A MULTI-LEVEL FEATURE EXTRACTION NETWORK FOR SOUND EVENT LOCALIZATION AND DETECTION IN DCASE 2025 TASK 3

Technical Report

QingJing Wan[1,2]*,Ying Hu*[1,2]*,Jie Liu*[1,2]*,Qiong Wu*[1,2]*,Qin Yang*[1,2]*,*
WenTao Zhou[1,2]*,Tianqing Zhou*[1,2] *,Nannan Teng*[1,2]*,Fangxu Chen*[1,2]*,Zijun Chen*[1,2]

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

ABSTRACT

This technical report describes our submission system for Task 3 of the DCASE 2025 Challenge: stereo sound event localization and detection (SELD) in regular video content. We participate in the audio-only track. Our system adopts a Multi-Level Feature Extraction Network, which consists of three main components. First, a Feature Extraction Enhancement module(FEEM) is used to extract fine-grained and meaningful features at multiple hierarchical levels, improving the model's ability to handle both sub-tasks: Direction of Arrival (DOA) estimation and Sound Event Detection (SED). Second, a Feature Fusion module(FFM) is employed to integrate multi-level features, further enhancing the representational capacity of the network. Finally, several data augmentation strategies are applied to improve the robustness of the network. Experimental results on the DCASE 2025 Task 3 stereo SELD dataset demonstrate the effectiveness of the proposed system.

***Index Terms*—** Stereo sound event localization and detection,Multi-Level Feature Extraction Network,Feature Extraction Enhancement, Feature Fusion, data augmentation

1. INTRODUCTION

The objective of the Sound Event Localization and Detection (SELD) task is to detect occurrences of sound events from specific target classes, track their temporal activity, and estimate their directions-of-arrival (DOA) or positions. Given multichannel audio input,a 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,including smart homes and audio surveillance[1, 2].

The Sound Event Localization and Detection (SELD) task was first introduced in DCASE2019 Task3[3], focusing on scenarios with a single, fixed-position sound source. It utilized multichannel audio synthesized by convolving mono audio files with impulse responses. Subsequent DCASE Challenges [4, 5, 6, 7, 8] gradually introduced more complex environmental settings, including moving sound sources, diverse impulse responses, overlapping sound events of the same class, lower signal-to-noise ratios (SNRs), real spatial acoustic scenes, and the estimation of the distance to the detected events.This year, the challenge tackles SELD using stereo audio data, converted from the previously used four-channel audio data[9],This change is intended to better reflect common audio and media scenarios, but it also increases the task's complexity due to the reduction in available spatial cues.

In this report, we present a system developed for the stereo SELD task in DCASE 2025 Task 3. To address the reduced spatial information in stereo audio, we propose a Multi-Level Feature Extraction Network that captures and integrates acoustic features across different hierarchical levels. This design aims to improve both Direction-of-Arrival estimation and Sound Event Detection. We further enhance the system's generalization through a set of data augmentation strategies. Experimental results on the official DCASE 2025 Stereo SELD dataset demonstrate the effectiveness of our approach in challenging acoustic environments.

2. METHOD

We propose a Multi-Level Feature Extraction Network for stereo sound event localization and detection. The input to the system is stereo audio, from which log-mel spectrograms are extracted as input features. We adopt the multi-ACCDOA representation[10].and use a track-wise output format to simultaneously predict temporal activity and direction-of-arrival (DOA) trajectories for each track. To address the track permutation problem during training, we employ the Auxiliary Duplicating Permutation-Invariant Training (ADPIT) strategy.

Our proposed architecture consists of an initial feature extraction module, a Feature Extraction Enhancement Module (FEEM) for multi-level fine-grained representation learning, a Feature Fusion Module (FFM) for cross-level feature integration, and a Conformer block for modeling long-range temporal context. Each FEEM block is followed by a max-pooling layer, which downsamples the feature map. Prior to fusion in the FFM, a 1×1 convolution is applied to unify the channel dimensions of multi-level features. The output of the FFM is then passed to the Conformer block, followed by two fully connected layers to produce the final SELD predictions.The overall network architecture is shown in Figure 1.

