INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Joint-Former: Jointly Regularized and Locally Down-sampled Conformer for Semi-supervised Sound Event Detection

Lijian Gao[1], *Qirong Mao*[1,*], *Ming Dong*[2]

[1]School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China
[2]Department of Computer Science, Wayne State University, Detroit, USA

ljgao@stmail.ujs.edu.cn, mao_qr@ujs.edu.cn, mdong@wayne.edu

Abstract

Semi-supervised Sound Event Detection (SSED) is to recognize the categories of events and mark their onset and offset times using a small amount of weakly-labeled and a large-scale of unlabeled data. To exploit unlabeled data effectively and reduce over-fitting, regularization techniques play a critical role in SSED. In this paper, we proposed a novel jointly regularized and locally down-sampled Conformer (Joint-Former) model for SSED. Joint-Former first locally down-samples the spectrogram and learns the token representations with high temporal resolution and low computational cost. Then, Joint-Former effectively exploits unlabelled data in SSED by integrating Mean-Teacher and Masked Spectrogram Modeling using joint regularization through a multitask learning framework. Extensive experiments on DCASE 2019, DCASE 2020, and DCASE 2021 task4 SSED datasets show that Joint-Former greatly outperformed existing methods.

Index Terms: Regularization, Masked spectrogram modeling, Multi-task learning, Sound event detection

1. Introduction

Semi-supervised Sound Event Detection (SSED) is a widely concerned task to recognize the categories of events and mark the onset and offset times for each event in a mixed audio signal using a small amount of weakly-labeled and a large-scale of unlabeled data. It generally contains two separate sub-tasks: audio tagging and audio localization [1].

In the literature, Convolutional Recurrent Neural Network (CRNN) [2, 3, 4, 5, 4] is often selected as the backbone network for SSED, which uses a convolutional neural network (CNN) as the front-end of a recurrent neural network (RNN) so that the necessary frequency domain features are learned together with temporal context features. Recently, benefiting from the superiority in modeling the time correlation of sound signals, Transformer-based models [6, 7, 8, 9] are introduced into SSED. The Convolution-augmented Transformer (Conformer) [10] was proposed and won the DCASE 2020 challenge task 4 [8]. To model long-time sequential data in SSED, Conformer uses several layers of CNN for global down-sampling. As a result, the sound representations will be smoothed with a lower temporal resolution, leading to less accurate boundary detection for sound events.

As manual-labeled datasets are scarce, SSED tasks require exploiting unlabeled data effectively to reduce over-fitting, typically through regularization techniques, e.g., consistency regularization. Recently, as one of the most popular consistency regularization techniques, Mean-Teacher (MT) [11] model was shown to be effective in learning robust representations for SSED [12, 13, 3, 14, 15, 16]. Given unlabeled data, the consistency regularization assumption in the MT framework requires the student model to predict consistently with the teacher model, an average of consecutive student models. Clearly, regularization plays a key role in solving SSED problems.

Another recent noteworthy work in audio representation is Masked Spectrogram Modeling (MSM), which learns general representation for audio with unlabeled data via self-supervised learning, and then fine-tunes the model for downstream tasks [17, 18, 19, 20, 21]. Models pre-trained by MSM can successfully capture the local correlations of the spectrogram when reconstructing the masked spectrogram using only the unmasked part. However, MSM pre-training usually comes with demanding training and optimization challenges, which involve large-scale training data and extremely long training time.

Aiming at learning a robust feature representation, in this paper, we propose Joint-Former, a novel Jointly regularized and locally down-sampled Conformer model for SSED. Specifically, Joint-Former first down-samples the spectrogram and learns the token representations with high temporal resolution and low computational cost through the proposed intra-patch local down-sampling strategy. Then, Joint-Former integrates MT and MSM to facilitate the downstream tasks using joint regularization through a multitask learning framework. The contributions of our work are summarized as follows,

- The joint (MT and MSM) regularization in Joint-Former allows us to effectively exploit unlabelled data in SSED by exploring both the local correlations and the task-aware global interactions for robust representation learning. Also, our multi-task learning framework provides more effective training than MSM-based pre-training, especially with limited (unlabelled) training samples.
- We propose a novel intra-patch local down-sampling strategy in Joint-Former, which is particularly helpful when handling the long-time spectrogram. High temporal resolution and low computational cost are critical for SSED tasks.
- Extensive experiments on benchmark SSED datasets show that Joint-Former greatly outperformed existing methods.

