A STEREO SOUND EVENT LOCALIZATION AND DETECTION METHOD BASED ON FEATURE FUSION AND TWO-STAGE TRAINING

Technical Report

Digao Wu[1], *Ming Zhu*[1]

[1] Huazhong University of Science and Technology
School of Electronic Information and Communications, Wuhan, China
{wudigao_, zhuming}@hust.edu.cn

ABSTRACT

This technical report presents our system for Task 3 of the DCASE 2025 Challenge: *Stereo Sound Event Localization and Detection in Regular Video Content*. The task requires predicting the activity, azimuth, and distance of sound events using stereo audio. We participate in the audio-only track. We propose a stereo SELD model based on the ResNet-Conformer structure, integrating channel-wise attention and feature fusion, with outputs represented in the ACCDOA format. To enhance model performance, we augment the training data with additional stereo audio segments sampled from the official DCASE 2024 synthetic dataset. We apply several data augmentation techniques and adopt a two-stage training strategy to improve generalization and performance on real data. A dynamic thresholding method is also introduced during inference to further boost the prediction accuracy. The experimental results on the official development dataset show that our proposed system outperforms the baseline in all evaluation metrics.

Index Terms— Sound event localization and detection, Source distance estimation, Stereo audio, Feature fusion, Training strategy

1. INTRODUCTION

Sound Event Localization and Detection with stereo audio data (referred to as **stereo SELD**) aims to perform Sound Event Detection (SED), Direction of Arrival Estimation (DOAE), and Source Distance Estimation (SDE) simultaneously using two-channel stereo audio. The SELD task was first introduced in the DCASE 2019 Challenge [1, 2], with the goal of jointly estimating the activity status and the direction of incoming sound events. In subsequent challenges [3], distance estimation was incorporated to enable a more comprehensive modeling of sound events, accompanied by the release of both real and synthetic datasets. SELD systems based on First-Order Ambisonics (FOA) or microphone array (MIC) audio formats have shown strong potential in practical applications such as machine listening, smart home systems, and wildlife monitoring.

In [4], a CRNN-based SELD model with two parallel output branches is proposed: one for SED and the other for DOAE, where the SED output serves as a soft mask to inform the DOA predictions. In [5], the Event-Independent Network V2 (EINV2) was introduced, which employs soft parameter sharing and multi-head self-attention (MHSA) to decode SELD outputs.

In [6], the Activity-Coupled Cartesian Direction of Arrival (ACCDOA) representation was proposed, where the sound event activity is assigned to the magnitude of the corresponding Cartesian DOA vector. As a result, the SED and DOA tasks are combined into a single regression task in Cartesian coordinates.However, the ACCDOA representation cannot handle the case of simultaneous occurrence of similar events. To address this issue, the ACCDOA representation was extended to Multi-ACCDOA by introducing Auxiliary Duplicating Permutation Invariant Training (ADPIT) [7].

To incorporate distance estimation, [8] proposed extended ACCDOA and extended Multi-ACCDOA formats. In this framework, sound event detection and localization are handled by the ACCDOA branch, while sound source distance estimation (SDE) is performed by a category-wise distance branch. Another solution is a unified approach, where the distance and direction information are embedded in a single output vector by extending the Multi-ACCDOA representation. In [9], three output structures were proposed for SELD models, along with three training strategies: (1) independently training two models for SED-DOA and SED-SDE estimation; (2) merging DOA and distance estimation into a unified Source Coordinate Estimation (SCE) task and training an SED-SCE model; and (3) directly using a three-branch model (SED-DOA-SDE) for joint training.

In previous DCASE SELD challenges, spatial audio was usually provided in four-channel formats such as FOA and MIC. This year, the challenge introduced a new input format—Mid/Side (M/S) stereo audio—where only two channels are used to perform SED, DOAE, and SDE.

