# Sound Event Localization and Detection Based on Adaptive Hybrid Convolution and Multi-scale Feature Extractor

DCASE 2021 Workshop

Xinghao Sun, Ying Hu, Xiujuan Zhu, Liang He

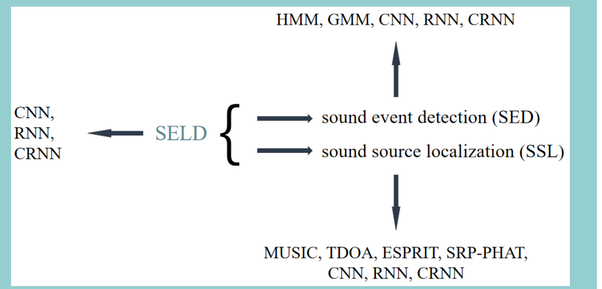
School of Information Science and Engineering, Xinjiang University, Urumqi, China

## Background

Sound Event Localization and Detection (SELD)

- identify the presence of independent or temporally-overlapped sound sources
- identify which sound class it belongs
- estimate their spatial directions while they are active.

## Present method

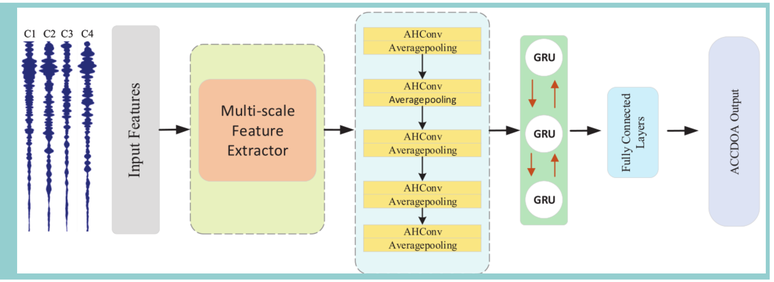



## Proposed method

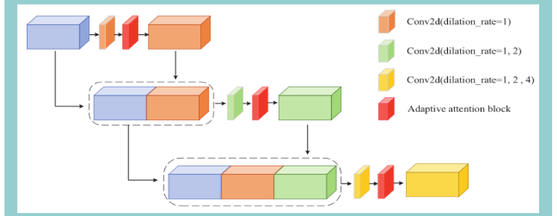
1.Multi-scale Feature Extractor

2.Adaptive Hybrid Convolution

3.Adaptive Attention block

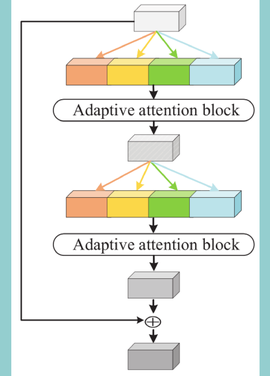



### 1.Multi-scale Feature Extractor

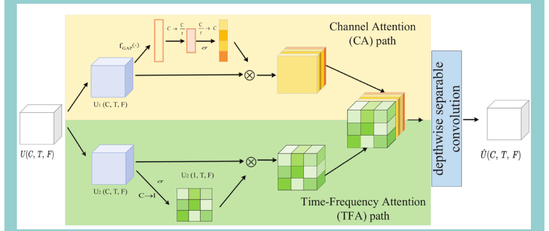



The maximum dilation rate in the last layer depends on the size of the input feature.

### 2.Adaptive Hybrid Convolution




### 3.Adaptive Attention block




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.

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

| The type of combination | $ER_{20^\circ}$ | $F_{20^\circ}(\%)$ | $LE_{CD}$ | $LR_{CD}(\%)$ |
|---|---|---|---|---|
| **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 |

## Experiment results

Table 2: The results of ablation experiments

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

## Conclusion

- 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.
- An adaptive attention block embodied in AHConv and multi-scale feature extractor.
- Next we will introduce data augmentation methods to improve the performance of our proposed method.