SOUND EVENT LOCALIZATION AND DETECTION BASED ON ADAPTIVE HYBRID CONVOLUTION AND MULTI-SCALE FEATURE EXTRACTOR

Xinghao Sun[1,3]*, Ying Hu*[1,3]*, Xiujuan Zhu*[1,3]*, Liang He*[1,2]*,*

[1] School of Information Science and Engineering, Xinjiang University, Urumqi, China
{xh_sun2019}@stu.xju.edu.cn
[2] Tsinghua National Laboratory for Information Science
and Technology, Department of Electronic Engineering, Tsinghua University, China
[3] Key Laboratory of Signal Detection and Processing in Xinjiang, China

ABSTRACT

Sound event localization and detection (SELD), which jointly performs sound event detection (SED) and sound source localization(SSL), detects the type and occurrence time of sound events as well as their corresponding direction-of-arrival (DoA) angles simultaneously. In this paper, we propose a method based on Adaptive Hybrid Convolution (AHConv) and multi-scale feature extractor. The square convolution shares the weights in each of the square areas in feature maps making its feature extraction ability limited. In order to address this problem, we propose a AHConv mechanism instead of square convolution to capture the dependencies along with the time dimension and the frequency dimension respectively. We also explore a multi-scale feature extractor that can integrate information from very local to exponentially enlarged receptive field within the block. In order to adaptive recalibrate the feature maps after the convolutional operation, we design an adaptive attention block that is largely embodied in the AHConv and multi-scale feature extractor. On TAU-NIGENS Spatial Sound Events 2021 development dataset, our systems demonstrate a significant improvement over the baseline system. Only the first-order Ambisonics (FOA) dataset was considered in this experiment.

Index Terms— DCASE2021, Sound source localization, Sound event detection, Adaptive hybrid convolution

1. INTRODUCTION

Sound Event Localization and Detection refers to the problem of identifying the presence of independent or temporally-overlapped sound sources, correctly identifying to which sound class it belongs, and estimating their spatial directions while they are active. In realistic aural environments, there are numerous co-occurring different sounds emitted from the sources distributed in space. Even humans cannot all correctly identify and locate multiple sources of sound, so it is very challenging for machines. To solve the SELD problem, two key issues denoted as sound event detection (SED) [1–5] and sound source localization (SSL) [6–13] have to be addressed.

The methodology proposed in this paper is based on the SELD-Net proposed by Adavanne et al [14]. A convolutional recurrent neural network (CRNN) model was proposed for joint SSL and SED of multiple overlapping sound events in three-dimensional space. The phase and magnitude of spectrogram were calculated separately on each audio channel as input features. In order to learn both inter-channel and intra-channel features, the input was fed through three consecutive convolutional blocks. Bidirectional Gate Recurrent Unit (BiGRU) was used for temporal context information learning. The output of the BiGRU is fed into two parallel branches of fully-connected blocks. The classes for all sound events would be output on each time-frame, and the sound source would be located in the three-dimensional Cartesian coordinate system.

Compared with DCASE2020 challenge task 3, the main difference is the emulation of scene recordings with a more natural temporal distribution of target events and, more importantly, the inclusion of directional interferences, meaning sound events out of the target classes that are also point-like in nature. For each reverberant environment and every emulated recording, Interferences are spatialized in the same way as the target events, resulting in recordings that are more challenging and closer to real-life conditions. The other difference is the elimination of the dedicated event classification output branch, by adopting the activity-coupled cartesian direction of arrival (ACCDOA) training target which unifies the localization and classification losses in a homogenous regression vector loss, pioneered by Shimada et al [15].

In this paper, we also propose a CRNN framework based on SELD-Net architecture. We adopt Adaptive Hybrid Convolution (AHConv) mechanism and multi-scale feature extractor to handle feature learning insufficiently. The logmel spectrogram and normalized sound intensity vector are extracted as input features. Instead of conventional square convolution, the AHConv structure is design to process richer spatial features and increase feature diversity by asymmetric convolution. We adopt a multi-scale feature to extract strategy that was designed to capture the longer temporal context information than the conventional convolution. Moreover, the parallel structure is applied in adaptive attention block which adaptive mitigates interference between the channel-wise and time-frequency-wise by exploring two different branches. Additionally, the adaptive attention block can also promote the robustness when a single branch is disturbed by the ambient noise without the presence of sound events. Furthermore, we conduct experiments on TAU-NIGENS Spatial Sound Events 2021 development dataset to verify the effectiveness of our proposed method.