In this technical report, we propose **Stereo-RCnet**, a stereo SELD framework that leverages channel-wise self-attention and feature fusion to effectively capture inter-channel spatial cues. The model adopts a ResNet-Conformer backbone to achieve SED, DOAE, and SDE simultaneously using two-channel stereo input. To further enhance performance, a post-processing strategy is applied to refine the model outputs.

To improve robustness and generalization, we apply several data augmentation techniques and incorporate external data, as permitted by the challenge rules. Specifically, we augment the training set with 90,000 five-second stereo audio clips sampled from the official DCASE 2024 synthetic dataset, adding approximately 125 hours of audio. This allows the model to learn from a wider range of spatial and temporal variations.

To mitigate the domain gap between real and synthetic data, we adopt a two-stage training strategy: the model is first trained on the combined dataset and then fine-tuned using real recordings only. Experimental results on the official development set demonstrate that our proposed method yields substantial improvements over the official audio-only baseline across all evaluation metrics.

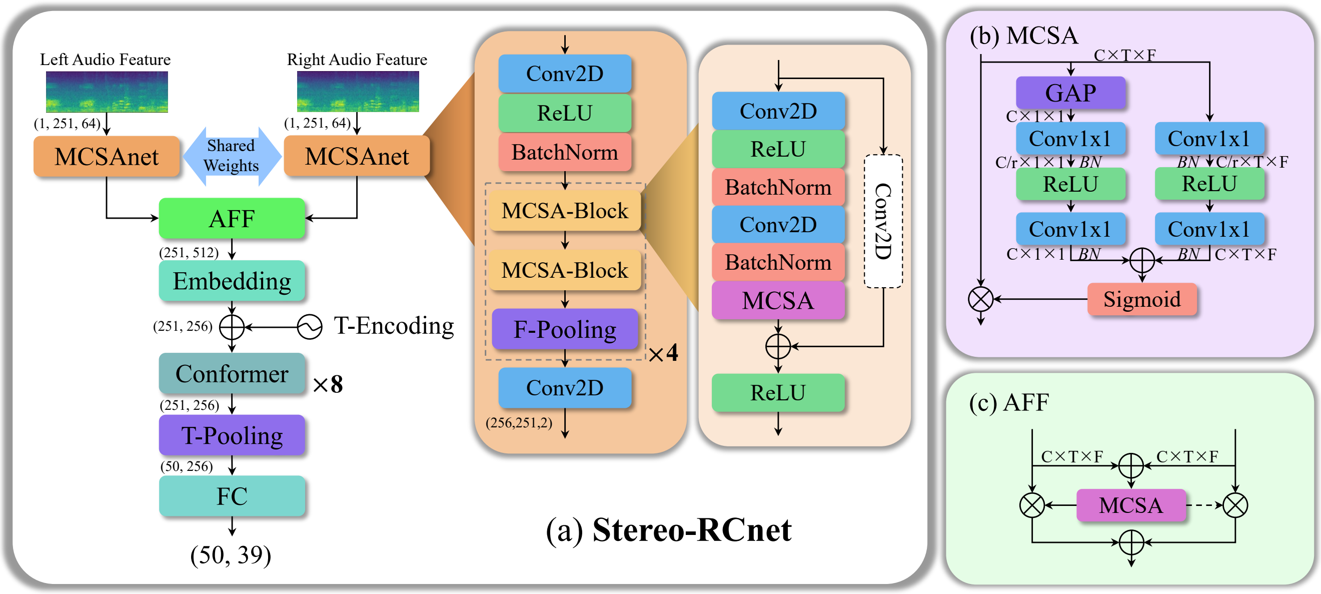

Figure 1: Architecture of the proposed Stereo-RCnet. (a) The overall framework, including stereo-channel feature extraction, fusion, and context modeling. (b) Structure of the Multi-Channel Self-Attention (MCSA). (c) Structure of the Attention Feature Fusion (AFF).

