# SOUND EVENT LOCALIZATION AND DETECTION BASED ON OMI-DIMENSIONAL DYNAMIC CONVOLUTION AND FEATURE PYRAMID ATTENTION MODULE

Technical Report

*Mengzhen Ma*[1,2], *Ying Hu*[1,2], *Mingyu Wang*[1,2], *Wenjie Fang*[1,2], *Jie Liu*[1,2], *Zunxue Niu*[1,2], *Xin Fan*[1,2]

[1] School of Information Science and Engineering, Xinjiang University, Urumqi, China
107552103621@stu.xju.edu.cn
[2] Key Laboratory of Signal Detection and Processing in Xinjiang, China

## ABSTRACT

In this report, we present our method for Detection and Classification of Acoustic Scenes and Events (DCASE) 2023 challenge task3: Sound Event Localization and Detection Evaluated in Real Spatial Sound Scenes. We propose a method based on Omi-dimensional dynamic convolution (ODConv) and Feature Pyramid Attention Module (FPAM). In order to enhance the ability of extracting features for convolution kernel, we introduce an attention mechanism to it along four dimensions in ODConv. In addition, we explore FPAM to recalibrate high-level features from Residual Omi-dimensional Dynamic Convolution (Res_ODConv) blocks, making the model pay more attention to significant positions and channels. We also design Bidirectional Conformer to realize modeling context information in time and frequency dimensions. On Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS2023) dataset, our system demonstrates a prominent improvement over the baseline system. Only the first-order ambisonics (FOA) dataset was considered in this experiment.

***Index Terms***— DCASE2023, Sound source localization, Sound event detection, Omi-dimensional dynamic convolution, Feature pyramid attention module

## 1. INTRODUCTION

Sound event localization and detection task is aimed at detecting occurrences of sound events belonging to specific target classes, tracking their temporal activity, and estimating their direction-of-arrival or positions during it. Given multichannel audio input, a sound event localization and detection (SELD) system outputs a temporal activation track for each of the target sound classes, along with one or more corresponding spatial trajectories when the track indicates activity. This results in a spatio-temporal characterization of the acoustic scene that can be used in a wide range of machine cognition tasks.

Sound event localization and detection task can be divided into two parts: Sound Event Detection (SED) and Sound Source Localization (SSL). Many classical frameworks for SED and SSL are parametric approaches. Recently, many methods based on deep neural network (DNN) have been greatly applied in SELD task. They were shown to improve the robustness of SSL in challenging conditions compared to traditional methods. Among deep learning models, different architectures have been proposed: convolutional neural networks (CNNs)[1], convolutional recurrent neural networks (CRNNs)[2][3], U-net architectures[4], autoencoders (AEs)[5] or attention-based neural networks[6].

In order to improve the performance of overlapping sound events localization and detection, a track-wise output format was proposed[7], and the types of sound events output by each trajectory were different. To further simplify the output format, multi-ACCDOA was proposed[8], this method uses a single-ACCDOA vector that represents the activity of sound events and their positional information in each track. Multi-ACCOA is an extension of single-ACCDOA in the trajectory dimension.

The Attention mechanism has emerged as a promising alternative to model temporal dependencies. Attention efficiently learns the interdependencies of elements (e.g.vectors) between two sequences. SALADNET[9] proposes to replace the bidirectional long short-term memory (BiLSTM) layers of a state-of-the-art CRNN with one or several self-attention encoders. By avoiding the recurrent layers, the proposed model lends themselves to parallel computing, which is shown to produce considerable savings in execution time. The Conformer architecture was first proposed in ASR, which has shown the superiority of self-attention. In Resnet-Conformer network[10], the convolution layers are effective in extracting local fine-grained features, while the transformer models are good at capturing long-range global context. It also proves that SELD task by use of self-attention can also get outstanding performance.

