



A Multi-grained based Attention Network for Semi-supervised Sound Event Detection

Ying Hu^{1,2}, Xiujuan Zhu^{1,2}, Yunlong Li^{1,2}, Hao Huang^{1,2}, and Liang He^{2,3}

¹Key Laboratory of signal detection and processing in Xinjiang, China

²School of Information Science and Engineering, Xinjiang University, Urumqi, China

³Department of Electronic Engineering, Tsinghua University, China

huying@xju.edu.cn, xjzhu@stu.xju.edu.cn, liyunlong@stu.xju.edu.cn

Abstract

Sound event detection (SED) is an interesting but challenging task due to the scarcity of data and diverse sound events in real life. This paper presents a multi-grained based attention network (MGA-Net) for semi-supervised sound event detection. To obtain the feature representations related to sound events, a residual hybrid convolution (RH-Conv) block is designed to boost the vanilla convolution's ability to extract the time-frequency features. Moreover, a multi-grained attention (MGA) module is designed to learn temporal resolution features from coarse-level to fine-level. With the MGA module, the network could capture the characteristics of target events with short- or long-duration, resulting in more accurately determining the onset and offset of sound events. Furthermore, to effectively boost the performance of the Mean Teacher (MT) method, a spatial shift (SS) module as a data perturbation mechanism is introduced to increase the diversity of data. Experimental results show that the MGA-Net outperforms the published state-of-the-art competitors, achieving 53.27% and 56.96% event-based macro F1 (EB-F1) score, 0.709 and 0.739 polyphonic sound detection score (PSDS) on the validation and public set respectively.

Index Terms: Sound Event Detection, Semi-supervised Learning, Multi-grained Attention

1. Introduction

Sound event detection (SED) aims to detect the onset and offset of sound events and identify the class of target events. Recently, there has been an increasing interest in semi-supervised SED in the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge Task4¹. Sound event detection has wide applications, including audio surveillance systems [1], monitoring systems [2] and smart homes[3].

In the real world, different sound events exhibit unique patterns reflected in the time-frequency distribution. As a consequence, it is necessary to obtain the effective feature representation related to sound events. Thanks to the development of deep learning approaches, recent advances [4, 5] have led to improved performance in SED task. Several standard convolutional neural network (CNN) blocks were stacked as the feature encoder to generate the high-level feature representations for the SED task [6, 7]. Lu et al. [8] proposed a multi-scale recurrent neural network (RNN) to capture the fine-grained and long-term dependencies of sound events. CNN is good at learning

¹ <https://dcase.community/challenge2019/task-sound-event-detection-in-domestic-environments>.

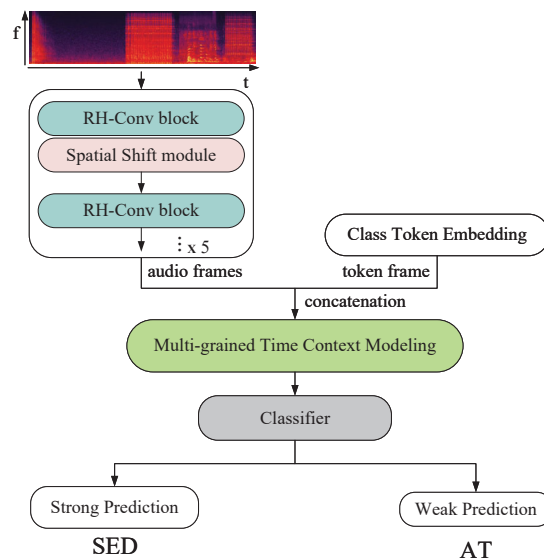


Figure 1: Illustration of the proposed MGA-Net.

features shifted in both time and frequency, while RNN models longer temporal context information.

Convolutional recurrent neural network (CRNN) approaches have shown their superiority in the estimation of onset and offset [9, 10]. For better-integrating information from different time resolutions, Guo et al. [11] proposed multi-scale CRNN to learn coarse or fine-grained temporal features by applying multiple RNNs. Recently, some works [12, 13] also proposed to combine CNN with the self-attention mechanism for the SED task that instead of applying RNN, that self-attention mechanism is used to model temporal context information. To be specific, Miyazaki et al. [12] incorporated the self-attention mechanism of the Transformer in SED to capture global time features and had shown its superior performance in SED. Then they further proposed the Conformer-based SED method [13] to capture both global and local time context information of an audio feature sequence simultaneously.

