.30%
	CNN-Transformer	33.46%	48.77%
+ICT	CRNN	36.38%	55.87%
	CNN-Transformer	33.39%	50.07%
+SCT	CRNN	**37.86%**	59.47%
	CNN-Transformer	35.61%	52.01%
+CCT	CRNN	37.64%	60.87%
	CNN-Transformer	**37.33%**	55.87%
+Multiscale	CRNN	37.51%	62.63%
	CNN-Transformer	34.75%	**61.10%**
+ECA-Net	CRNN	34.71%	**65.54%**
	CNN-Transformer	35.13%	60.27%

Table 2: Results of different pooling functions, based on above schemes without ECA. y_i and y means frame-level and clip-level.

Pooling Function	Formula	Model	PSDS 1	PSDS 2
Attention	$y = \frac{\sum_i y_i w_i}{\sum_i w_i}$	CRNN	**37.51%**	62.63%
		CNN-Transformer	**34.75%**	61.10%
Max pooling	$y = \max_i y_i$	CRNN	36.10%	64.59%
		CNN-Transformer	31.73%	59.77%
Average pooling	$y = \frac{1}{n}\sum_i y_i$	CRNN	5.34%	73.95%
		CNN-Transformer	4.53%	60.41%
Linear Softmax	$y = \frac{\sum_i y_i^2}{\sum_i y_i}$	CRNN	26.75%	60.17%
		CNN-Transformer	24.21%	60.57%
Exponential Softmax	$y = \frac{\sum_i y_i \exp(y_i)}{\sum_i \exp(y_i)}$	CRNN	5.82%	**75.35%**
		CNN-Transformer	4.13%	**61.31%**

Table 3: Results of different data augmentations, based on three single systems.

#	Model	Schemes	Data Augmentation	PSDS 1	PSDS 2
0	CRNN	-	Mixup ($\alpha = 0.2$)	34.04%	53.30%
1	CNN-Transformer	ICT, SCT, CCT, Multiscale	Mixup ($\alpha = 0.2$)	**34.75%**	61.10%
2			Shift	31.39%	55.05%
3			Masks	33.24%	59.04%
4			Mixup ($\alpha = 0.2$)+Shift	33.43%	58.68%
5			Mixup ($\alpha = 0.2$)+Masks	34.29%	**61.52%**
6			Shift+Masks	33.64%	55.46%
7	CRNN	ICT, SCT, CCT, Multiscale	Mixup ($\alpha = 0.1$)	37.69%	63.00%
8			Mixup ($\alpha = 0.2$)	37.51%	62.63%
9			Mixup ($\alpha = 0.4$)	36.71%	64.82%
10			Mixup ($\alpha = 0.5$)	36.84%	64.18%
11			Mixup ($\alpha = 0.6$)	36.55%	61.85%
12			Mixup ($\alpha = 0.7$)	36.70%	63.91%
13			Shift	35.71%	61.29%
14			Masks	36.96%	64.84%
15			Mixup ($\alpha = 0.2$)+Shift	37.03%	63.02%
16			Mixup ($\alpha = 0.2$)+Masks	**38.13%**	**65.32%**
17	CRNN	ICT, SCT, CCT, Multiscale, ECA, ExpSoftmax	Mixup ($\alpha = 0.1$)	**6.81%**	75.59%
18			Mixup ($\alpha = 0.2$)	5.71%	76.16%
19			Mixup ($\alpha = 0.7$)	5.37%	**76.29%**
20			Shift	4.46%	72.16%
21			Masks	5.29%	75.07%
22			Mixup ($\alpha = 0.2$)+Shift	5.12%	76.19%
23			Mixup ($\alpha = 0.2$)+Masks	4.82%	75.45%
24			Shift+Masks	4.83%	76.08%

Table 4: Results of the fusion systems on the two testing sets.

#	Model	Schemes	Validation		Public eval	
			PSDS 1	PSDS 2	PSDS 1	PSDS 2
7∼16	CRNN	ICT, SCT, CCT, Multiscale	**40.72%**	70.25%	37.22%	69.47%
17∼24	CRNN	ICT, SCT, CCT, Multiscale, ECA-Net, ExpSoftmax	6.08%	**80.80%**	8.30%	65.39%
1∼16	CRNN CNN-Transformer	ICT, SCT, CCT, Multiscale	38.79%	67.18%	**37.45%**	68.42%
1∼24	CRNN CNN-Transformer	ICT, SCT, CCT, Multiscale, ECA-Net, ExpSoftmax	37.02%	72.42%	33.56%	**69.73%**

5. CONCLUSION

Based on the mean-teacher framework of semi-supervised learning with RNNs and Transformer, we present a multi-scale CNNs with ECA-Net to capture various and important features of sound events. For the multiple consistency criteria, ICT helps the model discriminate the ambiguous samples to enhance the generalization ability, SCT assists the model to learn better temporal information, CCT promotes the model feature representation power. Then, an appropriate pooling function is applied to the specific scenario. The data augmentation and posterior-level score fusion further improve the performance. Finally, on the validation set and challenge's evaluation set, our proposed system significantly outperforms the baseline.

6. REFERENCES

[1] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *2005 IEEE International Conference on Multimedia and Expo*. IEEE, 2005, pp. 1306–1309.

[2] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, "Monitoring activities of daily living in smart homes: Understanding human behavior," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 81–94, 2016.

[3] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[4] http://dcase.community/challenge2021/task-sound-event-detection-and-separation-in-domestic-environments, 2020.

[5] N. Turpault, R. Serizel, J. Salamon, and A. P. Shah, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," 2019.

[6] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *arXiv preprint arXiv:1703.01780*, 2017.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[8] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.

[9] S. V. Katta, S. Umesh, *et al.*, "S-vectors: Speaker embeddings based on transformer's encoder for text-independent speaker verification," *arXiv preprint arXiv:2008.04659*, 2020.

[10] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 296–303.

[11] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.

[12] J. Wang and S. Li, "Self-attention mechanism based system for dcase2018 challenge task1 and task4," *Proc. DCASE Challenge*, pp. 1–5, 2018.

[13] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Weakly-supervised sound event detection with self-attention," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 66–70.

[14] M. Tang, L. Guo, Y. Zhang, W. Yan, and Q. Zhao, "Multi-scale residual crnn with data augmentation for dcase 2020 task 4."

[15] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks, 2020 ieee," in *CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE*, 2020.

[16] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," *arXiv preprint arXiv:1903.03825*, 2019.

[17] C.-Y. Koh, Y.-S. Chen, Y.-W. Liu, and M. R. Bai, "Sound event detection by consistency training and pseudo-labeling with feature-pyramid convolutional recurrent neural networks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 376–380.

[18] L. Yang, J. Hao, Z. Hou, and W. Peng, "Two-stage domain adaptation for sound event detection."

[19] https://github.com/DCASE-REPO/DESED_task/tree/master/recipes/dcase2021_task4_baseline, 2021.

[20] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Conformer-based sound event detection with semi-supervised learning and data augmentation," *dim*, vol. 1, p. 4, 2020.

[21] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu, "On layer normalization in the transformer architecture," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 524–10 533.

[22] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 31–35.

[23] A. Rasmus, H. Valpola, M. Honkala, M. Berglund, and T. Raiko, "Semi-supervised learning with ladder networks," *arXiv preprint arXiv:1507.02672*, 2015.

[24] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

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

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

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

[28] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 86–90.

[29] https://github.com/FFmpeg/FFmpeg, 2020.

[30] https://github.com/librosa/librosa, 2020.