2.1. Feature Extraction Module

In the initial feature extraction module, the input features are first processed by a conventional convolutional layer consisting of a standard convolution operation,followed by batch

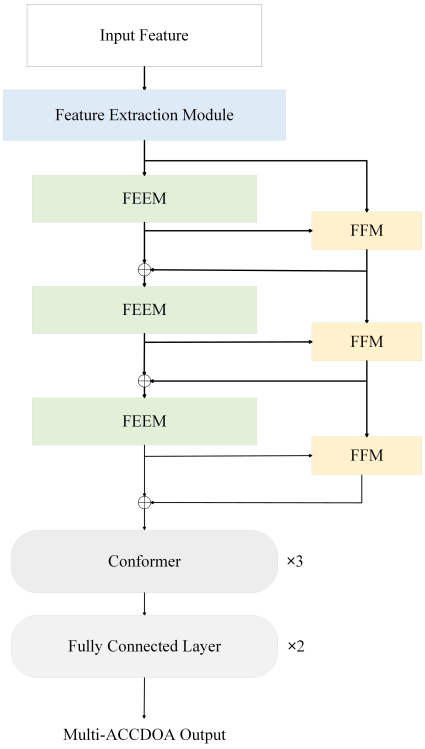

Figure 1: Overall architecture of the proposed network.

normalization[11]and a GELU activation function [12]. Max pooling is then applied for initial downsampling. This module increases the number of channels to obtain higher-dimensional feature representations, facilitating subsequent, more refined feature extraction.

2.2. Feature Extraction Enhancement Module(FEEM)

To obtain more fine-grained and meaningful features, we introduce a Feature Extraction Enhancement module (FEEM). This module consists of two components: a Multi-Branch Feature Extraction submodule and a Feature Enhancement submodule,as shown in Fig. 2.

The Multi-Branch submodule extracts detailed and diverse features through a parallel architecture consisting of three branches: a local branch that captures fine-grained spatial details using small non-overlapping patches, a global branch that aggregates broader contextual information using larger patches, and a serial convolution branch that replaces large convolution kernels with three consecutive 3×3 convolutions for efficient local structure modeling. The distinction between the local and global branches is controlled by the patch size parameter p where $p = 2$ corresponds to local features and $p = 4$ to global features.The input feature map is first divided into non-overlapping $p \times p$ patches using Unfold and reshape operations, then averaged along the channel dimension to obtain spatial tokens. These tokens are passed through a feed-forward network (FFN), followed by an activation function to generate spatial weights. The resulting weighted features are refined through a feature selection mechanism[13]that selects task-relevant tokens and channels to improve representation quality. The outputs of the three branches are summed along the channel dimension to form a unified representation, which is then processed by the Feature Enhancement submodule.This submodule applies attention mechanisms,including channel attention and spatial attention[14] to highlight important feature dimensions and regions. Finally, the enhanced features are passed through dropout, ReLU activation, and batch normalization to produce the final refined output.

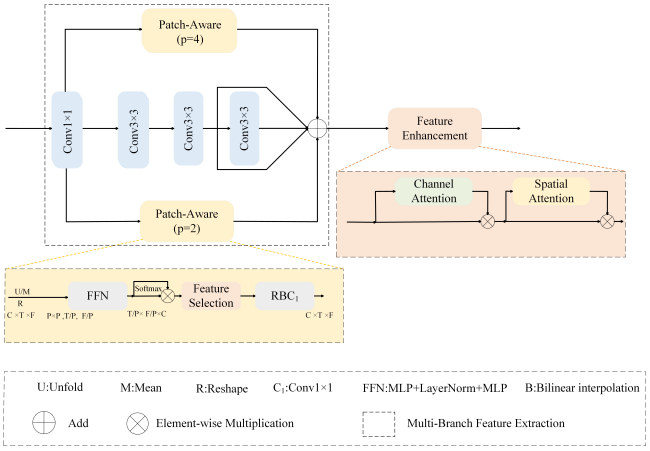

Figure 2: Structure of the Feature Extraction Enhancement Module (FEEM).

