SOUND EVENT LOCALIZATION AND DETECTION FOR REAL SPATIAL SOUND SCENES: EVENT-INDEPENDENT NETWORK AND DATA AUGMENTATION CHAINS

Jinbo Hu[1,2]*, Yin Cao*[3]*, Ming Wu*[1]*, Qiuqiang Kong*[4]*, Feiran Yang*[1]*, Mark D. Plumbley*[5]*, Jun Yang*[1,2]

[1]Key Laboratory of Noise and Vibration Research, Institute of Acoustics,
Chinese Academy of Sciences, Beijing, China, {hujinbo, mingwu, feiran, jyang}@mail.ioa.ac.cn
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Xi'an Jiaotong Liverpool University, Suzhou, China, yin.k.cao@gmail.com
[4]ByteDance Shanghai, China, kongqiuqiang@bytedance.com
[5]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
m.plumbley@surrey.ac.uk

ABSTRACT

Sound event localization and detection (SELD) is a joint task of sound event detection and direction-of-arrival estimation. In DCASE 2022 Task 3, types of data transform from computationally generated spatial recordings to recordings of real-sound scenes. Our system submitted to the DCASE 2022 Task 3 is based on our previous proposed Event-Independent Network V2 (EINV2) with a novel data augmentation method. Our method employs EINV2 with a track-wise output format, permutation-invariant training, and a soft parameter-sharing strategy, to detect different sound events of the same class but in different locations. The Conformer structure is used for extending EINV2 to learn local and global features. A data augmentation method, which contains several data augmentation chains composed of stochastic combinations of several different data augmentation operations, is utilized to generalize the model. To mitigate the lack of real-scene recordings in the development dataset and the presence of sound events being unbalanced, we exploit FSD50K, AudioSet, and TAU Spatial Room Impulse Response Database (TAU-SRIR DB) to generate simulated datasets for training. We present results on the validation set of Sony-TAu Realistic Spatial Soundscapes 2022 (STARSS22) in detail. Experimental results indicate that the ability to generalize to different environments and unbalanced performance among different classes are two main challenges. We evaluate our proposed method in Task 3 of the DCASE 2022 challenge and obtain the second rank in the teams ranking. Source code is released[1].

Index Terms— Sound event localization and detection, real spatial sound scenes, Event-Independent Network, data augmentation chains, simulated datasets

1. INTRODUCTION

Sound event localization and detection (SELD) consists of sound event detection (SED) and direction-of-arrival (DoA) estimation. SED aims to detect the presence and types of sound events, and DoA estimation predicts the spatial locations of different sound sources. SELD characterizes sound sources in a spatial-temporal manner. SELD plays an important role in a wide range of applications, such as robot auditory and surveillance of intelligent home.

SELD has received broad attention recently. Adavanne et al. [1] proposed a polyphonic SELD approach using an end-to-end network, SELDnet, which was utilized for a joint task of SED and regression-based DoA estimation. SELD was then introduced in Task 3 of the Detection and Classification of Acoustics Scenes and Events (DCASE) 2019 Challenge for the first time, which uses the TAU Spatial Sound Events 2019 dataset [2]. Most datasets of spatial sound events are computationally simulated and these recordings are generated by convolving randomly chosen sound event examples with a corresponding random real-life spatial room impulse response (SRIR) to spatially place them at a given position [2–4]. To bring each iteration of Task 3 of DCASE Challenge closer to real conditions, stronger reverberation, diversity of environment, dynamic scenes with both moving and static sound sources, ambient noise, sound events of the same type, and unknown directional interfering events out of the target classes were added into datasets to complicate the SELD task. In 2022, the challenge transforms from computationally simulated spatial recordings to real spatial sound scene recordings. The Sony-TAu Realistic Spatial Soundscapes 2022 (STARSS22) dataset is manually annotated and released to serve as the development and evaluation dataset of DCASE2022 Task 3 this year [5].

SELDnet is unable to detect sound events of the same type but with different locations [1], which is also called homogeneous overlap. An event-independent network (EIN) with a track-wise output format was proposed to detect the homogeneous overlap problem [6–8]. In EIN, there are several event-independent tracks, and each track can be of any event. The number of tracks needs to be pre-determined according to the maximum number of overlapping events. EINV2, an improved version of EIN, utilizes multi-head self-attention (MHSA) and a soft parameter-sharing strategy of multi-task learning to achieve better performance [7].