In addition, similar to [14, 15], Mean Teacher [16] method is adopted to perform semi-supervised learning (SSL) for SED in this paper. Under the cluster assumption that two samples close to each other in the input feature space are likely to belong to the same class [17], some SSL methods [18, 16, 9] introduced a consistency regularization based on perturbation techniques. Data perturbation methods [19, 20] play an essential role in introducing effective perturbation for SSL learning. Zheng et al.

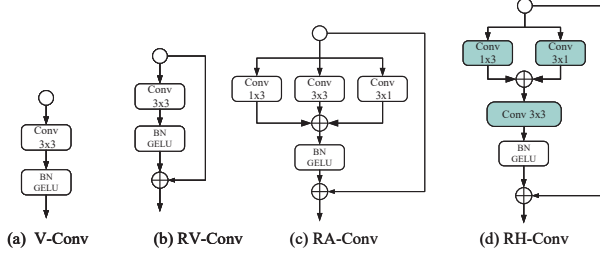


Figure 2: Four kinds of CNN feature extraction blocks. (a) Vanilla convolution block (V-Conv). (b) Residual Vanilla convolution (RV-Conv) block. (c) Residual Asymmetric convolution (RA-Conv) block. (d) Residual Hybrid convolution (RH-Conv) block.

[21] also showed that the MT method could benefit from suitable data and/or model perturbation.

Inspired by the above-mentioned works, we propose a multi-grained based attention network (MGA-Net) in this paper. For the time-frequency feature extraction, we explore four kinds of feature extraction blocks based on CNN and design residual hybrid convolution (RH-Conv) block to boost the representation power of vanilla convolution. We also design a multi-grained based attention (MGA) module to utilize the temporal information. The MGA module builds upon three stages of feature learning: global, local, and frame-level time context modeling. It can capture well the features of temporal resolution from coarse to fine-level. Similar to data augmentation, which can increase the diversity of data, a spatial-shift module is designed as a data perturbation mechanism to bring about data augmentation for the MT method. Experiments on the dataset of DCASE 2020 task4 demonstrate the superiority of our proposed methods.

2. Proposed Method

Our proposed MGA-Net is shown in Fig. 1. It employs six residual hybrid convolution blocks and one spatial shift module to extract time-frequency features, where each residual hybrid block is followed by an average pooling and dropout layer. Then the extracted features are fed into the multi-grained time context modeling to learn the temporal context information. A linear classifier based on a dense layer with sigmoid activation is followed to perform strong label prediction for the SED detection task. Similar to [12], a class token embedding is used to aggregate the whole sequence information that performs weak label prediction for the audio tagging (AT) classification task.

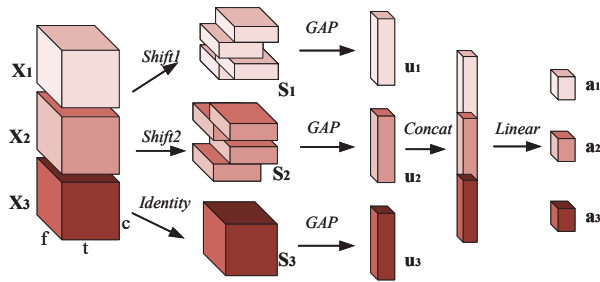


Figure 3: Illustration of the proposed Spatial Shift module.

The following subsections will describe the RH-Conv block, SS module, and MGA module.

2.1. Residual Hybrid Convolution Block

We build four kinds of CNN feature extraction blocks as shown in Fig. 2. Each CNN layer is followed by batch normalization (BN) and gaussian error linear unit (GELU) [22] activation. Fig. 2 (a) is the vanilla CNN with square kernels, i.e., 3×3 , referred to as “V-Conv”. Fig. 2 (b), referred to as “RV-Conv”, introduces identity mapping as the residual connection based on the “V-Conv” block. Fig. 2 (c) can be viewed as asymmetric convolution [23] comprising three parallel CNN layers with 3×3 , 1×3 and 3×1 kernels, respectively, referred to as “RA-Conv”. Fig. 2 (d) is our proposed residual hybrid convolution block, which is a combination of using two parallel CNN layers with 1×3 and 3×1 kernels followed by vanilla convolution with 3×3 kernels. It applies two asymmetric convolution kernels to strengthen the square convolution kernels and is referred to as the “RH-Conv” block. Four kinds of feature extraction blocks are explored with the goal of designing a better CNN structure to extract more robust features related to sound events.

