SOURCE SEPARATION FOR SOUND EVENT DETECTION IN DOMESTIC ENVIRONMENTS USING JOINTLY TRAINED MODELS

Diego de Benito-Gorrón[1], *Katerina Zmolikova*[2], *Doroteo T. Toledano*[1]

[1]AUDIAS Research Group, Escuela Politécnica Superior, Universidad Autónoma de Madrid
[2]Brno University of Technology, Faculty of IT, IT4I Centre of Excellence

2022 International Workshop on Acoustic Signal Enhancement (IWAENC) | 978-1-6654-6867-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IWAENC53105.2022.9914755

ABSTRACT

Sound Event Detection and Source Separation are closely related tasks: whereas the first aims to find the time boundaries of acoustic events inside a recording, the goal of the latter is to isolate each of the acoustic sources into different signals. This paper presents a Sound Event Detection system formed by two independently pre-trained blocks for Source Separation and Sound Event Detection. We propose a joint-training scheme, where both blocks are trained at the same time, and a two-stage training, where each block trains while the other one is frozen. In addition, we compare the use of supervised and unsupervised pre-training for the Separation block, and two model selection strategies for Sound Event Detection. Our experiments show that the proposed methods are able to outperform the baseline systems of the DCASE 2021 Challenge Task 4.

Index Terms— Sound Event Detection, Source Separation, DCASE, DESED

1. INTRODUCTION

An important amount of information about our surrounding environment is provided by sounds. The human ability to recognize them generally gives us an immediate idea of where we are, or what is happening near us. In computational intelligence, several research fields try to automatically retrieve this kind of information from audio recordings. Particularly, the goal of Sound Event Detection (SED) is to perform an automatic classification of the sound events and determine their time boundaries [1].

The recent research in automatic processing of acoustic environments and sound events has been noticeably supported by the yearly editions of the DCASE (Detection and Classification of Acoustic Scenes and Events) international challenges [2]. In particular, DCASE Challenge Task 4 tackles the problem of Sound Event Detection in domestic environments. The task considers a heterogeneous training dataset with different kinds of audio and labels [3], such as synthetic soundscapes with strong labels (i.e. annotations for the sound event categories and their time boundaries), and audio from web videos provided with weak (clip-level) annotations or without any annotation at all. In order to leverage unlabeled and weakly-labeled data, mean teacher semi-supervised training [4] is often employed [5, 6].

A different research field in audio computational intelligence is Source Separation (SSep), which aims to decompose a sound recording into the latent acoustic sources that form it, regardless of their type (speech, music, background noise, etc.) [7]. Considering a training dataset of audio mixtures for which the original sources are available, deep neural networks can be trained for Source Separation in a classic supervised scheme, using Permutation Invariant Training (PIT) [8]. Moreover, an unsupervised training method for Source Separation called Mixture Invariant Training (MixIT) has recently been introduced [9], allowing to train Source Separation systems when the original sources of the training mixtures are not available.

Recent research has suggested the idea that Source Separation and Sound Event Detection tasks can benefit from each other, for instance, using the predictions of a SED system to guide the separation of events into different sources [10, 11, 12], learning SSep as an intermediate representation for SED [13], or applying SSep as a pre-processing step for SED, either fine-tuning the SED system over automatically separated data [14] or training a SSep network as a front-end stage to a pre-trained SED system [15].

In this paper, we propose a Sound Event Detection system composed of two pre-trained blocks: a Source Separation network and a Sound Event Detection network. In contrast with previous work, we do not limit the training to just one of the two blocks. Instead, aiming for both tasks to learn from each other, we introduce a joint training setting, where the whole system is trained in an end-to-end fashion, and a two-stage training, in which the SED block is fine-tuned first while freezing the SSep block (Stage 1), and then the SSep block is fine-tuned while freezing the SED block (Stage 2). Apart from these training settings and their analysis, our experimental results provide two additional contributions. First, for the SSep block, we compare supervised pre-training on mismatched data with unsupervised pre-training on matched data, in the context of the SED task. Second, we compare two model selection strategies for the mean teacher training, i.e. student and teacher model selection.

