Background-aware Modeling for Weakly Supervised Sound Event Detection

Yifei Xin, Dongchao Yang, Yuexian Zou[*]

School of ECE, Peking University, Shenzhen, China
xinyifei@stu.pku.edu.cn

Abstract

Nowadays, a common framework for weakly supervised sound event detection (WSSED) is multiple instance learning (MIL). However, MIL directly optimizes the clip-level classification results, so it tends to localize the most distinct part rather than the entire sound event, making the indiscriminating parts of sound events mistakenly identified as background sounds. In this paper, we focus on adding background awareness for WSSED by proposing a learning structure called BA-WSSED. Our BA-WSSED first introduces a pseudo separator with softmax activation and two aggregators to purify and aggregate the event feature and the background feature, respectively. Then, with the help of the proposed background-aware staggered (BAS) loss, both the event classifier and the background classifier are learned to generate staggered classification scores for discerning and suppressing background sounds. Experiments show that our BA-WSSED significantly improves the performance of the general MIL-based WSSED method on multiple datasets and can be employed on various baseline models.

Index Terms: weakly supervised sound event detection, background awareness, multiple instance learning

1. Introduction

Sound event detection (SED) [1–3] consists of two subtasks, one is to tag the absence or presence of sound events in an audio clip, and the other is to locate their corresponding onset and offset times. SED has many potential applications (e.g., smart cities [4], surveillance [5]) and can also help enhance the performance of many other related tasks (e.g., audio caption [6], audio-text retrieval [7, 8], speech enhancement and separation [9–11]). Recently, weakly supervised sound event detection (WSSED) [12–14] has gained increasing attention as weak labels are much easier to gather than strong labels. A common framework for WSSED is multiple instance learning (MIL) [15, 16]. In MIL, we do not know the ground-truth label of each training instance; instead, the instances are grouped into bags, and we only know the label of bags. For WSSED, every training audio represents a bag, and its frames are treated as instances.

The general MIL-based WSSED method provides two ways to determine the clip-level predictions of an audio clip based on the frame-level information: the instance-level strategy and the embedding-level strategy. For the instance-level strategy, the aggregator integrates the frame-level probabilities generated by the classifier to produce the clip-level prediction result. For the embedding-level strategy, the aggregator pools the frame-level features output by the feature extractor into a clip-level feature, which is then fed into the classifier to obtain the clip-level probability. In [17], Lin et al. carried out a series of experiments and found that the embedding-level approach tends to perform better than the instance-level approach but receives less attention. Taking account of this, our work focuses more on the embedding-level approach and adopts it as our baseline.

However, the general MIL-based WSSED method only optimizes the global loss calculated from the aggregated clip-level predictions and weak clip-level labels, lacking direct constraints on the frame-level information, so it tends to localize the most distinct part but not the whole sound event, thus failing to catch the indiscriminating parts of sound events and misidentifying them as background sounds.

In this work, we attribute the above problems of the general MIL-based WSSED method to its unawareness of the background sound. Specifically, the goal in the WSSED training phase is to determine whether each category exists in an audio clip, rather than detecting sound events for each frame like the strongly labeled SED task, so the model tends to capture the most prominent features of each sound event category instead of learning to localize the whole sound event, thus failing to catch the indiscriminating parts of sound events and misidentifying them as background sounds. Besides, in many widely used WSSED datasets (e.g., DCASE2017 task4 and UrbanSED datasets), an input audio clip contains at least one sound event, so that the aggregated pure-background feature remains invisible for the clip-level supervised SED task. As a result, the general MIL-based WSSED method focuses more on discerning different sound event classes, but has limitations in simultaneously identifying whether one frame belongs to sound event parts or background sounds.

Moreover, the unawareness of background sounds also prevents the MIL-based WSSED method from suppressing the over-activation of the frames representing background sounds in an audio clip, especially the kind of background sounds that often co-occur with specific sound events, which are easily mistaken for being part of the co-occurring sound events. Therefore, if the background cues can be perceived, the over-activation of background sounds can be suppressed, and thus the performance of sound event localization can be improved.

In this paper, we focus on adding background awareness for WSSED by proposing an end-to-end learning structure called BA-WSSED. Our BA-WSSED attempts to generate the "unseen pure-background samples" by aggregating the frame-level features of sound events and background sounds, respectively. Then, with the assistance of our background-aware staggered (BAS) loss, an additional background classifier can be simultaneously learned with the event classifier to discern and suppress

This paper was partially supported by NSFC (No: 62176008) and Shenzhen Science & Technology Research Program (No:GXWD20201231165807007-20200814115301001).