The training set often deviates from real-scene spatial and acoustical environments, and mismatched distribution of locations and sound types between the training set and test set is common. A novel data augmentation method is used to generalize the model [8, 9]. The data augmentation method contains several data augmentation chains. These data augmentation chains consist of some randomly sampled data augmentation operations. The augmentation method can increase the diversity of augmented features.

In this study, our system is based on our previous proposed EINV2 with data augmentation chains. EINV2 is extended by

[1]https://github.com/Jinbo-Hu/DCASE2022-TASK3

Conformer, which is a combination structure of self-attention and convolution. The data augmentation method is composed of several augmentation operations. These data augmentation operations are sampled and layered randomly to combine to several data augmentation chains [8]. External data is allowed in this challenge. We generate simulated data by randomly convolving chosen samples of sound events from AudioSet [10] and FSD50K [11] with measured SRIRs from TAU Spatial Room Impulse Responses Database[2] (TAU-SRIR DB). The experimental results show the proposed model with the novel data augmentation method, which was trained on our simulated data, outperforms the DCASE2022 challenge Task 3 baseline model which was trained on official synthetic SELD mixtures[3]. In addition, we present class-wise and room-wise metric scores of the validation set of STARSS22 in detail. The proposed system obtains the second rank in Task 3 of DCASE 2022 Challenge[4].

2. THE METHOD

2.1. Input features

In this method, log-mel spectrograms and intensity vectors (IV) in log-mel space are used for features of the SELD task. First order ambisonics (FOA) include four-channel signals, i.e., omnidirectional channel $\mathbf{w}$, and three directional channels $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$. Log-mel spectrograms are computed from the mel filter banks and the short-time Fourier transform spectrograms, and IVs are cross-correlation of log-mel spectrograms of $\mathbf{w}$ with $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ [12]. These features are directly calculated online using a 1-D convolutional layer, which supports data augmentation on raw waveform.

2.2. Network Architecture

The track-wise output format was introduced in our previous works [6–8]. It can be defined as

$$\boldsymbol{Y}_{\text{Trackwise}} = \left\{ (y_{\text{SED}}, y_{\text{DoA}}) \mid y_{\text{SED}} \in \mathbb{O}_{\mathbf{S}}^{M \times K}, y_{\text{DoA}} \in \mathbb{R}^{M \times 3} \right\} \quad (1)$$

where M is the number of tracks, K is the number of sound-event types, $\mathbb{O}_{\mathbf{S}}^{M \times K}$ is one-hot encoding of K classes, and $\mathbf{S}$ is the set of sound events. Cartesian DoA estimation is used here.

The number of tracks is determined by the maximum polyphony. Each track can only detect a sound event with a corresponding direction of arrival. While a model with a track-wise output format is trained, sound events may be predicted in any track, instead of a fixed track. It may cause the track permutation problem that sound events predicted and their ground truth may not be aligned in a fixed track. Permutation-invariant training (PIT) is proposed to tackle the problem effectively. The PIT loss is defined as

$$\mathscr{L}_{PIT}(t) = \min_{\alpha \in \mathbf{P}(t)} \sum_{M} \left\{ \lambda \cdot \ell_{\alpha}^{\text{SED}}(t) + (1 - \lambda) \cdot \ell_{\alpha}^{\text{DoA}}(t) \right\} \quad (2)$$

where $\alpha \in \mathbf{P}(\mathbf{t})$ indicates one of the possible permutations and λ is a loss weight between SED and DoA. $\ell_{\alpha}^{\text{SED}}$ is binary cross entropy loss for the SED task, and $\ell_{\alpha}^{\text{DoA}}$ is mean square error for the DoA task. The lowest loss will be chosen by finding a possible permutation, and the back-propagation is then performed.