2.2. Spatial Shift Module

To provide a data perturbation mechanism for the MT semi-supervised method, we design a spatial shift module. It firstly conducts the spatial-shift operation, which is proposed by [24], helping to increase the diversity of features. And it further evaluates the degree of importance for spatial shift operation by generating the corresponding weights.

Given an input feature map $\mathbf{X} \in R^{C \times T \times F}$, we firstly expand the channels of \mathbf{X} from c to $3c$ by a linear layer. Then the expanded feature map is equally splitted into three parts: $\mathbf{X}_i \in R^{C \times T \times F}$ $i=1, 2, 3$. As shown in Fig. 3, \mathbf{X}_1 and \mathbf{X}_2 are shifted as \mathbf{S}_1 and \mathbf{S}_2 through the *Shift1* and *Shift2* operation, respectively. *Shift1* conducts the shift operations along the time and frequency dimension, respectively, as shown in Equation 1. In contrast, *Shift2* conducts an asymmetric spatial-shift operation with respect to *Shift1* as shown in Equation 2. Thus, they are complementary to each other. \mathbf{X}_3 is just identified as \mathbf{S}_3 . Then, we embed the global information vector by using global average pooling on \mathbf{S}_i . The global vectors $\mathbf{u}_i \in R^{C \times 1 \times 1}$ are concatenated together along the channel dimension. A linear layer is followed to generate weights \mathbf{a}_i , which is used to reweigh \mathbf{S}_i . Then the softmax function is applied on the weights \mathbf{a}_i to limit $\sum_{i=1}^3 \mathbf{a}_i = 1$. In all, the final output $\mathbf{X}_{out} \in R^{C \times T \times F}$ of this module can be writing as $\mathbf{X}_{out} = \sum_{i=1}^3 \mathbf{a}_i \times \mathbf{S}_i$.

$$\begin{aligned}
 \mathbf{X}_1[1:t, :, :c/4] &\leftarrow \mathbf{X}_1[0:t-1, :, :c/4]; \\
 \mathbf{X}_1[0:t-1, :, :c/4:c/2] &\leftarrow \mathbf{X}_1[1:t, :, :c/4:c/2]; \\
 \mathbf{X}_1[:, 1:f, c/2:3c/4] &\leftarrow \mathbf{X}_1[:, 0:f-1, c/2:3c/4]; \\
 \mathbf{X}_1[:, 0:f-1, 3c/4:] &\leftarrow \mathbf{X}_1[:, 1:f, 3c/4:]
 \end{aligned} \tag{1}$$

$$\begin{aligned}
 \mathbf{X}_2[:, 1:f, :c/4] &\leftarrow \mathbf{X}_2[:, 0:f-1, :c/4]; \\
 \mathbf{X}_2[:, 0:f-1, c/4:c/2] &\leftarrow \mathbf{X}_2[:, 1:f, c/4:c/2]; \\
 \mathbf{X}_2[1:t, :, :c/2:3c/4] &\leftarrow \mathbf{X}_2[0:t-1, :, :c/2:3c/4]; \\
 \mathbf{X}_2[0:t-1, :, :3c/4:] &\leftarrow \mathbf{X}_2[1:t, :, :3c/4:]
 \end{aligned} \tag{2}$$

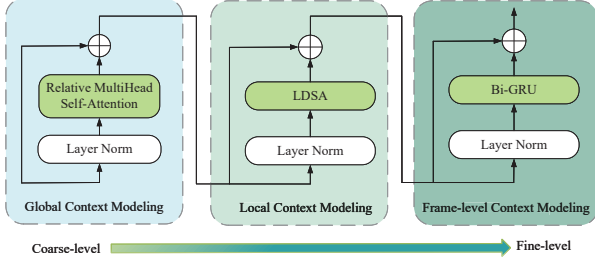


Figure 4: Illustration of the proposed Multi-grained Attention module. The green arrow denotes the time context is modeled from coarse to fine-level, and conversely, is modeled from fine to coarse-level.