[*] Yuexian Zou is the corresponding author.

10.21437/Interspeech.2023-330

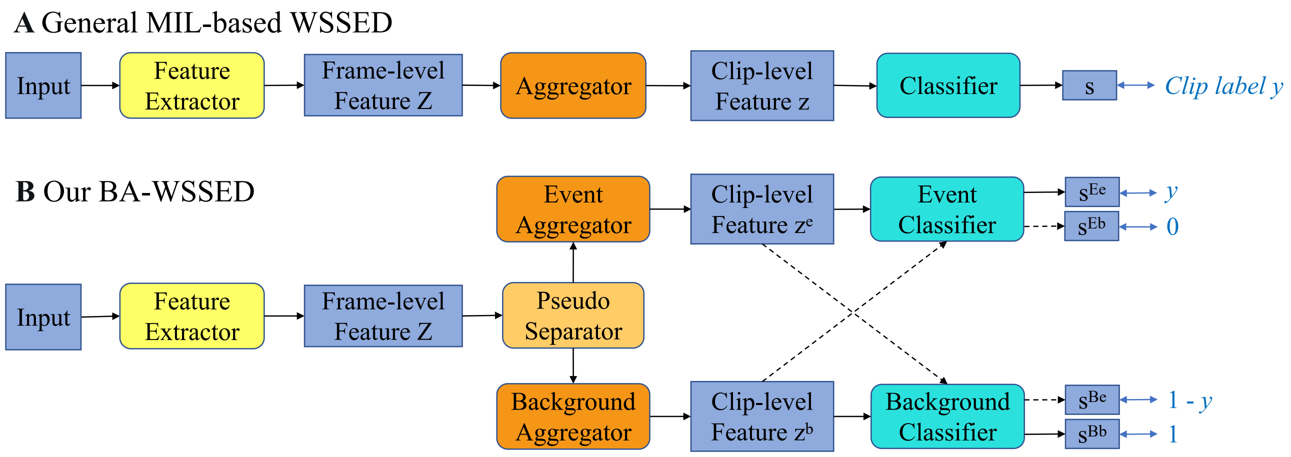

Figure 1: *The comparison of the general MIL-based WSSED pipeline and our BA-WSSED.*

the background sounds by considering background prediction as a multi-label classification task, thus preventing sound events from being misidentified as background sounds and helping the model to capture the indistinct parts of sound events.

In a nutshell, our contributions are threefold:

- To the best of our knowledge, our work is the first to supplement background awareness for WSSED and simultaneously learn both event and background classifiers with only clip-level labels.
- A novel structure called BA-WSSED is proposed to discern and suppress the background sounds and a BAS loss is elaborated to efficiently train our BA-WSSED in an end-to-end manner.
- Experiments show that our BA-WSSED significantly improves the performance of the general MIL-based WSSED method and can generalize to various baseline models.

2. Proposed method

2.1. Cause of the Unawareness of Background Sounds

The general MIL method only optimizes the global classification loss with the clip-level label y, so it tends to localize the most discriminating part but not the whole sound event, which leads to over-activation of background sounds and thus severely limits the localization performance. We attribute the problem to the inconsistent goals between MIL and SED, where MIL places too much emphasis on the clip-level sound event classification without taking into account the characteristic of SED, e.g., the location of background sounds, which is also critical and needs to be discerned for the WSSED task.

Besides, in many common WSSED datasets, since an audio clip contains at least one sound event, pure-background clip-level samples do not exist for the training of MIL in WSSED, so the model focuses more on learning to discern different sound event classes with only clip-level labels in the training stage, and thus cannot identify well whether a frame belongs to sound events or background sounds when performing frame-level predictions during inference. The inconsistency of training and inference objectives is a non-negligible gap for the classifier, which diminishes the capacity of the classifier in discerning and suppressing activation of background sounds to a great extent and seriously affects the accuracy of sound event localization.

Moreover, for the general MIL-based WSSED method, as shown in Fig. 1 A, the feature aggregator pools the frame-level features of sound events and background sounds mixedly. As a result, the clip-level feature z is unavoidably affected by the distribution of the background sound, especially those background sounds that often co-occur with sound events, causing undesirable background activation.