This paper is organized as follow: we will introduce the proposed method in Section 2. The experiment setup will be stated in Section 3. The development results compared with the baseline method will be described in Section 4. Finally, we draw a conclusion and future work in Section 5.

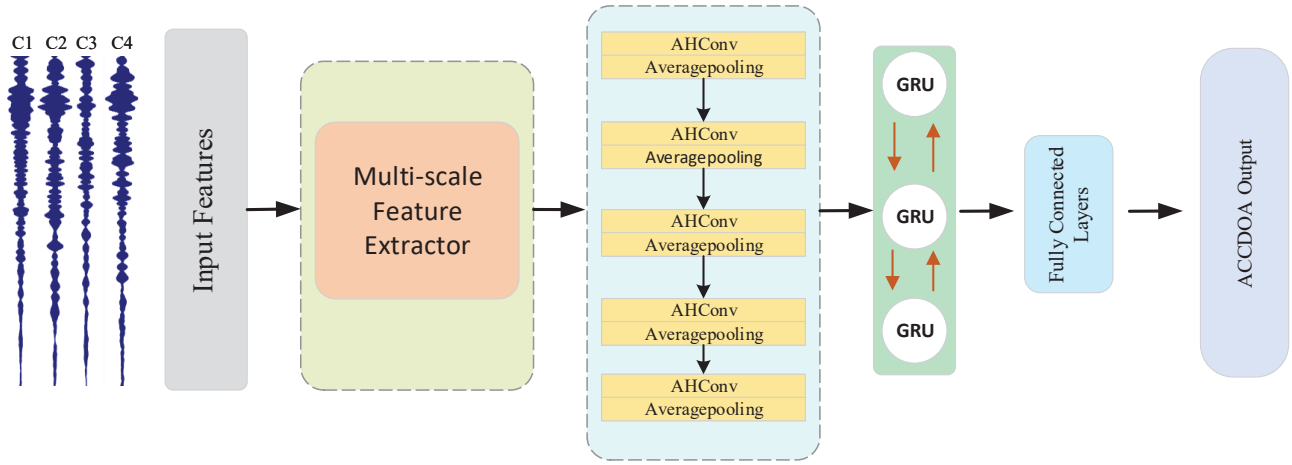

Figure 1: The overall of our proposed method.

2. PROPOSED METHOD

We proposed a method with Adaptive Hybrid Convolution (AHConv) and multi-scale feature extractor which achieves great performance to deal with SELD task in the noisy and reverberant scenes. The proposed network can predict the sound event classes active for each of the input frames along with their respective spatial location, and produce the temporal activity and DOA trajectory for each sound event class. The network diagram is illustrated in Fig. 1. For the multichannel audio, the logmel spectrogram and sound intensity vector are extracted as the input features of the network. The multi-scale feature extractor as depicted in Fig. 2, then followed five AHConv blocks and five average pooling layers. After that, the time dimension is downsampled 5 times and the frequency dimension is downsampled 32 times. Bidirectional Gated Recurrent Unit (Bi-GRU) is used to learn the temporal context information. This is followed by fully connected layers. We adopt the ACCDOA output which unifies the SED and SSL losses into a single homogeneous regression loss.

2.1. Multi-scale Feature Extractor

Among the various CNN architectures, if the network contains shorter connections between layers close to the input and those close to the output, it can be substantially deeper, more accurate, and efficient to train, to further improve the information flow between layers [16]. In this work, we combine the advantages of DenseNet and dilated convolution, and propose an extractor called multi-scale feature extractor. To properly combine DenseNet with the dilated convolution [17], we propose a multi-scale feature extractor that has a multiple dilation factor within a single layer. The dilation rate depends on which skip connection the channels come from, as shown in Fig. 2. The output of each dilated layer is fed into an adaptive attention block. The adaptive attention block reweighs the information of channel-wise and of spatial-wise dimension. That can enhance the important features and weaken the less important features. The outputs of the lth layer x_l receives the feature-maps of all preceding layers express as:

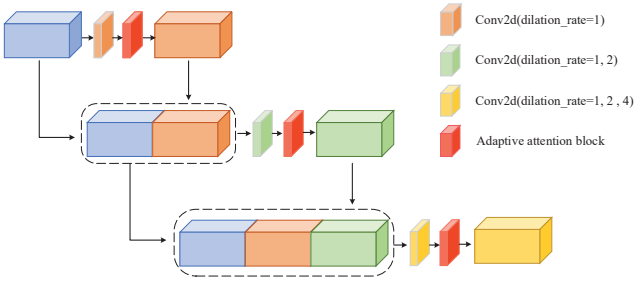

Figure 2: Multi-scale feature extractor, the feature maps of each layer are concatenated together, and the dotted box indicates the concatenate operation

$$x_l = \psi([x_0, x_1, x_2, ...x_{l-1}] \circledast k_l^{d=1,2,...2^{l-1}}) \quad (1)$$

where $[x_0, x_1, x_2, ...x_{l-1}]$ denotes the concatenation of the feature maps from $1, ...l-1$ layers, ψ is a nonlinear transformation consisting of batch normalization (BN) followed by ReLU and dilated convolution with the k_l kernel, $\circledast$ denotes convolution operation and d is the dilated rate in each layer.