2.3. Multi-Grained Attention Module

The multi-grained based attention module is designed to model the temporal context dependencies from coarse-level to fine-level. As shown in Fig. 4, there are three main processes in the multi-grained attention module: Global Context Modeling, Local Context Modeling, Frame-level Context Modeling. We also add residual connection and layer normalization (*LN*) operation at each modeling process.

2.3.1. Global Context Modeling

The global context modeling is built upon the multi-head self-attention mechanism [25]. Considering the sequential position of input features, we introduce relative positional encoding (RPE) [26] which has been shown effective in SED task [27] to encode position information of inter-frames. The length of attention weights is that of the entire time series, making the feature representation more global but coarser. Assuming the input sequence is $\mathbf{X} \in R^{T \times d}$, the global context modeling can be written as:

$$\mathbf{X}_{global} = RA(LN(\mathbf{X})) + \mathbf{X} \quad (3)$$

Where *RA* denotes the multi-head self-attention with relative positional encoding and *LN* the layer normalization.

2.3.2. Local Context Modeling

Local context modeling is designed to capture the local time dependencies within specific time frames rather than the entire time series, complementing the global context modeling. We use local dense synthesizer attention (LDSA) [28] to achieve local context modeling. The local context modeling is expressed as follows:

$$\mathbf{X}_{local} = LDSA(LN(\mathbf{X}_{global})) + \mathbf{X}_{global} \quad (4)$$

The LDSA firstly defines a context window *c* which restricts the attention scope to a local range around the current central frame. Attention weights of the other frames outside the context width are set to 0. *c* is set to 3 in our experiment. The current frame is restricted to only interact with its finite neighbouring frames, thus, achieving the learning of local features. The process of LDSA is calculated as follows:

$$A(\mathbf{X}_{global}) = Softmax(\sigma(\mathbf{X}_{global} \mathbf{W}_1) \mathbf{W}_2) \quad (5)$$

$$\mathbf{V} = \mathbf{X}_{global} \mathbf{W}_3 \quad (6)$$

where $\mathbf{W}_1 \in R^{d \times d}$, $\mathbf{W}_2 \in R^{d \times c}$ and $\mathbf{W}_3 \in R^{d \times d}$ are learnable weights.

Then it assigns the attention weights to the current frame and its neighboring frames:

$$\mathbf{Y}_t = \sum_{j=0}^{c-1} A_{(t,j)} (\mathbf{X}_{global}) \mathbf{V}_{t+j-\lfloor c/2 \rfloor} \quad (7)$$

Thus, the finally output of LDSA is obtained by:

$$LDSA(X) = [\mathbf{Y}_0, \dots, \mathbf{Y}_t, \dots, \mathbf{Y}_T] \mathbf{W}^o \quad (8)$$

where the $\mathbf{W}^o \in R^{d \times d}$ is learnable weight.

2.3.3. Frame-level Context Modeling

No matter the global or local context modeling, the close correlation among time frames is lacking. Thus, we introduce frame-level context modeling to learn the fine-grained inter-frame features. Compared to the self-attention mechanism, RNN can directly model the sequential information naturally present in a sequence of frames. We use Bi-GRU to perform frame-by-frame detection and capture the long-term context dependencies for both past and future frames of the time series. The calculation process is as follows:

$$\mathbf{X}_{frame} = Linear(\sigma(BiGRU(LN(\mathbf{X}_{local}))) + \mathbf{X}_{local}) \quad (9)$$

Where the σ denotes ReLU activation function.

3. Experiment Setup

3.1. Dataset

The experiments in this paper were conducted on the dataset of task 4 in the DCASE2020. It has ten classes of sound events from the domestic environment. The dataset contains three types of training data: weakly labeled data (1502 clips), unlabeled data (13723 clips), and strongly labeled data (2584 clips). We evaluate the performance of the SED network on the validation (1083 clips) and public (692 clips) set.

The input features were Log-Mel spectrograms extracted from the 10-sec audio clips resampled to 16000 Hz. The Log-Mel spectrogram was computed over 1024-point STFT windows with a hop size of 323 samples and 64 Mel-scale filters, resulting in an input feature matrix with 496 frames and 64 Mel-scale filters. More details of preprocessing and post-processing schemes used in our experiments were consistent with that setting in [13].