2.2. Network Design

To solve the above problems, our BA-WSSED is proposed as shown in Fig. 1 B. We first introduce event and background aggregators to replace the original single aggregator (i.e., the pooling function) [16, 18]. Specifically, after obtaining the frame-level features $Z \in \mathbb{R}^{C\times N}$ by the feature extractor (e.g., CNN, RNN or Transformer-based backbone), where C is the number of channels and N is the number of frames, we employ a 1 × 1 convolution with column-wise softmax activation as the pseudo separator, to project and generate the event prior A^e and the background prior A^b for each temporal position of the frame-level features Z:

$$\begin{cases} A^e_{:,i} = \frac{exp(W_1 * Z_{:,i})}{\sum_j^N exp(W_1 * Z_{:,j})} \\ A^b_{:,i} = \frac{exp(W_2 * Z_{:,i})}{\sum_j^N exp(W_2 * Z_{:,j})} \end{cases}, \tag{1}$$

where $W_1 \in \mathbb{R}^{M\times C}$ and $W_2 \in \mathbb{R}^{M\times C}$ are weight matrices for 1 × 1 convolution. Thus, each column-vector of the priors A^e and $A^b \in \mathbb{R}^{M\times N}$ can be viewed as an attention map to capture the temporal relations of frames by activating potential locations that belongs to sound events or background sounds, where M is the number of attention maps. Then, based on the two priors A^e and A^b, the event aggregator and the background aggregator are adopted to generate clip-level features z^e and $z^b \in \mathbb{R}^{C\times 1}$. In detail, we first utilize the corresponding column-vectors of A^e (or A^b) as the attention map to aggregate the frame-level features Z into M different aggregation features. Then, the final clip-level features are obtained by calculating the mean strength of these aggregation features:

$$\begin{cases} z^e = \frac{1}{M}\sum_m^M \sum_i^N A^e_{m,i} Z_{:,i} \\ z^b = \frac{1}{M}\sum_m^M \sum_i^N A^b_{m,i} Z_{:,i} \end{cases}. \tag{2}$$

Profited by the two priors, the clip-level event feature z^e is less affected by the background feature compared to the clip-level feature z of the general MIL-based WSSED method. Furthermore, the additional clip-level background feature z^b simulates features aggregated from "pure background audio clips", which

Table 1: *Performance comparison of our BA-WSSED and previous methods on the DCASE2017 task4 evaluation set.*

Method	AT-F1	Seg-F1	Event-F1
Winner SED [19]	0.526	0.555	-
CNN-I [20]	0.565	0.466	0.103
CDur-I [21]	0.553	0.508	0.152
CNN Transformer-I [20]	0.629	0.556	0.195
CNN biGRU-I [20]	0.625	0.564	0.193
HTSAT-I [22]	0.636	0.587	0.178
CNN-E	0.608	0.522	0.114
CDur-E	0.556	0.516	0.158
CNN Transformer-E	0.636	0.564	0.202
CNN biGRU-E	0.633	0.566	0.196
HTSAT-E	0.640	0.590	0.180
CNN-BA	**0.627**	**0.546**	**0.123**
CDur-BA	**0.568**	**0.532**	**0.166**
CNN Transformer-BA	**0.641**	**0.570**	**0.206**
CNN biGRU-BA	**0.658**	**0.586**	**0.205**
HTSAT-BA	**0.662**	**0.604**	**0.186**

assists the subsequent classifiers to discern and suppress the activation of background sounds.

Our BA-WSSED adopts two classifiers, which adds an additional background classifier $s^B(\cdot)$ upon the event classifier $s^E(\cdot)$. By feeding the two clip-level features z^e and z^b into our event and background classifiers, we can obtain four classification scores s^{Ee}, s^{Eb}, s^{Be}, and $s^{Bb} \in \mathbb{R}^{K\times 1}$, which are all used to supervise the network training with the clip-level labels y, where K is the number of classes. Specifically, the event classifier generates two event classification scores s^{Ee} and s^{Eb} respectively for the clip-level event feature z^e and the background feature z^b. Likewise, the background classifier produces two background classification scores s^{Be} and s^{Bb}. Based on the four classification scores, we will introduce the details of our background-aware staggered (BAS) loss in the next subsection.