2. Methodology

In this section, we propose the jointly regularized and locally down-sampled Conformer model (Joint-Former) for SSED. We first describe the multi-task learning framework and the intra-patch local down-sampling strategy in Joint-Former, and then formulate the joint regularization using MSM and MT.

[*] *Corresponding author.*

10.21437/Interspeech.2023-344

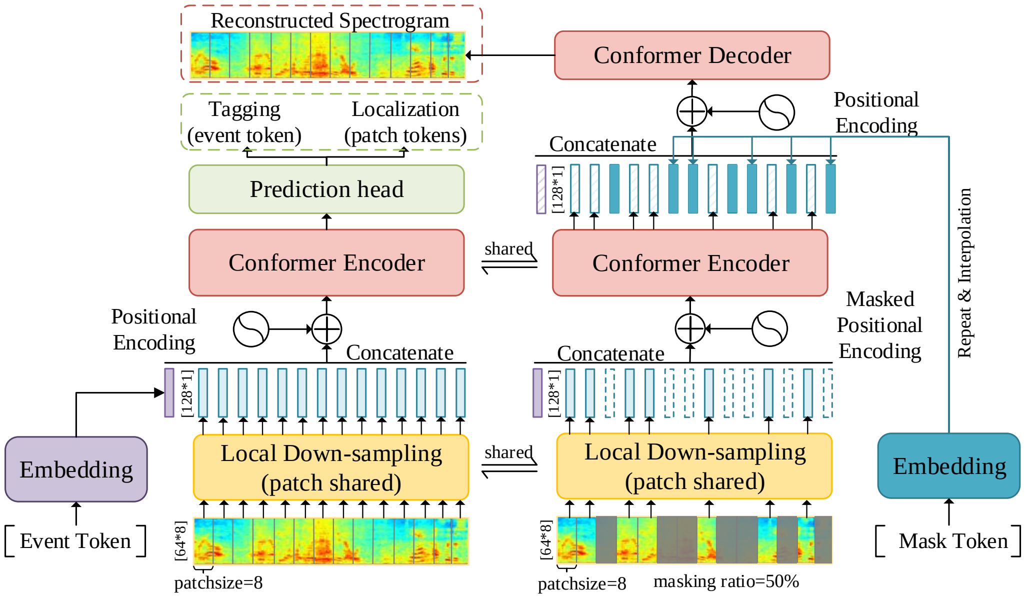


Figure 1: *The multi-task framework of the proposed Joint-Former, where red and green dashed rectangle represents the output of MSM branch and SED branch, respectively.*

2.1. Multi-task framework of Joint-Former

As shown in Fig. 1, the proposed Joint-Former contains two key branches: the major branch for SED and the auxiliary branch for MSM. The SED branch aims at learning event-aware representations from the raw spectrogram, while the MSM branch tries to capture the natural local correlations. We design a muti-task learning framework by sharing the CNN-based down-sampling layers and the Conformer encoder between the two branches so that event-aware features and natural local correlations are learned together. In the following, we introduce our intra-patch local down-sampling strategy, followed by detailed description of each downstream branch.

2.1.1. Intra-patch Local Down-sampling Strategy

To alleviate the temporal resolution reduction in the token representation, resulted from the global down-sampling in the vanilla Conformer, we propose to perform the intra-patch local down-sampling strategy.

Specifically, as shown in Fig. 1, given a D-dimensional spectrogram $\mathcal{X} \in \mathbb{R}^{T\times D}$ with T frames, we first patchify $\mathcal{X}$ along the time dimension with patch size t, resulting in $K = \frac{T}{t}$ numbers of patches $P = \{P_k|k = 1, 2, ..., K\}$. So, each patch contains t continuous frames of spectrogram features. Then, we use a shared CNN $\mathcal{F}$ to down-sample each patch and obtain the latent representations, i.e., *patch tokens* $Z = \{Z_k|k = 1, 2, ..., K\}$, where $Z_k = \mathcal{F}(P_k)$, and $Z_k \in \mathbb{R}^{1\times D'}$. As a result, each patch is down-sampled to 1-dimension along the time-axis and mapped into D'-dimensional token representations by D' convolutional kernels. In this case, the patch size t equals to the down-sampling scale. Besides, the structure of CNN $\mathcal{F}$ is the same as [8], which contains seven layers of convolutional blocks.