3.2. Experimental settings

Our proposed MGA-Net was trained using the RAdam optimizer [29], where the initial learning rate was set to 0.001. The size of the average pooling layer is set to 2×2 in the first two layers and 1×2 in the rest layers. The dropout rate was 0.1. In the multi-grained time context modeling, we applied 4 multi-grained attention modules, in which the dimension of features *d* was set to 144, the number of attention heads 4, and the hidden size of the Bi-GRU 512. The loss function is a weighted sum of the classification and consistency losses. The classification loss based on binary cross-entropy (BCE) is calculated by the predictions and the ground truth, while the consistency loss is based on the mean squared error (MSE) between the outputs of student and teacher network. Event-based macro F1 (EB-F1) [30] and polyphonic sound detection score (PSDS) [31] are used as the main evaluation metrics.

Table 1: Performance comparison between the proposed MGA-Net and the state-of-the-art SED methods. SS denotes the spatial shift module.

Method	Validation		Public	
	EB-F1	PSDS	EB-F1	PSDS
Conformer-SED [13]	47.70	0.637	49.00	0.681
ESA-Net [27]	47.80	0.688	52.10	0.712
MGA-Net(Coarse-Fine)	53.27	0.709	56.96	0.739
-SS	52.43	0.705	56.48	0.737

4. Results and Discussion

To investigate the effectiveness of the proposed MGA-Net, we compare it with the state-of-the-art methods [13, 27]. As shown in Table 1, the MGA-Net achieves 53.27%, and 56.96% EB-F1 score, 0.709 and 0.739 PSDS score for the validation and public set, respectively, significantly outperforming the compared methods. In addition, by removing the spatial shift (SS) module, the network performance degrades slightly on both datasets. This result shows that the SS module can help increase the diversity of features.

In the following subsections, we further verify the feature extraction capability of the RH-Conv block by comparing it with the other three kinds of feature extraction blocks and then evaluate the multi-grained attention (MGA) module.

4.1. Comparison Among Four Kinds of CNN Blocks

Table 2 shows the performance of MGA-Net with four different CNN feature extraction blocks introduced in Section 2.1. The “RV-Conv” can achieve better performance compared with “V-Conv”. This may be because introducing residual connection can preserve more of the original features, resulting in a better performance. Compared with “RV-Conv”, “RH-Conv” can achieve better performance. It reveals that the combination of CNNs with 1×3 and 3×1 kernels could enhance the feature extraction capability compared with vanilla CNN, especially when serially using asymmetric convolution (1×3 , 3×1) and 3×3 convolution. Finally, compared with “V-Conv”, the performance on both datasets are increased significantly when the network adopted “RH-Conv”. Especially when focusing on the EB-F1 score, the performance is improved by 1.21% on the validation and 1.6% public set.

4.2. Evaluation of Multi-grained Attention Module

We also investigated the effectiveness of the proposed multi-grained attention module, as shown in Table 3. We firstly explore the feature learning patterns from coarse-level to fine-level

Table 2: Comparison among four kinds of feature extraction blocks.

Method	Validation		Public	
	EB-F1	PSDS	EB-F1	PSDS
V-Conv	51.22	0.690	54.88	0.728
RV-Conv	52.31	0.703	55.31	0.728
RA-Conv	52.08	0.698	56.18	0.726
RH-Conv	52.43	0.705	56.48	0.737

Table 3: Evaluation of the multi-grained attention module.

Method	Validation		Public	
	EB-F1	PSDS	EB-F1	PSDS
MGA-Net(Fine-Coarse)	53.09	0.709	56.48	0.738
MGA-Net(Coarse-Fine)	53.27	0.709	56.96	0.739
-Global	52.93	0.711	56.78	0.748
-Local	51.91	0.705	55.59	0.734
-Global-Local	50.69	0.696	54.95	0.738
-Frame level	50.45	0.698	53.60	0.730

(Coarse-Fine) and from fine-level to coarse-level (Fine-Coarse), as shown in Fig. 4. The results show that the feature learning pattern from coarse-level to fine-level is slightly better than that from fine-level to coarse-level. Therefore, we adopt the Coarse-Fine feature learning pattern in the following experiments.