2.3. Background-aware Staggered Loss

In general, the background-aware staggered (BAS) loss serves as a multi-task loss that trains the event classification and background classification tasks with our z^e and z^b samples. The labels for both tasks are gathered based on the following properties:

The feature aggregated by sound events, i.e., z^e, is the positive sample for event classification, so the clip-level label $\boldsymbol{y}$ is the ground-truth of the event classification task for z^e, i.e., $\boldsymbol{y^{Ee}} = \boldsymbol{y}$. On the other hand, z^e is the negative sample for background classification, which is also the positive sample of other inactive sound event classes for the background classification task. Consequently, $\boldsymbol{1} - \boldsymbol{y}$ is the ground-truth of the background classification task for z^e, i.e., $\boldsymbol{y^{Be}} = \boldsymbol{1} - \boldsymbol{y}$.

The feature aggregated by background sounds, i.e., z^b, is the negative sample of all sound events for event classification. Therefore, 0 is the ground-truth of the event classification task for z^b, i.e., $\boldsymbol{y^{Eb}} = \boldsymbol{0}$. Meanwhile, z^b is the positive sample of all sound events for background classification. As a result, 1 is the ground-truth of the background classification task for z^b, i.e., $\boldsymbol{y^{Bb}} = \boldsymbol{1}$.

After obtaining the labels of the samples z^e, z^b on the event and background classification tasks, the BAS loss is designed to train our BA-WSSED. In detail, our BAS loss contains four terms:

Table 2: *Performance comparison of BA-WSSED and previous methods on the weakly labeled UrbanSED test set.*

Method	AT-F1	Seg-F1	Event-F1
Base-CNN [23]	-	0.560	-
CDur-I [21]	0.771	0.647	0.217
HTSAT-I [22]	0.771	0.644	0.210
CDur-E	0.775	0.650	0.218
HTSAT-E	0.778	0.648	0.216
CDur-BA	**0.784**	**0.658**	**0.221**
HTSAT-BA	**0.785**	**0.661**	**0.224**

$$\begin{aligned} L_{bas} = \lambda_1 * L_E(s^{Ee}, y) + \lambda_2 * L_B(s^{Be}, 1-y) + \\ \lambda_3 * L_E(s^{Eb}, 0) + \lambda_4 * L_B(s^{Bb}, 1), \end{aligned} \quad (3)$$

where L_E represents the event classification loss and L_B denotes the background classification loss, both of which are implemented by the cross-entropy loss.

In our BAS loss, the first term is used to supervise the accuracy of event classification as in the general MIL-based WSSED method, which ensures that z^e has a high probability of being recognized as active sound events, while the second term is leveraged to prevent z^e from being misidentified as background sounds. Furthermore, the third term forces z^b to be indiscriminate for all sound event classes, helping the event classifier to perceive pure-background samples and suppress the activation of background sounds. Lastly, the fourth term makes z^b have a high probability of being recognized as background sounds and cooperates with other BAS loss terms to jointly guide the pseudo separator and two aggregators to purify and aggregate frame-level features of sound events and background sounds to form z^e and z^b, respectively.

During the inference stage, since the event feature has been purified by the pseudo separator and the event classifier has been trained to recognize and suppress background sounds with the help of our BAS loss, we only need to feed the frame-level event feature into the event classifier to generate the final frame-level prediction results.

3. Experiments and Results

3.1. Datasets

We evaluate our method on the two publicly available sound event detection datasets: DCASE2017 task4 [24] and UrbanSED [23] datasets. The DCASE2017 task4 dataset "large-scale weakly supervised sound event detection for smart cars" is made up of a training subset with 51,172 audio clips, a validation subset with 488 audio clips, and an evaluation set with 1103 audio clips, including 17 sound events. The UrbanSED dataset has 10 event labels within an urban setting, which contains a total of 10,000 soundscapes generated by the Scaper soundscape synthesis library, divided into 6,000 training, 2,000 validation, and 2,000 evaluation clips.

3.2. Baseline Models and Training Details

To evaluate the effectiveness and generalization of our BA-WSSED, we apply our method to multiple baseline systems, including CNN [20], CDur [21], CNN-biGRU [20], CNN-Transformer [20] and HTSAT [22]. The CNN system is modeled by a 9-layer CNN, which consists of 4 convolutional blocks. The convolutional block includes 64, 128, 256 and 512

Table 3: *Ablation Study of our BAS loss.*

Method	AT-F1	Seg-F1	Event-F1
CNN biGRU-I [20]	0.625	0.564	0.193
CNN biGRU-E	0.633	0.566	0.196
Ours (loss1)	**0.628**	**0.562**	**0.194**
Ours (+loss2)	**0.639**	**0.568**	**0.199**
Ours (+loss2/3)	**0.649**	**0.577**	**0.201**
Ours (+loss2/3/4)	**0.658**	**0.586**	**0.205**

Table 4: *Ablation Study of* $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ *settings, and the number of attention maps* M.