The paper is organized as follows: In Section 2, the Sound Event Detection and Source Separation tasks are introduced, discussing the most relevant aspects for this work. Section 3 describes the proposed methods for joint Source Separation and Sound Event Detection. The experimental setup is presented in Section 4, describing the model settings and the datasets that are employed, along with the results of the experiments and their discussion. Finally, Section 5 highlights the conclusions of the paper and overviews some future work.

2. SOUND EVENT DETECTION AND SOURCE SEPARATION

2.1. Sound Event Detection in DCASE Challenge 2021 Task 4

The goal of SED is to determine, from a given audio signal $\mathbf{x}$, the onset and offset times (t_{on}, t_{off}) of the occurrences of a closed set of K acoustic event categories. These times are usually obtained by

[1] Funded by Project RTI2018-098091-B-I00 (Spanish Ministry of Science and Innovation and ERDF). [2]Work supported by Czech Ministry of Education, Youth and Sports from project no. LTAIN19087 "Multi-linguality in speech technologies".

978-1-6654-6867-1/22/$31.00 ©2022 IEEE

setting a threshold $\tau \in (0,1)$ for score sequences $\hat{\mathbf{D}} = \langle \hat{\mathbf{d}}_k \rangle \in (0,1), 1 \leq k \leq K$. In a neural network-based SED system, $\hat{\mathbf{D}}$ is the output of a sigmoid layer with K units. Such system, with parameters $\boldsymbol{\theta}_{sed}$, can be expressed as

$$\hat{\mathbf{D}} = f^{(sed)}(\mathbf{x}; \boldsymbol{\theta}_{sed}). \tag{1}$$

In contrast with previous editions, the primary metric for evaluation in DCASE 2021 Task 4 is PSDS (Polyphonic Sound Detection Score) [16]. This metric aims to overcome some issues of F1 score, such as the dependence to a specific threshold. Two parameterizations are proposed for PSDS, leading to metrics PSDS1 and PSDS2. These represent different application cases, prioritizing a finer temporal detection (PSDS1) or a more precise event classification (PSDS2). The challenge encourages to optimize each scenario with different systems [17].

A baseline system is provided by DCASE to benchmark the Sound Event Detection performance, based on a convolutional-recurrent neural network (CRNN) trained with mel-spectrogram features and mean teacher [3]. Mean teacher [4] is a semi-supervised method that consists on training two models (student and teacher) at the same time. The teacher has the same topology as the student, and its weights $\boldsymbol{\theta}_t$ are computed as an exponential moving average of the student's weights $\boldsymbol{\theta}_s$ (thus, the teacher does not learn from back-propagation). In addition to a binary cross-entropy supervised loss (L_{sup}), the teacher score sequences are used as self-supervised targets for the student by means of a mean squared error loss (L_{self}) between the student ($\hat{\mathbf{D}}_s$) and the teacher ($\hat{\mathbf{D}}_t$) predictions, allowing the system to learn from examples where ground truth annotations ($\mathbf{D}$) are not available. The total loss function (L_{sed}) of the system is the sum of the supervised and self-supervised losses. A weight α_{self} is used to control the contribution of L_{self} to the total loss.

$$L_{sup} = \mathrm{BCE}(\hat{\mathbf{D}}_s, \mathbf{D}) \tag{2}$$

$$L_{self} = \mathrm{MSE}(\hat{\mathbf{D}}_s, \hat{\mathbf{D}}_t) \tag{3}$$

$$L_{sed} = L_{sup} + \alpha_{self} L_{self} \tag{4}$$

Although the model (student or teacher) to use at test time is not defined in advance, the model selection criterion of the baseline system only tracks the performance of the student model.