[2] https://doi.org/10.5281/zenodo.6408611
[3] https://doi.org/10.5281/zenodo.6406873
[4] https://dcase.community/challenge2022

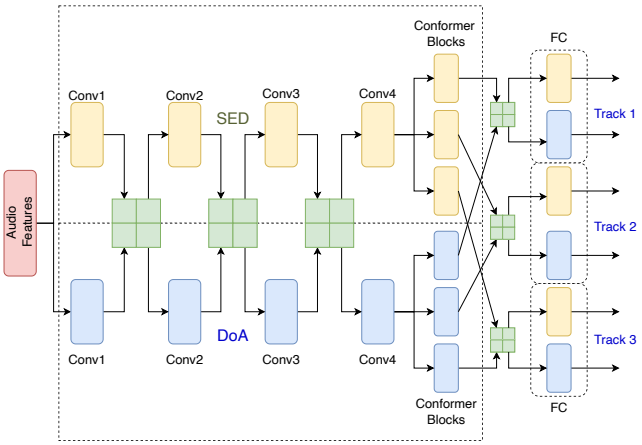

Figure 1: The architecture of the SELD network, which is a Conv-Conformer network. The upper half (yellow boxes) is the SED task. The lower half (blue boxes) is the DoA estimation task. The green boxes sandwiched between SED branch and DoA branch indicate soft connections between SED and DoA estimation.

From multi-task learning (MTL) perspective, joint SELD learning can be mutually beneficial. Hard parameter-sharing (PS) and soft PS are two typical methods to implement MTL. Hard PS means subtasks use the same feature layers, while soft PS means subtasks use their own feature layers with connections existing among those feature layers. In [7], experimental results show that soft PS using cross-stitch is more effective.

EINV2, which combines the track-wise output format, PIT, and soft PS, is utilized in our system. Three tracks are adopted to address up to three overlapped sound events. Multi-head self-attention (MHSA) blocks are replaced with Conformer blocks. Conformer consists of two feed-forward layers with residual connections sandwiching the MHSA and convolution modules, and hence has the ability to capture global and local patterns. [8, 13]. Our proposed network is shown in Fig. 1.

2.3. Data Augmentation Chains

The main characteristic of our data augmentation method is using some augmentation chains [8,9,14]. These augmentation chains are combined by some augmentation operations, which are randomly selected and linked in chain. We randomly sample $k = 3$ augmentation chains. Augmentation operations that are used here include Mixup [15], Cutout [16], SpecAugment [17], and frequency shifting [18]. Rotation of FOA signals [19] is an additional augmentation method, but excluded by data augmentation chains. The diagram of data augmentation chains is shown in Fig. 2.

Mixup utilize convex combinations of pairs of feature vectors and their labels to train the model. Mixup on both raw waveform and spectrograms is used here to improve the ability of detecting overlapping sound events. While random Cutout produces several rectangular masks on spectrograms, SpecAugment produces stripes masks on time and frequency dimension of spectrograms. Frequency shifting in the frequency domain is similar to pitch shift in the time domain, and it randomly shifts input features of all the channels up or down along the frequency dimension by several bands. We also use a spatial augmentation method, rotation of FOA signals. It rotates FOA format signals by channel swap to enrich DoA labels. This method does not lose physical relationships

Table 1: The SELD performance of our proposed system. The training set of STARSS22 is mixed into synthetic training set by default.

System	Datasets	Validation set				Evaluation (Blind test) set			
		ER_{20°	F_{20°	LE_{CD}	LR_{CD}	ER_{20°	F_{20°	LE_{CD}	LR_{CD}
Baseline FOA [5]	Official	0.71	21.0%	29.3°	46.0%	0.61	23.7%	22.9°	51.4%
EINV2 w/o dataAug chains	Official	0.75	32.3%	24.0°	56.1%	-	-	-	-
EINV2 w/ dataAug chains	Official	0.56	42.4%	19.3°	61.4%	-	-	-	-
System #1	A+B+C	0.50	48.4%	19.5°	65.7%	0.44	49.2%	16.6°	70.4%
System #2	A+B	0.50	51.0%	16.4°	65.9%	0.40	57.4%	15.1°	70.6%
System #3	A	0.53	48.1%	17.8°	62.6%	0.39	55.8%	16.2°	72.4%
System #4	B	0.53	45.4%	17.4°	62.5%	0.40	50.9%	15.9°	69.4%

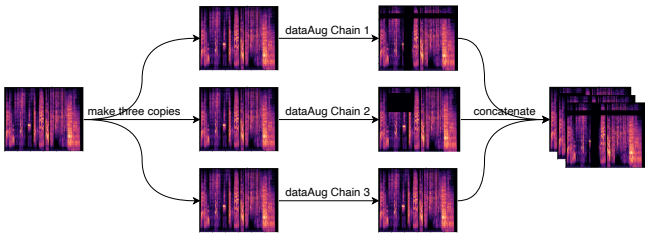

Figure 2: Diagram of data augmentation chains

between sound sources and observers. We use z-axis as the rotation axis to swap directional channel $\mathbf{x}$ and $\mathbf{y}$, which leads to 16 types of channel rotation.