2.2. Adaptive Hybrid Convolution

Some of the prior works [18, 19] have shown that a standard square convolutional layer with a filter size of k × k can be factorized as a sequence of two layers with k × 1 and 1 × k filters to reduce network complexity and lighten the computational burden. This asymmetrical convolutional [18] structure is better than a square convolutional structure for processing more and richer spatial features and increasing feature diversity. In addition, asymmetric convolution can obtain faster calculation speed and smaller parameter amounts while ensuring performance. The weight learning of the square convolution relies on the network but is limited by the size of the filter. Therefore, the square convolution is not captured fine-grained time-frequency features. In order to address this problem,

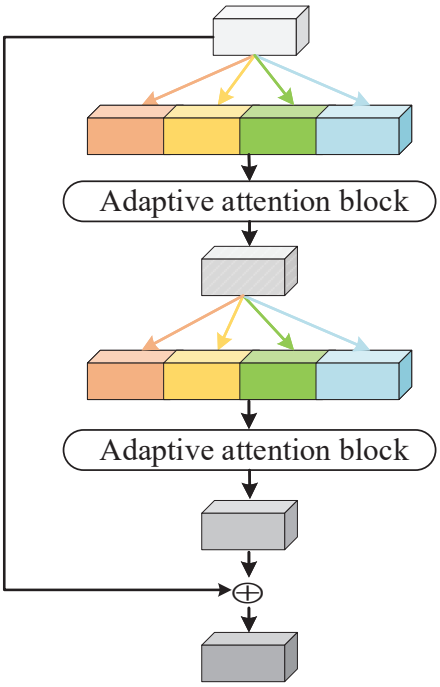

Figure 3: Adaptive Hybrid Convolution (AHConv). Each color represents a different convolution kernel, and the squares represent the convolution graph.

we propose a hybrid convolution mechanism based on the asymmetric convolutional structure, as shown in Fig. 3.

A parallel structure is composed of a filter size 1×3 and 1×5 for time frames, and a filter size 3×1 and 5×1 for frequency bins, thus the time dependency and frequency dependency are capture respectively. Then, the feature maps concatenated along the channel dimension will undergo an adaptive attention block to select the feature adaptively according to the importance. The output of the adaptive attention module will be fed into four asymmetric convolutions and one adaptive convolution again, and the importance of features will be marked more accurately. Finally, in order to preserve the original feature information, we add the original input to the recalibrated output.

2.3. Adaptive Attention block

We design an adaptive attention block as seen in Fig 4. The up half part denotes the path of channel attention (CA) [20], and the lower half part the time-frequency attention (TFA) [21]. In the channel attention path, Global Average Pooling (GAP) converts the information of the TF field of each channel into a value that has the overall information of the channel. To make full use of the aggregated information in the GAP operation, we follow it with fully connected convolution which aims to capture channel-wise dependencies. In the time-frequency attention path, a 2-D convolutional layer with (1,1) kernel size is employed to obtain the global feature maps across the time-frequency (TF) domain. Then sigmoid activation limits the values in the range of (0,1).

After that, different weights are applied to the channel and the TF domain, which can guide the network to pay different attention

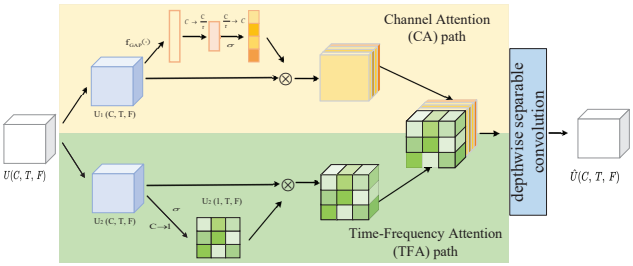

Figure 4: Adaptive attention block.

to the features of channel-wise and time-frequency-wise. The feature maps of each part will concatenate along the channel dimension and then pass through a depthwise separable convolution [22]. The depthwise separable convolution can not only adaptively capture useful information between channels, but also reduce the number of operational parameters. The adaptive attention block is largely embodied in the AHConv and multi-scale feature extractor.