2. PROPOSED METHOD

2.1. Input Features

We first apply the Short-Time Fourier Transform (STFT) to 24 kHz stereo audio signals using a 40-ms Hanning window, a 20-ms hop size, and a 1024-point DFT. From the resulting spectrograms, we extract 64-dimensional log-Mel features for each channel. As a result, each 5-second stereo audio clip is represented as a feature tensor of size $2 \times 251 \times 64$, which serves as the input to our model.

2.2. Network Architecture

The proposed **Stereo-RCnet** aims to enhance sound event localization and detection by fully leveraging the spatial cues embedded in stereo audio signals. The overall architecture is illustrated in Fig. 1(a).

Specifically, log-Mel features extracted from the stereo audio are fed into a shared-weight backbone, MCSAnet, for local feature extraction. MCSAnet is built upon a ResNet backbone and incorporates Multi-Channel Self-Attention (MCSA) module [10], as illustrated in Fig. 1(b), to enhance the modeling of inter-channel dependencies and time-frequency representations. As suggested by [11], only frequency pooling is applied within MCSAnet to preserve fine-grained temporal resolution.

The extracted features from both channels are then fused using an Attentional Feature Fusion (AFF) module [10], as illustrated in Fig. 1(c). This module integrates an internal MCSA module that dynamically adjusts fusion weights based on the relative importance of the left and right channel features, thereby facilitating more effective inter-channel feature integration. The fused feature is computed as:

$$\mathbf{F}_{\text{fused}} = w \cdot \mathbf{F}_{\text{left}} + (2 - w) \cdot \mathbf{F}_{\text{right}}, \tag{1}$$

where w is the weight generated by the MCSA module.

The fused feature is first passed through an embedding layer to obtain a hidden representation for each time frame. To better model the temporal dynamics of sound events—where temporal correlation often decays with increasing time lag—we incorporate temporal positional encoding (T-Encoding) after the embedding layer, which is implemented using the fixed (non-learnable) sinusoidal encoding. This process is formulated as:

$$\mathbf{H} = \text{Embed}(\mathbf{F}_{\text{fused}}) + \mathbf{P}, \tag{2}$$

where $\mathbf{P}$ denotes the T-Encoding.

This enriched representation is then fed into a stack of Conformer layers [12], which combine convolutional and self-attention mechanisms to effectively model both local and global temporal dependencies. This architecture enables the network to capture long-range contextual information while retaining precise temporal alignment, which is crucial for accurate event localization and detection.

Finally, temporal pooling (T-Pooling) is applied to the hidden representations $\mathbf{H}$, followed by a fully connected (FC) layer to produce the final prediction in the ACCDOA format, enabling joint sound event detection and localization.

$$\hat{\mathbf{y}}_{\text{ACCDOA}} = \text{FC}(\text{T-Pooling}(\mathbf{H})), \tag{3}$$

It is worth noting that although the ACCDOA output format cannot effectively handle overlapping sound events from the same class, we adopt it in this work due to its simpler output structure and training stability. Therefore, we do not adopt more complex output structures, such as the multi-track format Multi-ACCDOA or multi-branch output architectures.

2.3. Loss Function Design

To better train the model, we slightly modify the loss functions used in the official baseline. Specifically, we adopt the Mean Squared

Error (MSE) and the Mean Squared Percentage Error (MSPE) as our loss functions for DOA and distance estimation, respectively. The formulations are defined as follows:

$$\mathcal{L}_{\mathrm{DOA}} = \frac{1}{CT}\sum_{c,t}\left\|\mathbf{R}_{ct} - \hat{\mathbf{R}}_{ct}\right\|^2, \tag{4}$$

$$\mathcal{L}_{\mathrm{dist}} = \frac{1}{CT}\sum_{c,t}\left\|a_{ct}\cdot\frac{d_{ct} - \hat{d}_{ct}}{d_{ct}}\right\|^2, \tag{5}$$

where, $\mathbf{R}_{ct}$ and $\hat{\mathbf{R}}_{ct}$ denote the ground-truth and predicted DOA vectors for class c at time frame t, respectively. a_{ct} is the ground-truth activity indicator (1 for active, 0 for inactive), d_{ct} and $\hat{d}_{ct}$ represent the ground-truth and predicted distances. C is the number of sound event classes, and T is the number of time frames. $\|\cdot\|^2$ denotes the squared Euclidean norm. The predicted activity indicator $\hat{a}_{ct}$ is inferred from the magnitude of the predicted DOA vector $\hat{\mathbf{R}}_{ct}$. All classes use a fixed activity threshold of 0.5 during training.