Method	AT-F1	Seg-F1	Event-F1
HTSAT-I [22]	0.636	0.587	0.178
HTSAT-E	0.640	0.590	0.180
Ours (λ_1=0.5)	0.656	0.595	0.182
Ours (λ_1=1.0)	**0.662**	**0.604**	**0.186**
Ours (λ_2=0.2)	0.661	0.602	0.184
Ours (λ_2=0.3)	**0.662**	**0.604**	**0.186**
Ours (λ_2=0.4)	0.659	0.601	0.182
Ours (λ_3=0.2)	0.659	0.602	0.186
Ours (λ_3=0.3)	**0.662**	**0.604**	**0.186**
Ours (λ_3=0.4)	0.658	0.599	0.185
Ours (λ_4=0.1)	0.661	0.602	0.185
Ours (λ_4=0.2)	**0.662**	**0.604**	**0.186**
Ours (λ_4=0.3)	0.659	0.601	0.184
Ours (M=40)	0.658	0.601	0.182
Ours (M=60)	**0.662**	**0.604**	**0.186**
Ours (M=80)	0.663	0.603	0.186
Ours (M=100)	0.662	0.604	0.187

feature maps, respectively. For the CDur system, it consists of a 5-layer CNN followed by a bidirectional Gated Recurrent Unit (GRU) with 128 hidden units. The CNN-biGRU system is modeled by a 9-layer CNN and a bidirectional GRU with 256 hidden units and the CNN-Transformer consists of a 9-layer CNN with one transformer block. HTSAT adopts the Swin Transformer [25] backbone with ImageNet-pretraining, where we use 3 network groups with 2, 2, 6 swin-transformer blocks for the DCASE2017 task4 dataset while we only use two stages with 2, 2 swin-transformer blocks for the UrbanSED dataset.

In this work, we follow the training pipeline of the corresponding baseline models and keep the aggregators [16] (including event and background aggregators) consistent with the corresponding baseline models for a fair comparison. Specifically, the CDur system adopts the linear softmax pooling function. The CNN system uses the max pooling function. The CNN Transformer and HTSAT systems use the average pooling function. The CNN biGRU system uses the attention pooling function. The hyper-parameters are set as M = 60, λ_1 = 1, λ_2 = λ_3 = 0.3, and λ_4 = 0.2. We use audio tagging F1 score (AT-F1), Segment-F1 score (Seg-F1) and Event-F1 score [26] to evaluate our method.

3.3. Results on DCASE2017 Task4

Experimental results on the DCASE2017 Task4 dataset are shown in Table 1, where *-I indicates the instance-level strategy, *-E denotes the embedding-level strategy, and *-BA represents our background-aware pipeline. It can be seen that our BA-WSSED achieves significant performance boosts on various baseline models with different pooling functions (aggregators) for both instance-level and embedding-level strategies, especially on the AT-F1 and Seg-F1 metrics. The excellent performance gains mainly benefit from the trait that our BA-WSSED can perceive the unseen pure-background samples and suppress the activation of background sounds, which makes the model more discriminative between sound events and background sounds, thus classifying and localizing sound events more accurately.

3.4. Results on UrbanSED

We also compare our BA-WSSED with previous approaches on the weakly labeled UrbanSED corpus. As shown in Table 2, our BA-WSSED also achieves consistent improvements compared with the corresponding baseline models on the UrbanSED dataset, which further demonstrates the effectiveness, robustness, and generalization of our method.

3.5. Ablation Study

In this part, we discuss the influence of our BAS loss and the selection of hyper-parameter settings. The experiments are carried out on the DCASE2017 task4 dataset.