Different from a vanilla Conformer which down-samples long-time data globally by adapting a CNN at the whole sequence, Joint-Former uses the CNN inside each non-overlapping patch only, which keeps the patch token higher temporal resolution. Besides, the intra-patch local down-sampling can be conducted in parallel, which greatly reduces the computational cost.

2.1.2. Sound Event Detection Branch

After down-sampling, the patch tokens are concatenated with a learnable event token, and then the positional encoding will be embedded into the token sequence. Next, the token sequence is delivered to the Conformer encoder to learn the global interactions among tokens. Following [8], the encoder contains three layers of the Conformer block. Finally, we use a prediction head containing one layer of linear connection and a Sigmoid activation to get the SED predictions. More specifically, we use the event token to get the predictions of audio tagging and adopt the patch tokens to locate each sound event, i.e., audio localization.

2.1.3. Masked Spectrogram Modeling Branch

In the MSM branch, partial spectrogram patches are first masked randomly, and then the unmasked patches are down-sampled and embedded with the positional encoding. Note that the positional encoding of the unmasked patches should be consistent with their absolute position in the original patches. After encoding, we get the latent embeddings of the unmasked patch tokens. To reconstruct the original spectrogram, a shared and learnable mask token will be repeated and interpolated into the latent embeddings at the corresponding position.

Previously, MSM-based self-supervised learning only reconstructs the masked part of the input to reduce the computation cost in the pre-training, and then fine-tunes the model for downstream tasks. In Joint-Former, instead, we optimize the MSM and downstream tasks together through multi-task learning, which is considered as another way of regularization in addition to MT model. As only reconstructing the masked parts forces the model to pay too much attention to local correlations, we reconstruct the whole spectrogram in Joint-Former.

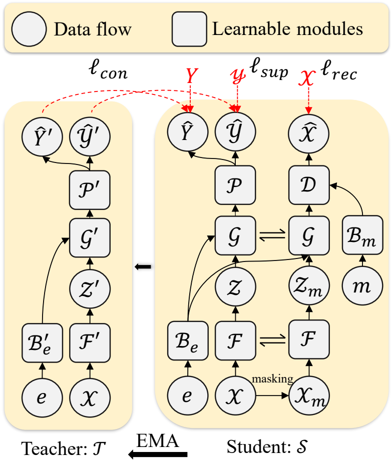


Figure 2: *The computation graph of Joint-Former, where the circles denote the data flow, and the rounded rectangles are learnable modules.*

2.2. Formulating the joint regularization

To effectively exploit unlabeled data in SSED, we jointly regularize the model training in Joint-Former through both MT and MSM techniques. Specifically, the consistency regularization assumption requires the student model to predict consistently with the teacher model, where the teacher model is the ensembled history of students by Exponential Moving Average (EMA). Meanwhile, the student model is required to reconstruct the spectrogram against *masking perturbation*. So, the MSM task can be considered as an auxiliary regularization task. Following the computation graph of our Joint-Former as shown in Fig. 2, we now formulate the joint regularization in details.

For the student model $\mathcal{S}$, given the input spectrogram $\mathcal{X}$, we get the predictions $\hat{Y}$ for audio tagging and $\hat{y}$ for localization (shown in Eq. (1)), where $\mathcal{P}$, $\mathcal{G}$, $\mathcal{B}_e$, e, and $\mathcal{F}$ presents prediction head, Conformer Encoder, event embeddings, initial event tokens, and down-sampling layers, respectively.

$$\hat{Y}, \hat{y} = \mathcal{P}(\mathcal{G}(\mathcal{B}_e(e), \mathcal{F}(\mathcal{X}))) \tag{1}$$

For consistency regularization, we ensemble the consecutive student models over the previous training steps by EMA with a smoothing decay α as the current teacher model $\mathcal{T}$,

$$\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t, \tag{2}$$

where θ_t and θ'_t denotes the parameters set of student $\mathcal{S}$ at t training step and teacher $\mathcal{T}$, respectively. Then, a consistency cost ℓ_{con} is applied between the predictions of $\mathcal{S}$ and $\mathcal{T}$ when given the unlabeled data,

$$\ell_{con}(\hat{Y}', \hat{y}'; \hat{Y}, \hat{y}) = BCE(\hat{Y}, \hat{Y}') + BCE(\hat{y}, \hat{y}'), \tag{3}$$

where $\hat{Y}'$ and $\hat{y}'$ denotes the prediction of the teacher model in audio tagging and localization, respectively.