Accordingly, the overall loss function of the proposed stereo SELD model is formulated as follows:

$$\mathcal{L}_{\mathrm{stereo\ SELD}} = \alpha\mathcal{L}_{\mathrm{DOA}} + \beta\mathcal{L}_{\mathrm{dist}}, \tag{6}$$

where α and β are hyperparameters that control the relative weights of the DOA and distance loss terms. In our implementation, we set $\alpha = 1$ and $\beta = 2$.

2.4. Data Augmentation

At the early stage of the challenge, the organizers released a development dataset consisting of 30,000 stereo audio segments from real recordings, among which 16,214 segments (approximately 22 hours) were designated for training and 13,786 segments (approximately 19 hours) for validation. However, this real dataset suffers from an evident class imbalance, which may hinder model generalization.

Due to the class imbalance in the real audio dataset, we sampled additional synthetic stereo audio segments to improve model generalization ability and performance. First, we obtained the 20-hour official synthetic dataset from DCASE 2024 Task 3 [3]. The synthetic audio is in FOA format and was generated using the Spatial Scaper library [13], with sound samples drawn from the FSD50K dataset [14] and spatial room impulse responses (SRIR) [15]. Following the official stereo synthesis procedure, we sampled approximately 90,000 five-second stereo audio clips from the 20-hour FOA format data, resulting in about 125 hours of training data, which was used to augment the development set.

More specifically, for FOA signals following an ACN/SN3D convention ordered as $[W(n), Y(n), Z(n), X(n)]$, the corresponding stereo signals $[L(n), R(n)]$ are derived using the following linear transformation:

$$L(n) = W(n) + Y(n), \quad R(n) = W(n) - Y(n)$$

To enhance model robustness, we also applied several data augmentation techniques during training, including random cutout [16], time-frequency masking [17], frequency shifting [18], and AugMix [19].

2.5. Two-stage training strategy

To mitigate the distributional discrepancy between synthetic and real audio data, we employ a two-stage training strategy to improve the model's robustness in real scenarios. In the first stage, the model is trained on a mixture of synthetic and real recordings for 20 epochs with an initial learning rate of 1×10^{-4}. In the second stage, fine-tuning is performed exclusively on the real dataset with a reduced learning rate of 2×10^{-5}, until validation performance plateaus for 20 consecutive epochs.

Compared to training solely on real data, this two-stage training strategy allows the model to first learn generalizable acoustic patterns from a larger and more diverse synthetic dataset, and then adapt to the nuances of real-world environments through fine-tuning. This strategy improves generalization without sacrificing performance in realistic acoustic conditions.

2.6. Post-processing

During inference, We adopt a post-processing strategy, referred to as dynamic threshold (DT), to further enhance the performance of the trained model. Instead of using a fixed default threshold (typically set to 0.5), we apply class-specific decision thresholds for different sound event categories in the SED task.

3. RESULTS ON DEVELOPMENT DATASET

We evaluate our proposed stereo SELD model on the official development dataset, and the experimental results are presented in Table 1. "Baseline-A" refers to the official audio-only baseline system [20, 21]. "Stage 1" indicates the model performance after training on the combined synthetic and real dataset. "Stage 2" shows the results after further fine-tuning on the real dataset. "Stage 2 + PP" represents the final performance obtained by applying the DT post-processing method on top of Training Stage 2.