Results of each term of our BAS loss. To show the effectiveness of each term in our BAS loss, we present ablation results using CNN-biGRU system in Table 3. Ours (loss1) denotes that we use our BA-WSSED pipeline, but only includes the term of loss1, i.e., $\lambda_1 * L_E$ in our BAS loss. Ours (+loss2) represents that we add the loss2 term, i.e., $\lambda_2 * L_B$ based on Ours (loss1). Ours (+loss2/3) means to add loss2 and loss3 terms, and Ours (+loss2/3/4) is in the same way. It can be seen that only using the loss1 term harms the performance compared with the baseline system. This is because in such a condition, the event feature is only coarsely formed without any restrictions, which may undesirably contain excessive background sounds or missing sound event parts. By combining loss2, loss3, and loss4 step by step, our BA-WSSED achieves consistent performance improvements, thus demonstrating the effectiveness of our background awareness modeling for WSSED.

Results of the hyper-parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, and the number of attention maps M. Here, we show the influences of hyper-parameter selection using the HTSAT system in Table 4. Ours (*) means that we only change the setting of *, and the rest of the settings are the same as the experimental best settings, which we show in bold. We can see that different λ settings would affect the results, but not too much. Besides, when changing the number of attention maps M from 60 to 40, the performance has a comparatively large drop. When M becomes larger, there is little performance gain. To save costs, we finally choose M = 60 for our BA-WSSED.

4. Conclusions

In this paper, we focus on adding background awareness for WSSED by proposing a BA-WSSED pipeline to suppress the over-activation of background sounds, which can not only help capture indistinct sound events but also distinguish sound events from their frequently co-occurring background sounds, thus effectively reducing the sound event classification errors and localization bias. Experiments show that our BA-WSSED yields significant performance gains compared to the general MIL-based WSSED method on multiple datasets and can generalize to various baselines.

5. References

[1] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.

[2] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in the dcase 2017 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 992–1006, 2019.

[3] Y. Xin, D. Yang, and Y. Zou, "Audio pyramid transformer with domain adaption for weakly supervised sound event detection and audio classification," *Proc. Interspeech 2022*, pp. 1546–1550, 2022.

[4] J. P. Bello, C. Mydlarz, and J. Salamon, "Sound analysis in smart cities," in *Computational Analysis of Sound Scenes and Events*. Springer, 2018, pp. 373–397.

[5] R. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio analysis for surveillance applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.* IEEE, 2005, pp. 158–161.

[6] A. Ö. Eren and M. Sert, "Audio captioning using sound event detection," *arXiv preprint arXiv:2110.01210*, 2021.

[7] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[8] Y. Xin, D. Yang, and Y. Zou, "Improving text-audio retrieval by text-aware attention pooling and prior matrix revised loss," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2023, pp. 1–5.

[9] Q. Kong, H. Liu, X. Du, L. Chen, R. Xia, and Y. Wang, "Speech enhancement with weakly labelled data from audioset," *arXiv preprint arXiv:2102.09971*, 2021.

[10] Q. Kong, Y. Wang, X. Song, Y. Cao, W. Wang, and M. D. Plumbley, "Source separation with weakly labelled data: An approach to computational auditory scene analysis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2020, pp. 101–105.

[11] Y. Xin, X. Peng, and Y. Lu, "Improving speech enhancement via event-based query," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2023, pp. 1–5.

[12] 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.

[13] M. R. Izadi, R. Stevenson, and L. Kloepper, "Affinity mixup for weakly supervised sound event detection," in *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP).* IEEE, 2021, pp. 1–6.

[14] Y. Xin, D. Yang, F. Cui, Y. Wang, and Y. Zou, "Improving weakly supervised sound event detection with causal intervention," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2023, pp. 1–5.

[15] S.-Y. Tseng, J. Li, Y. Wang, F. Metze, J. Szurley, and S. Das, "Multiple instance deep learning for weakly supervised small-footprint audio event detection," in *Proc. Interspeech 2018*, 2018, pp. 3279–3283.

[16] 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.

[17] 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.

[18] L. Zhu, Q. She, Q. Chen, X. Meng, M. Geng, L. Jin, Z. Jiang, B. Qiu, Y. You, Y. Zhang *et al.*, "Background-aware classification activation map for weakly supervised object localization," *arXiv preprint arXiv:2112.14379*, 2021.

[19] D. Lee, S. Lee, Y. Han, and K. Lee, "Ensemble of convolutional neural networks for weakly-supervised sound event detection using multiple scale input," *Detection and classification of acoustic scenes and events (DCASE)*, 2017.

[20] 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, pp. 2450–2460, 2020.

[21] 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.

[22] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2022, pp. 646–650.

[23] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).* IEEE, 2017, pp. 344–348.

[24] 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," in *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[26] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.