2.4. Simulated Data

Manual annotations are expensive and the duration of STARSS22 (about 5 hours of the development set) is limited compared with the synthetic datasets (about 13 hours synthetic recordings in DCASE 2021) used in previous years, therefore, external datasets are used to improve the model performance. We generated simulated data using the generator code[5] provided by DCASE 2022.

Samples of sound events are mainly sourced from FSD50K dataset, based on affinity of the labels in that dataset to the target classes. The target class *background music* and the interference class *shuffling cards* are not in FSD50K dataset, therefore, we use AudioSet as a supplement. Spatial events were spatialized in 9 unique rooms, using collected SRIRs from the TAU-SRIR DB dataset. The ambient noise from the same room was additionally mixed at varying signal-to-noise ratios (SNR) ranging from 30 dB to 6 dB. The maximum polyphony of target classes is 3, excluding additional polyphony of interference classes.

We select sound event samples whose labels significantly corresponded to the target classes. Each sound event sample also has a different energy gain for mixing. By setting different ranges of gain and choosing different samples, we generate three datasets, A, B, and C. All of these synthetic datasets have 2700 1-minute clips.

3. EXPERIMENTS

3.1. Datasets

The STARSS22 dataset contains recordings of real scenes, and the density of sound event samples and the presence of each class varies greatly. The maximum number of the overlaps is 5, but those cases are very rare [5]. The overlap of 4 and 5 accounts for the proportion of 1.8% in total. Occurrences of up to 3 simultaneous events are fairly common, so we ignore the case scenarios that the number of overlapping events is more than 3. During the development stage, we train our proposed model on mixed datasets of synthetic recordings and the training set of STARSS22, and evaluate those systems using the validation set of STARSS22. During the evaluation stage, both synthetic recordings and all of the development set of STARSS22 are used for training.

[5] https://github.com/danielkrause/DCASE2022-data-generator

3.2. Hyper-parameters

Audio clips are segmented to have a fixed length of 5 seconds with no overlap for training and inference. Log-mel spectrograms and intensity vectors features, with 24 kHz sampling rate, a 1024-point Hanning window with a hop size of 400, and 128 mel bins, are extracted from these audio segments. AdamW optimizer is used. The learning rate is set to 0.0003 for the first 70 epochs and then decreased to 0.00003 for the following 20 epochs. The threshold for SED is set to 0.5 to binarize predictions. The loss weight λ is 0.5.

3.3. Evaluation Metrics

We use the official evaluation metrics to evaluate the SELD performance [20, 21]. The evaluation metrics use a joint metric of localization and detection: location-sensitive F-score ($F_{\leq T^\circ}$), error rate ($ER_{\leq T^\circ}$), and class-sensitive localization recall (LR_{CD}), localization error (LE_{CD}). T° means spatial threshold and is set to 20° in this challenge. $F_{\leq T^\circ}$ and $ER_{\leq T^\circ}$ consider true positives predicted under a spatial threshold T° from the ground truth. For LE_{CD} and LR_{CD}, the detected sound class has to be correct in order to count the corresponding localization predictions.

In the previous challenges, the evaluation metrics were micro-averaged, which gives equal weight to each individual decision and the performance is affected by the classes with more samples. In this challenge, macro-averaging of evaluation metrics is used. Macro-averaging gives equal weight to each class and emphasizes the system performance on the smaller classes [22].