For MSM auxiliary regularization, we compute a reconstruction cost $\ell_{rec}(\hat{\mathcal{X}}, \mathcal{X}) = MSE(\hat{\mathcal{X}}, \mathcal{X})$ between the inputs $\mathcal{X}$ and the reconstructed spectrogram $\hat{\mathcal{X}}$,

$$\hat{\mathcal{X}} = \mathcal{D}(\mathcal{G}(\mathcal{B}_e(e), \mathcal{F}(\mathcal{X}_m)), \mathcal{B}_m(m)), \tag{4}$$

where $\mathcal{D}$, $\mathcal{X}_m$, $\mathcal{B}_m$, and m presents decoder, masked spectrogram, mask embedding, and initial mask tokens, respectively.

Table 1: *Results of ablation study, where "Down." and "Reg." denotes down-sampling strategies and regularization technologies, respectively. Best results are* ***bolded****.*

Model ablations	Down.	Reg.	Event-F1
(1#) ConformerSED [8]	global	consistency	46.6%
(2#) 1# - CNN	none	consistency	31.4%
(3#) 2# + local down.	**local**	consistency	**48.4%**
(Joint-Former) 3# + MSM	**local**	**joint**	**51.3%**

Finally, we join the above regularization terms with the supervised SED loss ℓ_{sup} for robust representation learning in SSED,

$$\mathcal{L} = \ell_{sup} + w(t)(\ell_{con} + \lambda\ell_{rec}), \tag{5}$$

where $w(t)$ is the weight of joint regularization at the $t-th$ training step, ramped up during the training, and λ is a fixed weight of the MSM auxiliary regularization.

3. Experimental Evaluation

3.1. Experiment Setup

To evaluate the performance of our proposed model, we compare Joint-Former with ConformerSED [8], the winner in DCASE 2020 challenge task 4, on the DCASE 2019 [22], 2020 [23], and 2021 [24] task 4 challenge datasets. The training sets in these tasks all contain the same 1578 audio clips with a weak label and 14,412 unlabeled real-recorded audio clips, but a different number of synthetic strongly labeled samples (2,045, 2,584, and 10,000 for the three tasks, respectively). For the metrics, we use event-based macro F1 (Event-F1), used in DCASE 2019 and 2020 challenges, and Polyphonic Sound Detection Scores (PSDS-1 and PSDS-2) [25], used in DCASE 2021.

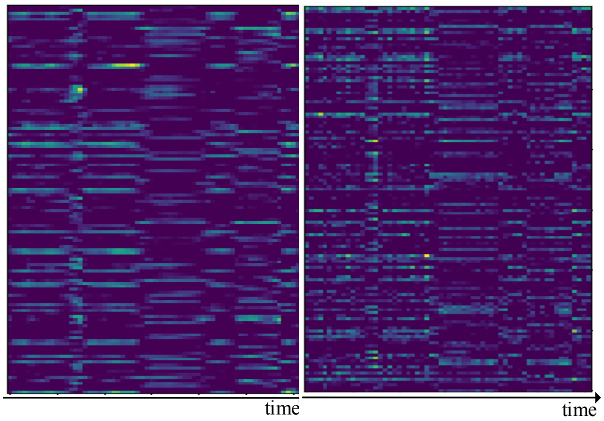
For a fair comparison, the SED branch in Joint-Former has the same architecture as ConformerSED [8]. Also, we extract the 64-dimensional log Mel spectrogram features with the 323 hop size and 1024 sampling width for each clip down-sampled to 16kHz, leading to a total of 496 frames. The mask ratio and auxiliary regularization weight in Eq. (5) is chosen as 50% and 0.1 by grid search, respectively. The down-sampling scale (patch size) is set as 8. ConformerSED won in DCASE 2020 challenge task 4. For a more general comparison, we reproduce ConformerSED [8] results on DCASE 2019 and 2021 datasets following the same experiment setup they adopted on the DCASE 2020 dataset. Ablation studies are conducted on the DCASE 2019 dataset to investigate the influences of each component in Joint-Former. The source code of our work is available[1].