A different baseline system is provided for the Sound Event Detection+Separation (SSep+SED) subtask, using the SED baseline system as a pre-trained model, in addition to a pre-trained Source Separation network. The SSep+SED baseline is a weighted combination of two branches: the first branch fine-tunes the pre-trained SED block over the output sources of the pre-trained SSep model, whereas the second branch consists only of the pre-trained SED block. The SSep model in the first branch and the SED model in the second branch are frozen, thus they do not learn during the fine-tuning process. The weight of the combination of the two branches is learnt during the training process.

2.2. Source Separation

Universal Sound Separation [7] aims to separate arbitrary types of sounds in a mixture $\mathbf{x}$ to M different outputs, $\hat{\mathbf{S}} = \langle \hat{\mathbf{s}}_m \rangle, 1 \leq m \leq M$. Therefore, it is a particular task in the field of Source Separation (SSep), and can be expressed as:

$$\hat{\mathbf{S}} = f^{(sep)}(\mathbf{x}; \boldsymbol{\theta}_{sep}) \tag{5}$$

A typical issue when training SSep systems, compared to other machine learning tasks, is that the order of the M output channels is usually irrelevant for the adequacy of the result. This fact is known as the permutation problem, and it is solved by the Permutation Invariant Training (PIT) paradigm [8]. PIT computes the loss function L for all possible permutations of the output ($\hat{\mathbf{S}}$) and target ($\mathbf{S}$) sources using the permutation matrix $\mathbf{P}$, and considers only the minimum loss, assuming that it corresponds to the correct permutation:

$$L_{PIT}(\mathbf{S}, \hat{\mathbf{S}}) = \min_{\mathbf{P}} \sum_{m=1}^{M} L(\mathbf{s}_m, [\mathbf{P}\hat{\mathbf{S}}]_m) \tag{6}$$

In order to train SSep when target sources are not available for the training data, Mixture Invariant Training (MixIT) [9] proposes to train over the sum of two audio examples $(\mathbf{x}_1, \mathbf{x}_2)$. In an analogous way to PIT, MixIT considers the best assignment of the outputs to each example, employing a $2 \times M$ binary matrix, $\mathbf{A}$, in which each column sums to 1:

$$\hat{\mathbf{S}} = f^{(sep)}(\mathbf{x}_1 + \mathbf{x}_2; \boldsymbol{\theta}_{sep}) \tag{7}$$

$$L_{MixIT}(\mathbf{x}_1, \mathbf{x}_2, \hat{\mathbf{S}}) = \min_{\mathbf{A}} \sum_{i=1}^{2} L(\mathbf{x}_i, [\mathbf{A}\hat{\mathbf{S}}]_i) \tag{8}$$

In both cases (PIT and MixIT), the loss function L is a negative signal-to-noise ratio (SNR) between a target and an output source:

$$L_{SNR}(\mathbf{s}, \hat{\mathbf{s}}) = -10 \log_{10} \left(\frac{||\mathbf{s}||^2}{||\mathbf{s} - \hat{\mathbf{s}}||^2} \right) \tag{9}$$

3. PROPOSED METHODS

3.1. Joint Source Separation + Sound Event Detection

We propose a Joint SSep+SED (JSS) model composed of pre-trained SSep and SED blocks, as described in Figure 1. The input audio clip $\mathbf{x}$ is fed to the SSep block, which outputs M waveforms with the estimated sources, $\hat{\mathbf{S}}$. Then, the SED block is applied to each source, obtaining M source-level SED predictions, $\hat{\mathbf{D}}^{(src)}_{1\ldots M}$. Finally, a max-pooling function is applied to the source-level predictions, obtaining clip-level predictions, $\hat{\mathbf{D}}$. The loss function L_{sed} (eq. 4) is employed for training, back-propagating its gradients to update the parameters of the SED block ($\boldsymbol{\theta}_{sed}$) and the SSep block ($\boldsymbol{\theta}_{sep}$). This enables both blocks to adjust to each other, namely the SSep block, pre-trained with the signal-level objective, is now fine-tuned to produce optimal signals for the SED task, whereas the SED block learns to process the separated signals.