3.4. Experimental Results

Table 1 summarizes the performance of our proposed systems. The official dataset means the synthetic mixtures for baseline training. The system baseline, EINV2 without dataAug chains, and EINV2 with dataAug chains all use the same dataset for training. EINV2

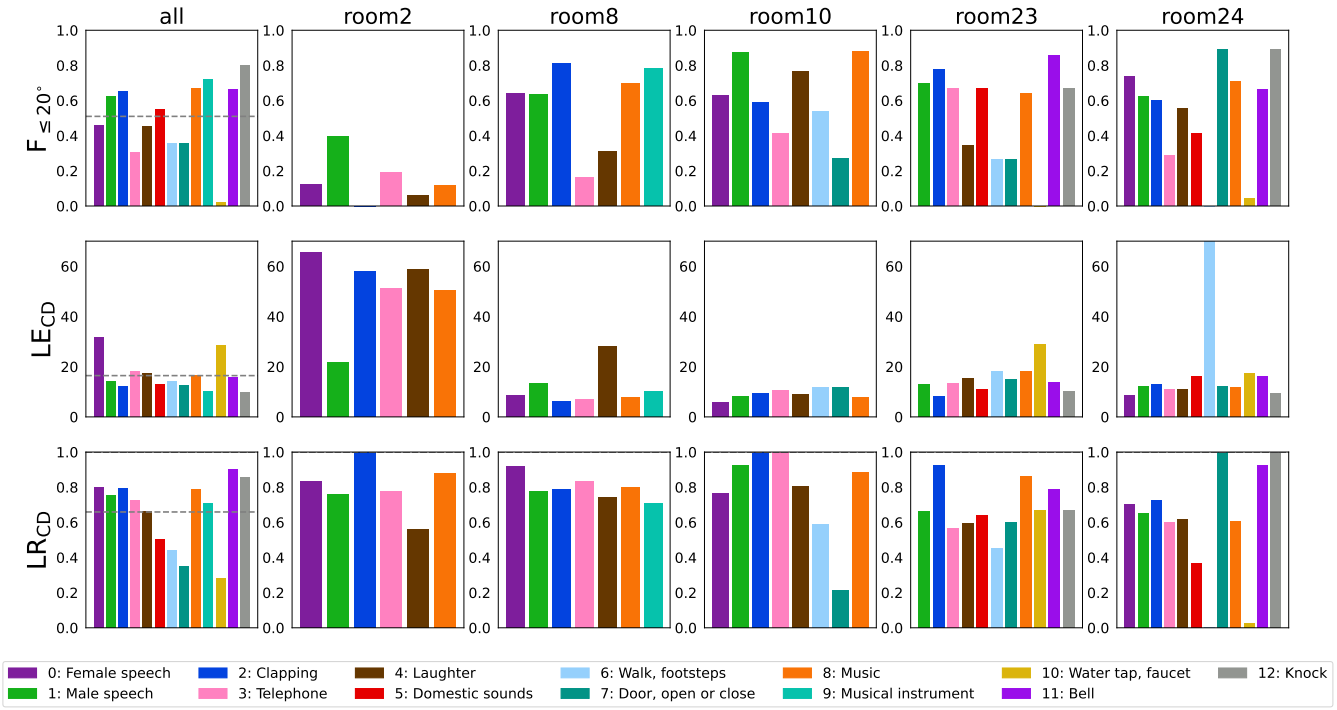

Figure 3: Metric scores of System #2 on validation set of STARSS22 in detail. The first column shows metric scores of the whole validation set. The following columns present metric scores of each room of validation set.

without data augmentation chains outperforms the baseline model, whereas EINV2 with data augmentation chains performs better.

All configurations of systems #1 - #4 are the same as system EINV2 with dataAug chains, except for the training set used. The results also demonstrate the effectiveness of our simulated data over the official dataset.

The first column of Fig. 3 shows class-wise metric scores of System #2 on the validation set of STARSS22. The class-wise performance on the whole validation set is highly skewed, with $F_{\leq 20^\circ}$ of *knock* class being 80.0%, whereas $F_{\leq 20^\circ}$ of *water tap and faucet* class being 2.2%. LE_{CD} of *female speech* class and *water tap and faucet* class is a lot higher than average. Other columns of Fig. 3 present class-wise performance for each room. Unbalanced class-wise performance among different rooms results in the skewed class-wise performance on the whole validation set.