3.2. Ablation studies

3.2.1. *Intra-patch local down-sampling*

As shown in Table. 1, we conduct three groups of experiments to evaluate the impact of down-sampling in Joint-Former, i.e., (1#) the global down-sampling in vanilla Conformer [8], (2#) without down-sampling (by removing the CNN-based down-sampling layers from (1#)) and (3#) our intra-patch local down-sampling. Clearly, removing the down-sampling layers (2#) leads to great performance degradation for SSED when compared to (1#) the vanilla Conformer. In contrast, there is a

[1] https://github.com/mastergofujs/Joint-Former



(a) Global down-sampling (b) Local down-sampling

Figure 3: *Visulizations of the token embeddings from (a) global down-sampling strategy in ConformerSED [8] and (b) intra-patch local down-sampling strategy in Joint-Former (ours).*

Table 2: *Performance comparison between one-/two-stage training for Joint-Former on DCASE 2019 dataset.*

Models	Event-F1	Training steps
ConformerSED	46.6%	30,000
Two-stage training	47.2%	30,000 * 2
One-stage training	**51.3%**	30,000

1.8% performance gain on Event-F1 by our (3#) intra-patch local down-sampling strategy.

Additionally, in Fig. 3 we visualize the token representations gained from the global and local down-sampling, respectively. Compared with smoothed tokens resulted from the global down-sampling in the vanilla Conformer (Fig. 3 (a)), the qualitative results in Fig. 3 (b) clearly show that the resolution of token representations obtained by our local down-sampling is improved greatly, especially along the time-axis, which is critical for SSED tasks.

3.2.2. Joint regularization

Next, we evaluate the influence of the proposed joint regularization. Clearly, when compared with single consistency regularization, the MSM-based auxiliary regularization achieves another 2.9% performance gain (The last row of Table. 1).

3.2.3. Two-stage vs. one-stage training

Typically, MSM-based representation learning is a two-stage process: pre-training and fine-tuning, where the pre-training stage usually requires a huge amount of training data. In Joint-Former, instead, we exploit unlabeled data with our novel joint regularization, a one-stage multi-task learning framework.

As a comparison, we also implemented Joint-Former with the two-stage training on the DCASE 2019 dataset. Specifically, in the first stage, we freeze the SED branch and pre-train the MSM branch in Joint-Former. Then, we freeze the MSM branch and fine-tune the Conformer model (the SED branch in Joint-Former) through consistency regularization in the second stage.

Table 3: *Performance comparison with SOTAs methods on DCASE 2019, 2020, and 2021 challenge datasets. "–" denotes the results are not reported, and the underline means the reproduced results. Best results are **bolded**.*

Models	2019	2020	2021	
	Event-F1	Event-F1	PSDS1	PSDS2
GL [2]	42.7%	–	–	–
Sparse-Trans [7]	–	47.6%	–	–
CNN-Trans [6]	–	–	29.2%	55.0%
ConformerSED [8]	46.6%	46.0%	29.7%	52.0%
Joint-Former	**51.3%**	**49.5%**	**33.9%**	**55.1%**

As shown in Table 2, compared to ConformerSED [8], the two-stage pre-trained Joint-Former only improves 0.6% (from 46.6% to 47.2%) on Event-F1, while the Joint-Former with one-stage training improves 4.7%. Recall that the DCASE 2019 dataset only contains about 10,000 unlabeled training samples. We believe this is not sufficient to support an MSM-based model pre-training. In contrast, the remarkable performance of Joint-Former with one-stage training clearly demonstrates the effectiveness of our multi-task learning framework in exploiting unlabeled data for SSED.

3.3. Performance comparison with SOTAs

To extensively evaluate the performance of Joint-Former, we compare it with the state-of-the-art (SOTA) models on DCASE 2019, 2020, and 2021 challenge datasets, including GL [2] (winner of DCASE 2019), Sparse-Trans [7], CNN-Trans [6], and ConformerSED [8] (winner of DCASE 2020). Here, we compare only with none-ensemble methods as Joint-Former is not an ensemble-based method. As shown in Table 3, Joint-Former greatly improves the performance on all the metrics on these benchmark datasets. Specifically, the performance gain of Joint-Former is 4.7%, 3.5%, 4.2%, and 3.1% on DCASE 2019, 2020, and 2021 datasets, respectively. The superior performance of Joint-Former over the SOTAs clearly demonstrates the effectiveness of joint regularization and local down-sampling.