The JSS model introduces two main differences with the DCASE SSep+SED baseline: First, JSS is composed of a single branch, omitting the SED-only branch of the baseline. This simplifies the training process, however, the combination with a SED-only model can still be performed in test time. Second, whereas the SSep+SED baseline fine-tunes the SED system over already separated mixtures, in JSS the SSep block is a part of the model itself, meaning that its parameters can be optimized during the training process.

In order to train JSS models, we propose two methods. On the one hand, Joint Training (JT) loads the two pre-trained blocks and trains them together in a single additional process. Alternatively, a Two-stage Training is considered: In Stage 1, only the SED block is updated, fine-tuning the SED block on separated data, as done in the SSep+SED baseline. Afterwards, Stage 2 updates only the SSep block, by back-propagating gradients of L_{sed} through the SED

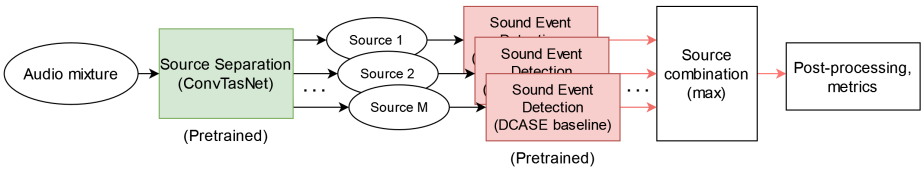

Fig. 1. Structure of the Joint Source separation and Sound event detection (JSS) system. The input is an audio waveform, which is separated into M sources by the Source Separation (SSep) block. Each of the estimated sources is fed into the Sound Event Detection (SED) block, obtaining M sets of SED source-level predictions that are combined into a single set of SED mixture-level predictions by means of a max-pooling function. SED metrics and loss functions are computed over the mixture-level predictions.

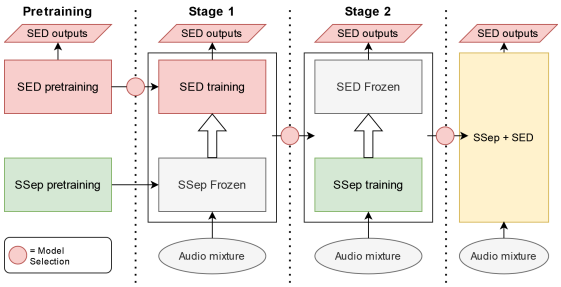

Fig. 2. Diagram of the Two-stage Training method proposed for Joint Source separation and Sound event detection (JSS). Horizontal arrows indicate that the model obtained in a given stage is used as the starting point for the next one.

system. Dividing the training into two stages allows us to better control the process (e.g. check the convergence of each block) and to analyze the contribution of each component in more detail. The Two-stage Training process is illustrated in Figure 2.

3.2. Model selection

Our experiments involve several training processes that require a criterion for model selection. For this purpose, the SED baseline proposes choosing the student model that holds the best performance in validation. However, due to the nature of the mean teacher training, the teacher model often holds a better performance in test time. We also observed that the teachers consistently outperform their corresponding student models in the validation objective metric, thus, we propose applying the model selection criterion over the teacher model.

4. EXPERIMENTS AND RESULTS

4.1. Datasets

Our experimental setup involves three datasets, one for Sound Event Detection (DESED) and two for Source Separation (FUSS and YFCC100M).

DESED (Domestic Environment Sound Event Detection) [18, 19] is the development dataset for DCASE 2021 Task 4. It contains real audio clips (retrieved from Google AudioSet [20]) and synthetic audio soundscapes, which are generated using the Scaper toolkit [21], mixing foreground event recordings from the Freesound Dataset (FSD) [22] and background recordings from SINS [23]. Several subsets are defined, according to the origin of the audio clips and their available annotations: Unlabeled training set (14412 clips), Synthetic training (12500), Weak training (1578), and Validation set (1168 clips).