In this paper, the attention mechanism also plays an important role in the network. Firstly, we introduce Omi-dimensional Dynamic Convolution[11], which applies attention mechanism like SE to convolution operation along spatial, input channel, and filter dimensions to enhance the ability of extracting features for convolution kernel. Secondly, we propose a Feature Pyramid Attention Module[12], a combination of Channel Attention and Spatial Attention, further recalibrating the features at different resolutions from previous layers. Thirdly, inspired from[13], we design a novel Bidirectional Conformer for modeling temporal context, considering the time and frequency dimensions in it. We conduct experiments on the development dataset to verify the effectiveness of our proposed method.

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

## 2. PROPOSED METHOD

We propose an architecture based on Omi-dimensional Dynamic Convolution and Feature Pyramid Attention Module, and our method achieves good performance in SELD task in real spatial sound scenes. The input to the method is multichannel audio, and the logmel spectrogram was extracted as input feature. We use a track-wise output format in the representation of multi-ACCDOA (a class- and track-wise output format). This network produces the temporal activity and DOA trajectory for each track. Meanwhile, we adopt ADPIT for the training process as the solution of the track permutation problem. Our proposed architecture is composed of a feature extraction module, a feature recalibration module, and Bidirectional Conformer. This is followed by two fully connected layers. The network diagram is presented in Fig.1.

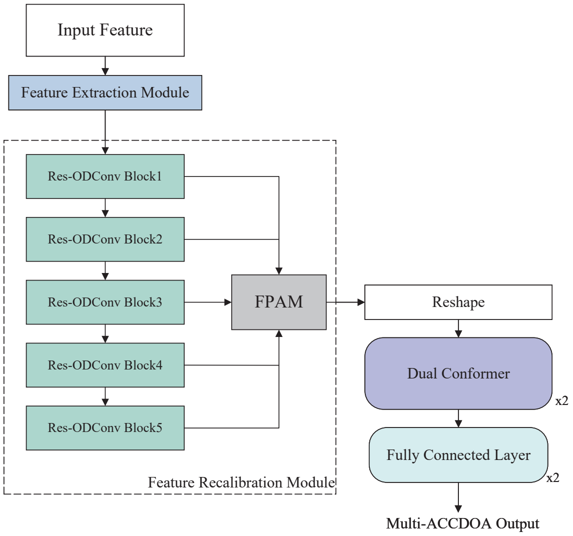

Figure 1: Diagram of our model.

### 2.1. Feature Extraction Module

In the feature extraction module, we preprocess the input features by two traditional convolution layers, each of which consists of vanilla convolution operation, batch normalization operation, and GELU activation function. After the feature is convolved, the average pooling is used to realize the preliminary downsampling. The feature extraction module makes input features sparse, increases the number of channels to achieve high-dimensional mappings of features, and facilitates the next step of more refined feature extraction.

### 2.2. Feature Recalibration Module

To obtain high-level features containing more context information, we design a feature recalibration module. The module includes two parts: Residual Omi-dimensional Dynamic Convolution (Res_ODConv) blocks and Feature Pyramid Attention Module (FPAM). Res_ODConv plays a role of extracting high-level features. FPAM refines features from five Res_ODConv blocks to aggregate features across different resolutions.

#### *2.2.1. Residual Omi-dimensional Dynamic Convolution Block*

A vanilla convolutional layer has only one static convolution kernel applied to all input samples. For dynamic convolution, it dynamically weights *n* kernels based on the attention mechanism. The weights of the convolution kernels are calculated linearly from the attention function conditioned on the input. However, previous researches on dynamic convolutions, such as DyConv[14], CondConv[15], only assigns an attention scale to the weight matrix of the convolution kernel, and the weight of each filter shares the same attention weight for all inputs. In a word, the previous dynamic convolutions ignore the spatial, the input channel , and the output channel dimensions, which leads to a coarse utilization of the kernel space when they design the attention mechanism that endows the *n* convolution kernels with dynamic properties.