4. Conclusion

In the field of SSED, it is critical to learn robust representation from semi-labeled data, i.e., a training set that contains a small amount of weakly labeled and a large-scale of unlabeled data. In this paper, we proposed a novel jointly regularized and locally down-sampled Conformer (Joint-Former) model. Joint-Former can model long-time sequential data efficiently by an intra-patch local down-sampling strategy, and exploit unlabeled data effectively through multitask joint (MT and MSM) regularization. Extensive experiments on benchmark datasets clearly demonstrate the superior performance of Joint-Former for SSED.

5. Acknowledgements

This work is partially supported by the Key Project of National Nature Science Foundation of China (Grant No. U1836220), the Jiangsu key research and development plan (industry foresight and key core technology, Grant No. BE2020036), and the general project of National Nature Science Foundation of China (Grant No. 62176106) (to QM), the Post-graduate Research & Practice Innovation Program of Jiangsu Province (Grant No. KYCX22_3668) (to LG), and the MTRAC Grant for Advanced Computing Technologies (to MD).

6. References

[1] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, pp. 67–83, 2021.

[2] L. Lin, X. Wang, H. Liu, and Y. Qian, "Guided learning for weaklylabeled semisupervised sound event detection," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 626–630.

[3] 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 *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 376–380.

[4] L. Gao, L. Zhou, Q. Mao, and M. Dong, "Adaptive hierarchical pooling for weakly-supervised sound event detection," in *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 2022.

[5] D. de Benito-Gorrón, D. Ramos, and D. T. Toledano, "A multi-resolution crnn-based approach for semi-supervised sound event detection in dcase 2020 challenge," *IEEE Access*, vol. 9, pp. 89 029–89 042, 2021.

[6] K. Wakayama and S. Saito, "Cnn-transformer with self-attention network for sound event detection," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 806–810.

[7] Y. Guan, J. Xue, G. Zheng, and J. Han, "Sparse self-attention for semi-supervised sound event detection," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 821–825.

[8] 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," Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE), Tech. Rep., 2020.

[9] S.-J. Kim and Y. Chung, "Multi-scale features for transformer model to improve the performance of sound event detection," *Applied Sciences*, 2022.

[10] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proceedings of INTERSPEECH*, vol. abs/2005.08100, 2020.

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

[12] X. Zheng, Y. Song, I. Mcloughlin, L. Liu, and L. Dai, "An improved mean teacher based method for large scale weakly labeled semi-supervised sound event detection," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 356–360.

[13] J. Yan, Y. Song, L. Dai, and I. Mcloughlin, "Task-aware mean teacher method for large scale weakly labeled semi-supervised sound event detection," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 326–330.

[14] X. Zheng, Y. Song, J. Yan, L. Dai, I. Mcloughlin, and L. Liu, "An effective perturbation based semi-supervised learning method for sound event detection," in *Proceedings of INTERSPEECH*, 2020.

[15] Y. Hu, X. Zhu, Y. Li, H.-M. Huang, and L. He, "A multi-grained based attention network for semi-supervised sound event detection," in *Proceedings of INTERSPEECH*, vol. abs/2206.10175, 2022.

[16] N. Shao, E. Loweimi, and X. Li, "Rct: Random consistency training for semi-supervised sound event detection," in *Proceedings of INTERSPEECH*, 2021.

[17] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

[18] A. Baade, P. Peng, and D. F. Harwath, "Mae-ast: Masked autoencoding audio spectrogram transformer," in *Proceedings of INTERSPEECH*, vol. abs/2203.16691, 2022.

[19] H. Dinkel, P. Zhang, M. Wu, and K. Yu, "Depa: Self-supervised audio embedding for depression detection," in *Proceedings of ACM International Conference on Multimedia (ACM MM)*, 2021.

[20] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. R. Glass, "Ssast: Self-supervised audio spectrogram transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[21] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

[22] N. Turpault, R. Serizel, A. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Tech. Rep., 2019.

[23] 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," Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE), Tech. Rep., 2020.

[24] S. Wisdom, H. Erdogan, D. P. W. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. R. Hershey, "What's all the fuss about free universal sound separation data?" in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 186–190.

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