Table 1: Experimental results of the audio-only stereo SELD systems on the development dataset.

System	**F_{20° ↑**	**$\mathrm{DOAE}_{\mathrm{CD}}$ ↓**	**$\mathrm{RDE}_{\mathrm{CD}}$ ↓**
Baseline-A	22.78%	24.5°	0.41
Stage 1	34.56%	16.5°	0.36
Stage 2	41.32%	**14.9°**	**0.30**
Stage 2 + PP	**42.5%**	15.0°	**0.30**

As shown in Table 1, our method achieves substantial improvements over the baseline across all three evaluation metrics. Stage 1 training, which leverages both real and synthetic data, raises the F_{20° from 22.78% to 34.56% and reduces the $\mathrm{DOAE}_{\mathrm{CD}}$ from 24.5° to 16.5°, demonstrating the effectiveness of incorporating synthetic data for representation learning.

Stage 2 further improves performance through fine-tuning on real data, increasing the F_{20° to 41.32%, reducing the $\mathrm{DOAE}_{\mathrm{CD}}$ to 14.9°, and lowering the $\mathrm{RDE}_{\mathrm{CD}}$ to 0.30. These results confirm the advantage of our two-stage training strategy in adapting the model to real spatial distributions.

Finally, applying DT post-processing in Stage 2 + PP slightly improves the F_{20° to 42.5%, while maintaining the same low localization and distance errors. This indicates that the DT post-processing contributes to improving the model's accuracy in SED.

4. CONCLUSION

In this paper, we propose Stereo-RCnet, a stereo sound event localization and detection method based on feature fusion and a two-stage training strategy. The model leverages stereo channel information and employs a Multi-Channel Self-Attention (MSCA) module together with an Attentional Feature Fusion (AFF) module to effectively model inter-channel spatial differences. For temporal modeling, a stack of Conformer layers with temporal positional encoding is used to capture long-range contextual dependencies. Furthermore, we adopt a combined training scheme using both synthetic and real data, along with a dynamic threshold (DT) post-processing method to enhance model generalization and detection accuracy in real-world environments. Experimental results show clear improvements over the baseline, validating the effectiveness of our approach.

5. REFERENCES

[1] http://dcase.community/challenge2019/.

[2] 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, 2021.

[3] http://dcase.community/challenge2024/.

[4] 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, March 2019.

[5] 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 *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, 2021, pp. 885–889.

[6] 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)*, Toronto, ON, Canada, 2021, pp. 915–919.

[7] 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)*, Singapore, Singapore, 2022, pp. 316–320.

[8] D. A. Krause, A. Politis, and A. Mesaros, "Sound event detection and localization with distance estimation," in *2024 32nd European Signal Processing Conference (EUSIPCO)*, Lyon, France, 2024, pp. 286–290.

[9] Y. Dong, Q. Wang, H. Hong, Y. Jiang, and S. Cheng, "An experimental study on joint modeling for sound event localization and detection with source distance estimation," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, 2025, pp. 1–5.

[10] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 2021, pp. 3559–3568.

[11] S. Niu, J. Du, Q. Wang, L. Chai, H. Wu, Z. Nian, L. Sun, Y. Fang, J. Pan, and C. Lee, "An experimental study on sound event localization and detection under realistic testing conditions," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*. IEEE, 2023, pp. 1–5.

[12] 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 *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 5036–5040.

[13] 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 *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1221–1225.

[14] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[15] 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 Workshop*, 2020, pp. 165–169.

[16] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 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 logspectrogram features for polyphonic sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1749–1762, 2022.

[19] 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," *arXiv preprint*, 2019.

[20] D. Diaz-Guerra, A. Politis, P. Sudarsanam, and et al., "Baseline models and evaluation of sound event localization and detection with distance estimation in dcase2024 challenge," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2024 Workshop (DCASE2024)*, Tokyo, Japan, October 2024, pp. 41–45.

[21] http://dcase.community/challenge2025/.