Inspired by the above problems, Omi-dimensional Dynamic Convolution (ODConv)[11] introduces a multi-dimensional attention mechanism and uses a parallel strategy to learn different attention values of convolution kernels along the four dimensions of convolution kernel space. The ODConv can be expressed by the following formula:

$$y = \sum_{1}^{n} (\alpha_{wi} \odot \alpha_{fi} \odot \alpha_{ci} \odot \alpha_{si} \odot W_i) * x \tag{1}$$

The attention weights $\alpha_s, \alpha_c, \alpha_f$ are calculated along the spatial, the input channel, and the output channel dimensions of the kernel space for the convolution kernel *W* respectively. The attention weight $\alpha_\omega$ assigns different weights to *n* convolution kernels. In this paper, we only use one convolution kernel, so $\alpha_\omega$=1. This type of multi-dimensional attention mechanism greatly enhances the ability of feature extracting for convolution kernel.

When extracting high-level features, we use the basic block with two ODConv layers followed by average pooling. We also introduce a residual structure[16] after each layer of ODConv to retain the original information and reduce the possibility of gradient disappearance.

#### *2.2.2. Feature Pyramid Attention Module (FPAM)*

After high-level feature extraction, the features need to be refined. Inspired by[12], we feed the output of each Res_ODConv block into Feature Pyramid Attention Module (FPAM) at the same time. Because the output size of each block is different, we call this attention mechanism Feature Pyramid Attention. FPAM mainly consists of Spatial Attention Module (SAM) and Channel Attention Module (CAM). The overall structure of FPAM is shown in Fig.2. The outputs of five Res_ODConv blocks are firstly convolved to have the same number of channels and then sent to the SAM respectively. The SAM assigns different weight values to every spatial position of each feature. The weight matrixes are calculated using the following formula:

$$w = \sigma(Conv7\_7(Concat(MaxPool(x), StdPool(x)))) \tag{2}$$

SAM is composed of global max pooling, global standard pooling, and a convolution layer with a convolution kernel size of 7x7. The operations of pooling calculate the various statistical characteristics of the features, and the pooled feature are concatenated into

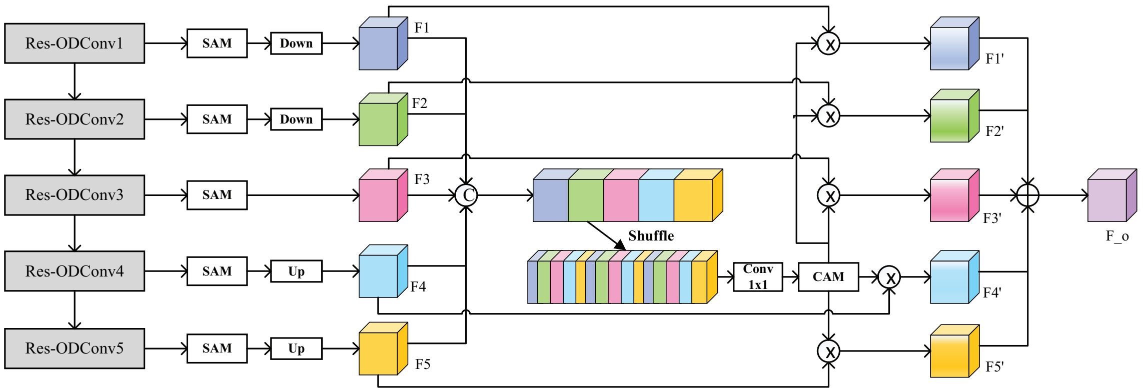

Figure 2: The overall structure of FPAM.

a convolution layer to halve the number of channels, which is convenient for multiplying them with the original features. The convolution output is activated by the Sigmoid function to obtain the attention values of each spatial position in the feature.

The features of different sizes after spatial refinement are upsampled and downsampled respectively to ensure that all features have the same size. Then, the features from five layers at diverse resolutions are concatenated for the shuffle operation to be mixed. In order to ensure the number of channels is consistent with the original features, we reduce the number of channels by convolution operation with the convolution kernel size of 1x1.

The features after spatially refined processing and accuracy mixing contain more information, and then the refined features are fed into the CAM for further recalibration. The internal structure of CAM is shown in Fig.3.

When calculating channel attention, similar to spatial attention, the global max pooling and global average pooling of features are carried out at first. We believe that the two pooling operations play individual role in the process of seeking attention. Therefore, two parameters that can be learned are designed to select the two pooled features adaptively and make a weighted summation. It is seen as an adaptive mechanism. The added features also pass through the Sigmoid function. Thus the channel attention weights are obtained and assigned to channels of the original features.