3. EXPERIMENT SETUP

3.1. Dataset

The development set of TAU-NIGENS Spatial Sound Events 2021 has two types of data, one is 4 channel directional microphone array (MIC) from the tetrahedral array and the other one is first-order ambisonic (FOA) data. We used the FOA format for the challenge. The SELD development dataset consists of 600 one-minute audio clips divided into training, validation, and test set of size 400, 100, and 100 clips, respectively. The development dataset is distributed between 12 classes of alarm, crying baby, crash, barking dog, footsteps, knocking on door, female speech, male speech, female scream, male scream, ringing phone and piano. Additionally, dry recordings of disparate sounds not belonging to any of those classes are also spatialized in the same way to serve as directional interference. The sounds are sourced from the running engine, burning fire, and general classes of NIGENS database [23]. The source position for a static event is drawn randomly from the pool of spatial room impulse responses (SRIRs) of a single room used in that recording, while moving events are synthesized for one of the measured trajectories in the room.

3.2. Evaluation metrics

The performance of our proposed model is evaluated by the individual metrics for SED task and SSL task. Standard polyphonic SED metrics, F-score (F1) and error rate (ER) across segments of one second without overlapping are utilized [24]. The DOA estimation in the SSL task was evaluated using frame-wise metrics [25] of DOA error (DE) and frame recall (FR). Considering that a TP is predicted only when the spatial error for the detected event is within the given threshold of 20° deviates from the reference, ER and F1 replaced with ER_{20° and F_{20°. Classification-dependent localization metrics are computed only across each class, instead of across all outputs, DE and FR are replaced with LE_{CD} and LR_{CD}. A more detailed description can be obtained in [25, 26].

3.3. Training procedure

The sampling frequency was used at 24 kHz in our method. Extracting log-mel spectrograms in 64 melbands from 1024-point FFTs, using a 40 ms window and 20 ms hop length. We use a batchsize of 64. Moreover, to ensure a fair comparison, all models were trained for 500 epochs with the Adam optimizer of the same initialized parameters. An early stopping mechanism is used to avoid overfitting during training, where the training is stopped if no improvements on validation split for 50 epochs.

4. RESULT AND DISCUSSION

In this section, we will describe and discuss the experimental results. Firstly, we explored the most appropriate combination of asymmetric convolution for AHConv, and then analyzed the effect of dilated convolution and adaptive attention block in the multi-scale feature extractor with ablation experiments. All the experiments were performed without data augmentation.

Table 1: Explore the combination type of AHConv (+A denotes adding adaptive attention block)

The type of combination	ER_{20°	F_{20°**(%)**	LE_{CD}	LR_{CD}**(%)**
Baseline(3×3)	0.73	30.7	24.5	44.8
1×3,3×1	0.68	42.2	22.6	51.6
(1×3,3×1)+A	0.61	44.7	21.0	54.4
1×5,1×3,3×1,5×1	0.64	43.7	21.9	52.4
(1×5,1×3,3×1,5×1)+A	**0.56**	**46.0**	**20.7**	**55.7**
1×7,1×5,1×3,3×1,5×1,7×1	0.66	43.1	23.1	50.7
(1×7,1×5,1×3,3×1,5×1,7×1)+A	0.58	44.8	20.8	53.3

In order to explore the AHConv in Fig. 3, we performed the experiments without the multi-scale feature extractor. That is the input features of log-mel spectrum and sound intensity vector were directly fed into AHConv. In table. 1, we have explored many combinations of asymmetric convolution. Only using 1×3 and 3×1 can't get enough features on frequency domain and time domain, while using 1×7, 1×5, 1×3, 3×1, 5×1 and 7×1 will degrade performance which may result in too many useless features being captured. The effect is best when the combinations of asymmetric convolution are 1×5, 1×3, 3×1 and 5×1. The results show that this combination of hybrid convolution can fully learn the features of different frequency domains and time domains simultaneously, which is very effective for SELD task. In addition, we also add adaptive attention block to the experiment. The experimental comparison of the same kind of hybrid convolution shows that the performance can be improved by adding the adaptive convolution.