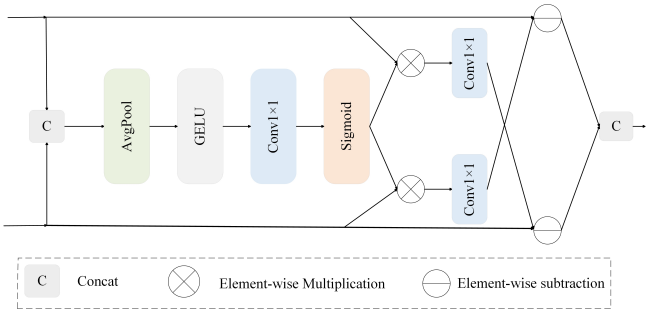

Figure 3: Structure of the Feature Fusion Module (FFM).

2.3. Feature Fusion module(FFM)

To integrate detailed information across different levels and construct a more robust feature representation, we employ a Feature Fusion Module (FFM) that effectively combines multi-level features to enhance overall representational capacity. This module incorporates a global attention mechanism to adaptively reweight and reorganize features from various levels, emphasizing key features while suppressing irrelevant information, as illustrated in Fig. 3.

2.4. Data Augmentation

To increase the diversity of our dataset, we applied frequency shifting [15] and synthesized 40 hours of multi-channel audio using the Spatial Scaper library[16, 17], which was then segmented into 5-second stereo audio clips.

Table 1: Performance comparison on the development set.

Methods	F1 score ↑	DOAE ↓	RDE ↓
Baseline 2025	22.8%	24.5°	41%
Ours' system	**37.1%**	**18.3°**	**30%**

3. EXPERIMENTS

In this section, we show our results on the development dataset.

3.1. Experimental settings

In our experiments, we used stereo audio. The sampling rate was set to 24 kHz, the STFT frame length was 40 ms, and the hop length was 20 ms. We used 128 mel filters. The input length was 5 seconds, or 250 frames. The model was trained with the Adam optimizer for 200 epochs, with a learning rate set to 0.001.

We evaluated our SELD system using official metrics, including the location-dependent F1 score, direction-of-arrival error (DOAE), and relative distance error (RDE).

3.2. Experimental result

Table 1 presents the performance of our proposed methods on the development set. As shown, our approach significantly outperforms the baseline in terms of the location-dependent F1 score , as well as the DOA error (DOAE) and relative distance error (RDE) metrics.

4. CONCLUSION

We presented a stereo sound event localization and detection (SELD) system developed for Task 3 of the DCASE 2025 Challenge. Our system leverages a Multi-Level Feature Extraction Network that integrates a Feature Extraction Enhancement Module (FEEM) and a Feature Fusion Module (FFM) to extract and combine fine-grained, hierarchical features for improved performance in both sound event detection and direction-of-arrival (DOA) estimation. In addition,we apply data augmentation methods to enhance the model's robustness . Experimental results demonstrate that our system significantly outperforms the baseline across multiple metrics.

5. REFERENCES

[1] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 279–288, 2015.

[2] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, 2007, pp. 21–26.

[3] http://dcase.community/challenge2019/, 2019.

[4] http://dcase.community/challenge2020/, 2020.

[5] http://dcase.community/challenge2021/, 2021.

[6] http://dcase.community/challenge2022/, 2022.

[7] http://dcase.community/challenge2023/, 2023.

[8] http://dcase.community/challenge2024/, 2024.

[9] K. Shimada, A. Politis, P. Sudarsanam, and et al., "Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *Advances in Neural Information Processing Systems*, vol. 36, pp. 72 931–72 957, 2023.

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

[11] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*. PMLR, 2015, pp. 448–456.

[12] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[13] B. Shi, S. Gai, T. Darrell, and X. Wang, "Refocusing is key to transfer learning," *arXiv preprint arXiv:2305.15542*, 2023.

[14] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[15] T. N. T. Nguyen, K. N. Watcharasupat, N. K. Nguyen, D. L. Jones, and W.-S. Gan, "Salsa: Spatial cue-augmented logspectrogram features for polyphonic sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1749–1762, 2022.

[16] I. R. Roman, C. Ick, S. Ding, A. S. Roman, B. McFee, and J. P. Bello, "Spatial scaper: A library to simulate and augment soundscapes for sound event localization and detection in realistic rooms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1221–1225.

[17] H. Hong, Q. Wang, R. Wei, and et al., "Mvanet: Multi-stage video attention network for sound event localization and detection with source distance estimation," in *ICASSP 2025 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.