Through the above series of operations, features acquire more attention weight in some important spatial positions and channels. During training, the network will give more attention to these features. This is the main purpose of feature recalibration in FPAM.

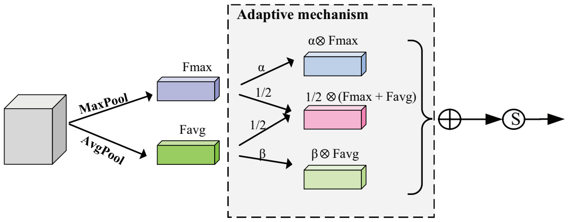

Figure 3: The details of CAM.

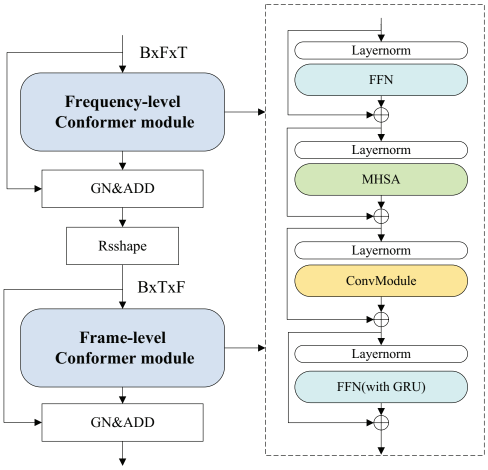

Figure 4: The global architecture and internal composition of Bidirectional Conformer.

## 2.3. Bidirectional Conformer

When we get recalibrated high-level features, we reshape them by multiplying frequency and channel dimension. In order to avoid the complexity of the model, we cut the number of channels in half before multiplying. Then, the Bidirectional Conformer module is sent for time context modeling to simulate the time structure of sound events. The specific structure is shown in Fig.4.

Bidirectional Conformer calculates attention along time and frequency dimensions by frequency-level Conformer and frame-level Conformer. Each Conformer includes two forward feedback layers, Multi-head Self-attention (MHSA) and ConvModule. Especially, the last forward feedback layer contains BiGRU and a linear layer. Similar to the Conformer structure, we also use the ”macaron” structure, with two attention modules sandwiched between two forward feedback layers. At the end of each module in a single Conformer, the layer normalization operation is carried out, and the residual structure is introduced to retain the original information. Each Conformer is finally followed by a group normalization layer, which is added with the original features as the output of the next layer. Bidirectional Conformer considers global and local information in two dimensions, respectively. MHSA is used to realize long-range sequence modeling, and ConvModule with kernel size of 31x31 achieves the function of obtaining local feature attention.

The features of time modeling through two Bidirectional Conformer layers are sent to the two fully connected layers, and the last fully connected layer outputs multi-ACCDOA vectors to represent the active state and corresponding position information of sound events, thus completing the classification and positioning task of sound events.

## 3. EXPERIMENT SETUP

### 3.1. Dataset

The Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) dataset contains multichannel recordings of sound scenes in various rooms and environments, together with temporal and spatial annotations of prominent events belonging to a set of target classes. Detailed information can be found in [17], we also use synthetic recordings generated through convolution of isolated sound samples with real spatial room impulse responses (SRIRs) in DCASE 2022 as external data to train our system.

### 3.2. Evaluation metrics

To evaluate SELD performance, we used the official evaluation metrics[18] that were introduced in the 2022 DCASE Challenge as our default metrics. The evaluation metrics of SELD can be divided into SED metrics and DOA metrics individually. For SED task, we use location-dependent error rate $ER_{20^\circ}$ and F1-score $F_{20^\circ}$. Contrary to the previous challenges, in this challenge we perform micro averaging of the location-dependent F1-score, an ideal method will have an F1-score of one and ER of zero. A DOA method is evaluated using a class-dependent localization error $LE_{CD}$, which is computed as the mean angular error of the matched true positives per class. In addition, we compute a localization recall $LR_{CD}$ metric per class to describe the performance of DOA. Similar to SED metrics, a good method will have an LR of one and LE of zero in the ideal case.