The FUSS (Free Universal Sound Separation) dataset [24] consists on synthetic audio mixtures, which are artificially generated by overlapping from 2 to 4 isolated recordings of foreground and background sounds. The individual sources are provided as well, in order to use them as targets for Source Separation. FUSS is comprised of 20000 mixtures for training, 1000 mixtures for validation and 1000 mixtures for evaluation.

YFCC100M (Yahoo-Flickr Creative Commons 100 Million) [25] is a dataset of pictures and videos obtained from web sources and licensed for free public use. The dataset contains approximately 0.8 million videos with their corresponding audio. In contrast with FUSS, and due to the origin of the videos, the individual audio sources are not available.

4.2. Model settings

The SED baseline of DCASE 2021 is a CRNN with 7 convolutional layers and 2 Bi-GRU (Bidirectional Gated Recurrent Units), trained using mean teacher [3]. The unlabeled data from DESED, together with a 90% of the weak set and an 80% of the synthetic set, are used for training, whereas the remaining of the weakly-labeled and synthetic sets are employed as validation data for model selection.

The Source Separation model of the SSep+SED baseline is an improved time-domain convolutional network (TDCN++), trained using MixIT over the YFCC100M audio data. The training and validation data is distributed in the same way as in the SED baseline.

Alternatively, the SSep block in our JSS models is a ConvTasNet model [26] with $R = 1$ repeat, $X = 4$ convolutional blocks and $M = 4$ outputs, whereas the SED block is a replica of the SED baseline, with no changes in the structure or the training process. The SED pre-training for JSS uses the same data distribution as the SED baseline. This configuration is also used for the Joint Training and Two-stage Training methods.

We try two pre-training settings for the SSep block: supervised training over FUSS, and unsupervised training with MixIT using DESED. In the latter case, Synthetic and Unlabeled training sets are used for training, and the Weak training set for validation. Both models were trained using the Asteroid toolkit [27].

4.3. Results and discussion

We have performed experiments comparing the Joint Training and Two-stage Training methods, taking as starting point for the SED block the DCASE baseline system, and for the SSep block the ConvTasNet model [26] trained over FUSS (with supervised training) or

	Student model sel.			Teacher model sel.		
	PSDS1	PSDS2	F1(%)	PSDS1	PSDS2	F1(%)
SED baseline	0.338	0.522	40.12	0.357	0.552	41.65
FUSS-S0	0.241	0.336	31.42	0.281	0.380	33.15
FUSS-S1	0.329	0.517	41.03	0.349	0.534	41.47
FUSS-S2	0.344	0.549	42.36	0.356	0.552	42.75
FUSS-JT	0.336	0.535	41.50	0.358	0.547	41.46
DESED-S0	0.249	0.352	33.89	0.273	0.373	35.69
DESED-S1	0.346	0.539	39.25	0.355	0.550	42.41
DESED-S2	0.328	0.529	40.85	0.362	**0.572**	**43.40**
DESED-JT	0.337	0.504	40.77	**0.365**	0.555	43.14

Table 1. Sound Event Detection results of the Two-stage Training (S0, S1, S2) and the Joint Training (JT) methods over the DESED Validation set in terms of PSDS1, PSDS2 and collar-based F1 score. For each method, FUSS and DESED pre-training for the Source Separation block are included, as well as Student and Teacher model selection for the mean teacher training. Our result with the SED baseline system of DCASE 2021 is provided for comparison. The best result for each metric is highlighted in bold.

	PSDS1	PSDS2	F1(%)
SSep-SED baseline	0.363	0.532	44.34
SED Bs + DESED-S2	0.379	**0.590**	43.74
SED Bs + DESED-JT	0.366	0.563	43.02
DESED-S2 + DESED-JT	**0.380**	0.589	**45.52**
SED Bs+DESED-S2+DESED-JT	0.379	0.587	45.05

Table 2. Sound Event Detection results of model fusions over the DESED Validation set in terms of PSDS and collar-based F1 score. All the models in each fusion use Teacher model selection. Our result with the SSep-SED baseline of DCASE 2021 is provided for comparison. The best result for each metric is highlighted in bold.