We then investigated how much the proposed global/local or frame-level context modeling contributes to the MGA-Net. As shown in Table 3, when the global context modeling is removed, the performance of SED is only slightly decreased on the EB-F1 metric. When the local context modeling is removed, the performances on both datasets are all decreased. It seems that local context modeling plays a more critical role than global context modeling in time context modeling. When both the global and local context modeling is removed, only frame-level context modeling is used to extract the fine temporal information, the performance on both datasets is further decreased. Results reveal that it is necessary to first conduct the global context modeling before the local context modeling. In particular, the EB-F1 score is decreased by 2.4% on the validation and by 2.8% on the public set. It also demonstrates that global and local context modeling plays a vital role in capturing event-specific onset and offset information. When the frame-level context modeling is removed while preserving the global and local context modeling, we can see that the performance on both datasets is all decreased.

5. Conclusions

In this paper, we propose a multi-grained attention network for sound event detection. Four kinds of CNN feature extraction blocks are investigated, and the RH-Conv block has shown it superior to the vanilla CNN block in obtaining features related to the sound events. The spatial shift (SS) module provides a data perturbation and shows its effect on increasing features’ diversity. In addition, a multi-grained attention (MGA) module is designed to progressively model the time context information from coarse-level to fine-level. Ablation experiments show that a better performance can be achieved when combining the global, local, and frame-level modeling, clearly demonstrating the effectiveness of the proposed method. In the future, we hope to design more effective feature extraction structures to improve sound event detection performance.

6. Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC) (U1903213), Tianshan Innovation Team Plan Project of Xinjiang (202101642)

7. References

- [1] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2007, pp. 21–26.
- [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] N. Turpault, R. Serizel, J. Salamon, and A. P. Shah, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," 2019.
- [4] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 6440–6444.
- [5] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.
- [6] L. Lin, X. Wang, H. Liu, and Y. Qian, "Specialized decision surface and disentangled feature for weakly-supervised polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1466–1478, 2020.
- [7] Y. Huang, X. Wang, L. Lin, H. Liu, and Y. Qian, "Multi-branch learning for weakly-labeled sound event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 641–645.
- [8] R. Lu, Z. Duan, and C. Zhang, "Multi-scale recurrent neural network for sound event detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 131–135.
- [9] J. Yan, Y. Song, L.-R. Dai, and I. McLoughlin, "Task-aware mean teacher method for large scale weakly labeled semi-supervised sound event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 326–330.
- [10] 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, pp. 887–900, 2021.
- [11] Y. Guo, M. Xu, Z. Wu, J. Wu, and B. Su, "Multi-scale convolutional recurrent neural network with ensemble method for weakly labeled sound event detection," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 2019, pp. 1–5.
- [12] 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.
- [13] 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," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020, pp. 100–104.
- [14] L. JiaKai, "Mean teacher convolution system for dcase 2018 task 4," *Detection and Classification of Acoustic Scenes and Events*, 2018.
- [15] N. Turpault and R. Serizel, "Training sound event detection on a heterogeneous dataset," *arXiv preprint arXiv:2007.03931*, 2020.
- [16] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.
- [17] Y. Luo, J. Zhu, M. Li, Y. Ren, and B. Zhang, "Smooth neighbors on teacher graphs for semi-supervised learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8896–8905.
- [18] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.
- [19] 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.
- [20] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.
- [21] X. Zheng, Y. Song, J. Yan, L.-R. Dai, I. McLoughlin, and L. Liu, "An effective perturbation based semi-supervised learning method for sound event detection," in *INTERSPEECH*, 2020, pp. 841–845.
- [22] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.
- [23] X. Ding, Y. Guo, G. Ding, and J. Han, "Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1911–1920.
- [24] T. Yu, X. Li, Y. Cai, M. Sun, and P. Li, "S2-mlpv2: Improved spatial-shift mlp architecture for vision," *ArXiv*, vol. abs/2108.01072, 2021.
- [25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [26] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.
- [27] H. Sundar, M. Sun, and C. Wang, "Event specific attention for polyphonic sound event detection," in *Interspeech*, 2021.
- [28] M. Xu, S. Li, and X.-L. Zhang, "Transformer-based end-to-end speech recognition with local dense synthesizer attention," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5899–5903.
- [29] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," *arXiv e-prints*, pp. arXiv–1908, 2019.
- [30] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.
- [31] Č. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 61–65.