The performance of the localization in room 2 is the worst among all the rooms, resulting in a directly significant increase of LE_{CD} of *female speech* class. It may be attributed to small room size of room 2 compared with other rooms. LR_{CD} of *walk, footsteps* (0.0%) class and *water tap and faucet* (2.4%) class in room 24 is very low. A possible reason is the low quality of synthetic training samples, because we ignore the natural temporal occurrences and spatial connections of some types of sounds happening in real scenes when simulating data [5]. For example, the target class *water tap and faucet* and the directional interference class *dishes, pots, and pans* often occur simultaneously in room 24, which leads to many observed false negatives of the class *water tap and faucet* in the system output. It is difficult to synthesis training samples that contains the temporal and spatial relationships of sound events in real scenes. These factors can lead to performance degradation.

4. CONCLUSION

We have presented an approach using an Event-Independent Network V2 (EINV2) with a novel data augmentation method for real-life sound event localization and detection. EINV2 is extended by conformer blocks. The novel data augmentation method contains several augmentation chains, which are stochastic combinations of data augmentation operations. For this challenge, we synthesized more training samples which are convolved using sound events from FSD50k and AudioSet with measured room impulse responses from TAU-SRIR DB. Our model with data augmentation chains performs better than the baseline model. Furthermore, experimental results show further improvement with our synthetic datasets. We also show results on the validation set of STARSS22 in detail. Our proposed method is evaluated in the evaluation set of STARSS22, and obtained the second best team in Task 3 of DCASE 2022 Challenge. The study of the generalization ability to different environments and the performance for unbalanced classes will be analyzed further in the future work.

5. ACKNOWLEDGEMENT

This work was partly supported by Frontier Exploration project independently deployed by Institute of Acoustics, Chinese Academy of Sciences (No. QYTS202009), UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/T019751/1 “AI for Sound”. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

6. REFERENCES

[1] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Sel. Top. Signal Process.*, vol. 13, pp. 34–48, 2018.

[2] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Proc. DCASE 2019 Workshop*, 2019, pp. 10–14.

[3] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Proc. DCASE 2020 Workshop*, 2020, pp. 165–169.

[4] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," in *Proc. DCASE 2021 Workshop*, 2021, pp. 125–129.

[5] A. Politis, K. Shimada, P. Sudarsanam, S. Adavanne, D. Krause, Y. Koyama, N. Takahashi, S. Takahashi, Y. Mitsufuji, and T. Virtanen, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *arXiv:2206.01948*, 2022.

[6] Y. Cao, T. Iqbal, Q. Kong, Y. Zhong, W. Wang, and M. D. Plumbley, "Event-independent network for polyphonic sound event localization and detection," in *Proc. DCASE 2020 Workshop*, 2020, pp. 11–15.

[7] Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang, and M. D. Plumbley, "An improved event-independent network for polyphonic sound event localization and detection," in *Proc. IEEE ICASSP 2021*, 2021, pp. 885–889.

[8] J. Hu, Y. Cao, M. Wu, Q. Kong, F. Yang, M. D. Plumbley, and J. Yang, "A track-wise ensemble event independent network for polyphonic sound event localization and detection," in *Proc. IEEE ICASSP 2022*, 2022, pp. 9196–9200.

[9] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "AugMix: A simple data processing method to improve robustness and uncertainty," in *Proc. ICLR 2020*, 2020.

[10] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, 2017, pp. 776–780.

[11] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: an open dataset of human-labeled sound events," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 30, pp. 829–852, 2021.

[12] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, no. 1, pp. 107–151, 2022.

[13] A. Gulati, J. Qin, 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 *Proc. Interspeech 2020*, 2020, pp. 5036 – 5040.

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML 2020*, 2020, pp. 1597–1607.

[15] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. ICLR 2018*, 2018.

[16] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. of AAAI 2020*, vol. 34, no. 07, 2020, pp. 13 001–13 008.

[17] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613 – 2617.

[18] T. T. N. Nguyen, K. N. Watcharasupat, K. N. Nguyen, D. L. Jones, and W.-S. Gan, "SALSA: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 30, pp. 1749–1762, 2022.

[19] L. Mazzon, Y. Koizumi, M. Yasuda, and N. Harada, "First order ambisonics domain spatial augmentation for DNN-based direction of arrival estimation," in *Proc. DCASE 2019 Workshop*, 2019, pp. 154–158.

[20] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in DCASE 2019," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 29, pp. 684–698, 2020.

[21] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *Proc. IEEE WASPAA 2019*, 2019, pp. 333–337.

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