DESED (with MixIT unsupervised training). Moreover, we considered separate experiments with Student model selection or Teacher model selection, leading to four different settings: FUSS+Student, FUSS+Teacher, DESED+Student and DESED+Teacher.

We divide the results of the Two-stage Training into three steps: Stage 0 (S0) is the initial state of the JSS model, loading the pre-trained blocks without any further training, Stage 1 (S1) is a partial result after the first stage of the training, and Stage 2 (S2) is the final result after both stages.

Following the evaluation setup of DCASE 2021 Task 4, results are provided in terms of PSDS1 and PSDS2 (primary metrics) and collar-based F1 score (supplementary metric) over the DESED Validation set. Our results with the SED baseline system are included as a benchmark of performance.

Table 1 shows the results of Joint Training and Two-stage Training for the four considered settings. It is worth noting that the S0 models hold lower performance than the baseline. This is caused by the domain mismatch between the pre-trained SED block and the separated sources. However, S1 results are comparable to the baseline performance, suggesting that the domain mismatch is solved when the SED block is tuned on separated data. In addition, S2 generally provides further improvement, showing that a fine-tuning of the SSep block can be helpful, even using SED objective functions. On the other hand, the performances of JT and S2 are similar.

When comparing the two proposed pre-training settings for the SSep block, it is found that the unsupervised pretraining on matched data (DESED) allows for similar or better performance than the supervised training on unmatched data (FUSS). Between the student and teacher model selection strategies, the teacher model selection provides better results. Overall, the best results are obtained using the DESED-pre-trained SSep network and teacher model selection: 0.365 PSDS1 (JT), 0.572 PSDS2 and 43.40% F1 score (S2).

In order to compare our methods with the SSep-SED baseline, we computed a score fusion of our proposed models and the SED baseline. For this purpose, we chose the DESED-pre-trained SSep network with Teacher model selection. Results are provided in Table 2, including as a benchmark our results with the SSep-SED baseline.

In contrast with the SSep-SED baseline, which trains the weight of the model combination, our score fusion is a fixed average of the network predictions. Moreover, whereas the SSep-SED baseline requires external data to train SSep, our DESED-SSep block is pre-trained using the training data of the SED task.

Our fusion of the SED baseline and the Two-stage Training outperforms the SSep-SED baseline in terms of PSDS, reaching 0.379 PSDS1 and 0.590 PSDS2, while holding a lower F1 (43.74%). When combining the S2 model with the JT model, similar PSDS results are obtained (0.380 PSDS1 and 0.589 PSDS2), whereas F1 increases to 45.52%, outperforming the SSep-SED baseline . Nevertheless, combining the three models (SED baseline, DESED-S2 and DESED-JT) does not provide further improvements.

Although the results show that JSS is beneficial for SED, the winner systems of the DCASE 2021 challenge yield clearly better performance.[1] This was expected, however, due to the use of large model ensembles and data augmentation in the challenge.

5. CONCLUSIONS

In this paper, we have introduced two training methods for joint Sound Event Detection and Separation. Additionally, we compared supervised and unsupervised pre-training for SSep, as well as two different model selection criteria for mean teacher (student or teacher model selection).

Our experiments show that the proposed methods outperform the DCASE Task 4 baseline system, reaching our best results when using unsupervised pre-training for SSep and teacher model selection. Moreover, score fusion allows to further improve the results.

In future work, the use of specific loss functions for SSep in combination with SED objectives could be explored. Furthermore, it would be interesting to perform experiments with high levels of event overlap and to analyze the SSep performance of the models.

[1] https://dcase.community/challenge2021/task-sound-event-detection-and-separation-in-domestic-environments-results

6. REFERENCES

[1] Toni Heittola, Emre Çakır, and Tuomas Virtanen, *The Machine Learning Approach for Analysis of Sound Scenes and Events*, pp. 13–40, Springer International Publishing, Cham, 2018.