### 3.3. Training procedure

The sampling frequency was used at 24 kHz in our method. STFT was applied with configurations of 20 ms frame length and 10 ms frame hop. We use a batch size of 64.When extracting logmel spectrogram, we set the number of frequency bins is 128. With the aim of ensuring a fair comparison, all models were trained for 300 epochs with the Adam optimizer of the same initialized parameters by early stopping strategy.

### 3.4. Our challenge submissions

## 4. RESULT AND DISCUSSION

Our proposed model results outperform the DCASE 2023 baseline model, **Ma_XJU_task3a_1** achieve the improvement of 3.7, 10.5% in $LE_{CD}$ and $LR_{CD}$. The result of the baseline is acquired by ourselves. Our methods have outstanding improvement in $LR_{CD}$, but the metric of $F_{20^\circ}$ is lower than the baseline. There is still potential for our model to realize greater improvement.

Table 1: The performance comparison for different methods on the development dataset.

| method | $ER_{20^\circ}$ | $F_{20^\circ}\%$ | $LE_{CD}$ | $LR_{CD}\%$ |
|---|---|---|---|---|
| DCASE2023baseline | 0.69 | 42.7 | 29 | 52.8 |
| Ma_XJU_task3a_1 | 0.69 | 36.4 | 25.3 | 63.3 |

## 5. CONCLUSIONS

In this paper, we propose a SELD method based on Omi-directional Dynamic Convolution (ODConv) and Feature Pyramid Attention Module (FPAM). The ODConv was designed to enhance the ability of extracting features for convolution kernel and kernel weights can be dependent on input dynamically. The FAPM in the high-level feature extraction stage was designed to pay more attention to significant positions and channels. The results on the development dataset show that our proposed method outperforms the baseline method. In the future, we will explore more innovative methods to improve the performance of the model.

## 6. REFERENCES

[1] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, “Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates,” *Sensors*, vol. 18, no. 10, p. 3418, 2018.

[2] 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.

[3] 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.

[4] S. E. Chazan, H. Hammer, G. Hazan, J. Goldberger, and S. Gannot, “Multi-microphone speaker separation based on deep doa estimation,” in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[5] Z.-M. Liu, C. Zhang, and S. Y. Philip, “Direction-of-arrival estimation based on deep neural networks with robustness to array imperfections,” *IEEE Transactions on Antennas and Propagation*, vol. 66, no. 12, pp. 7315–7327, 2018.

[6] C. Schymura, T. Ochiai, M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, and D. Kolossa, “Exploiting attention-based sequence-to-sequence architectures for sound event localization,” in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 231–235.

[7] Y. Cao, T. Iqbal, Q. Kong, Y. Zhong, W. Wang, and M. D. Plumbley, “Event-independent network for polyphonic sound event localization and detection,” *arXiv preprint arXiv:2010.00140*, 2020.

[8] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji, "Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 316–320.

[9] P.-A. Grumiaux, S. Kitić, P. Srivastava, L. Girin, and A. Guérin, "Saladnet: Self-attentive multisource localization in the ambisonics domain," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 336–340.

[10] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1251–1264, 2023.

[11] C. Li, A. Zhou, and A. Yao, "Omni-dimensional dynamic convolution," *arXiv preprint arXiv:2209.07947*, 2022.

[12] L. Zhou, Y. Zhou, X. Qi, J. Hu, T. L. Lam, and Y. Xu, "Feature pyramid attention based residual neural network for environmental sound classification," *arXiv preprint arXiv:2205.14411*, 2022.

[13] L. Wang, W. Wei, Y. Chen, and Y. Hu, "D 2 net: A denoising and dereverberation network based on two-branch encoder and dual-path transformer," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 1649–1654.

[14] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 030–11 039.

[15] B. Yang, G. Bender, Q. V. Le, and J. Ngiam, "Condconv: Conditionally parameterized convolutions for efficient inference," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] 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," *arXiv preprint arXiv:2206.01948*, 2022.

[18] 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*, vol. 29, pp. 684–698, 2020.