Table.2 shows the results of ablation experiments of our proposed method. The first row denotes the scores of the baseline method. This method is the official baseline system of DCASE 2021 challenge task 3, and all of our experiments are based on it. The second row denotes the scores of the baseline method that adding the multi-scale feature extractor. After the multi-scale feature extractor, the AHConv is replaced by conventional CNN similar to the baseline method. Compared with the results of the first row, the scores in the second row decrease 0.16 and 4.5 on ER_{20° and LE_{CD}, and increase 18.7% and 12.0% on F_{20° and LR_{CD}, respectively. This proves the usefulness of multi-scale feature extractor. The last row denotes the scores of the method that adding multi-scale feature extractor and AHConv. This is the network that we proposed in Fig. 1. Compared with the results of the second row, the scores in the last row further decrease 0.04 and 1.2 on ER_{20° and LE_{CD}, and increase 5.7% and 4.8% on F_{20° and LR_{CD}, respectively. These results verified the effectiveness of the AHConv.

Table 2: The results of ablation experiments

Method	ER_{20°	$F_{20^\circ}(\%)$	LE_{CD}	$LR_{CD}(\%)$
baseline	0.73	30.7	24.5	44.8
$+Extractor$	0.57	49.4	20.0	56.8
$+Extractor + AHConv$	**0.53**	**55.1**	**18.8**	**61.6**

Table 3: The results of exploring the validity of depthwise separable convolution (DSConv)

Method	ER_{20°	$F_{20^\circ}(\%)$	LE_{CD}	$LR_{CD}(\%)$
Conv(1×1)	0.57	52.6	19.6	58.1
DSConv	0.53	55.1	18.8	61.6

In addition, we also explored the effectiveness comparison of conventional 2-D convolution and DSConv in the adaptive attention block. The method in the first row denotes adding a 2-D convolution with 1×1 kernel at the end of each path and then adding it. The second row is using depthwise separable convolution in our proposed method as seen in Fig. 4. Comparing two adaptive methods, the results showed that DSConv performed better.

5. CONCLUSIONS

In this paper, we propose a SELD method based on Adaptive Hybrid Convolution (AHConv) and multi-scale feature extractor. AHConv is designed to capture the time and frequency dependencies. Multi-scale feature extractor is designed to extract the multi-scale feature maps. We also propose an adaptive attention block embodied in AHConv and multi-scale feature extractor. Through a series of ablation experiments on the development dataset, we verify the effectiveness of AHConv and multi-scale feature extractor respectively. The results also show that our proposed method outperforms the baseline method on four evaluation metrics. Next we will introduce data augmentation methods to improve the performance of our proposed method.

6. ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (NSFC) (61761041,U1903213), Tianshan Innovation Team Plan Project of Xinjiang (202101642)

7. REFERENCES

[1] Y. Li, M. Liu, K. Drossos, and T. Virtanen, "Sound event detection via dilated convolutional recurrent neural networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 286–290.

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

[3] 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.

[4] H. Wang, Y. Zou, D. Chong, and W. Wang, "Environmental sound classification with parallel temporal-spectral attention," *Proceedings of INTERSPEECH 2020*, 2020.

[5] X. Zheng, Y. Song, J. Yan, L.-R. Dai, I. McLoughlin, and L. Liu, "An effective perturbation based semi-supervised learning method for sound event detection," *Proc. Interspeech 2020*, pp. 841–845, 2020.

[6] S. Chakrabarty and E. A. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 136–140.

[7] N. Ma, J. A. Gonzalez, and G. J. Brown, "Robust binaural localization of a target sound source by combining spectral source models and deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2122–2131, 2018.

[8] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1462–1466.

[9] T. N. T. Nguyen, W.-S. Gan, R. Ranjan, and D. L. Jones, "Robust source counting and doa estimation using spatial pseudo-spectrum and convolutional neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2626–2637, 2020.

[10] Z. Tang, J. D. Kanu, K. Hogan, and D. Manocha, "Regression and classification for direction-of-arrival estimation with convolutional recurrent neural networks," *Proc. Interspeech 2019*, pp. 654–658, 2019.

[11] H. Sundar, W. Wang, M. Sun, and C. Wang, "Raw waveform based end-to-end deep convolutional network for spatial localization of multiple acoustic sources," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4642–4646.

[12] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Crnn-based multiple doa estimation using acoustic intensity features for ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[13] W. He, P. Motlicek, and J.-M. Odobez, "Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 770–774.

[14] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2018.

[15] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Accdoa: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 915–919.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[17] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[18] X. Ding, Y. Guo, G. Ding, and J. Han, "Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1911–1920.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[21] W. Xia and K. Koishida, "Sound event detection in multichannel audio using convolutional time-frequency-channel squeeze and excitation," *Proc. Interspeech 2019*, pp. 3629–3633, 2019.

[22] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[23] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, "The nigens general sound events database," *arXiv preprint arXiv:1902.08314*, 2019.

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

[25] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 333–337.

[26] 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 Transactions on Audio, Speech, and Language Processing*, 2020.