[2] "Detection and classification of acoustic scenes and events (dcase community)," http://dcase.community/.

[3] Nicolas Turpault and Romain Serizel, "Training sound event detection on a heterogeneous dataset," in *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Tokyo, Japan, Nov. 2020, pp. 200–204.

[4] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, 2017, pp. 1195–1204.

[5] Koichi Miyazaki et al., "Conformer-based sound event detection with semi-supervised learning and data augmentation," in *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Tokyo, Japan, Nov. 2020, pp. 100–104.

[6] Xu Zheng, Yan Song, Ian McLoughlin, Lin Liu, and Li-Rong Dai, "An improved mean teacher based method for large scale weakly labeled semi-supervised sound event detection," in *IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021, pp. 356–360.

[7] Ilya Kavalerov, Scott Wisdom, et al., "Universal sound separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 175–179.

[8] Morten Kolbaek, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[9] Scott Wisdom, Efthymios Tzinis, et al., "Unsupervised sound separation using mixture invariant training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3846–3857, 2020.

[10] Qiuqiang Kong, Yuxuan Wang, et al., "Source separation with weakly labelled data: An approach to computational auditory scene analysis," in *IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 101–105.

[11] Efthymios Tzinis, Scott Wisdom, John R. Hershey, Aren Jansen, and Daniel P. W. Ellis, "Improving universal sound separation using sound classification," in *IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 96–100.

[12] Scott Wisdom, Aren Jansen, Ron J. Weiss, Hakan Erdogan, and John R. Hershey, "Sparse, efficient, and semantic mixture invariant training: Taming in-the-wild unsupervised sound separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021, pp. 51–55.

[13] Qiuqiang Kong, Yong Xu, Iwona Sobieraj, Wenwu Wang, and Mark D. Plumbley, "Sound event detection and time–frequency segmentation from weakly labelled data," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 777–787, 2019.

[14] Nicolas Turpault, Scott Wisdom, et al., "Improving sound event detection in domestic environments using sound separation," in *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Tokyo, Japan, Nov. 2020, pp. 205–209.

[15] Samuele Cornell, Michel Olvera, et al., "Task-aware separation for the dcase 2020 task 4 sound event detection and separation challenge," in *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Tokyo, Japan, Nov. 2020, pp. 31–35.

[16] Çağdaş Bilen, Giacomo Ferroni, Francesco Tuveri, Juan Azcarreta, and Sacha Krstulović, "A framework for the robust evaluation of sound event detection," in *IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2020, pp. 61–65.

[17] Diego de Benito-Gorrón, Sergio Segovia, Daniel Ramos, and Doroteo T. Toledano, "Multiple feature resolutions for different polyphonic sound detection score scenarios in dcase 2021 task 4," in *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Barcelona, Spain, Nov. 2021, pp. 65–69.

[18] Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, New York City, United States, 2019.

[19] Romain Serizel, Nicolas Turpault, Ankit Shah, and Justin Salamon, "Sound event detection in synthetic domestic environments," in *IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcelona, Spain, 2020.

[20] Jort F. Gemmeke, Daniel P. W. Ellis, et al., "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017.

[21] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello, "Scaper: A library for soundscape synthesis and augmentation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 344–348.

[22] Eduardo Fonseca, Jordi Pons, et al., "Freesound datasets: a platform for the creation of open audio datasets," in *18th International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017, pp. 486–493.

[23] Gert Dekkers, Steven Lauwereins, et al., "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," in *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2017.

[24] Scott Wisdom, Hakan Erdogan, et al., "What's all the FUSS about free universal sound separation data?," in *IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021, pp. 186–190.

[25] Bart Thomee, David A. Shamma, et al., "YFCC100M: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[26] Yi Luo and Nima Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[27] Manuel Pariente, Samuele Cornell, et al., "Asteroid: the PyTorch-based audio source separation toolkit for researchers," in *Proc. Interspeech*, 2020.