THE NERC-SLIP SYSTEM FOR STEREO SOUND EVENT LOCALIZATION AND DETECTION IN REGULAR VIDEO CONTENT OF DCASE 2025 CHALLENGE

Technical Report

Qing Wang[1]*, Hengyi Hong*[1]*, Ruoyu Wei*[3]*, Lin Li*[2]*, Yuxuan Dong*[1]
Mingqi Cai[2]*, Xin Fang*[2]*, Jiangzhao Wu*[3]*, Jun Du*[1]

[1] University of Science and Technology of China, Hefei, China
{qingwang2, jundu}@ustc.edu.cn, {hyhong, yxdong0320}@mail.ustc.edu.cn
[2] iFLYTEK, Hefei, China, {linli13, mqcai, xinfang}@iflytek.com
[3] National Intelligent Voice Innovation Center, Hefei, China, {rywei, jzwu}@iflytek.com

ABSTRACT

This technical report details our submission system for Task 3 of the DCASE 2025 Challenge, which focuses on sound event localization and detection (SELD) in regular video content with stereo audio. In addition to estimating the direction of arrival (DOA) and distance of sound sources, the audio-visual SELD task requires predicting whether the sound source is on-screen. For the audio-only track, we used two-channel log-Mel spectrogram features from stereo audio as model inputs. We adapted the audio-visual pixel swapping (AVPS) technique from first-order Ambisonics (FOA) to stereo format through left-right channel swapping coupled with horizontal video pixel transposition, effectively doubling the training data. Our architecture implemented three specialized models for DOA, distance, and source coordinates estimation tasks, subsequently integrated through a joint prediction framework. The audio-visual track utilized a ResNet-50 model pre-trained on ImageNet for visual feature extraction, enhanced by a teacher-student learning paradigm for cross-modal knowledge distillation. To improve on-screen event detection, we developed a novel two-stage visual post-processing method. Our methods were evaluated using the development set of the DCASE 2025 Task 3.

Index Terms— Sound event localization and detection, audio-visual fusion, Conformer, visual post-processing

1. TRACK A: AUDIO-ONLY INFERENCE

Sound event localization and detection (SELD) refers to the ability of a machine to automatically recognize the temporal activity trajectory of each sound category from a multi-channel audio input and to track the spatial position of the active sound source. In this technical report, we try to address the task with an additional source distance estimation (SDE), formatting a 3D SELD framework that jointly detects sound events, estimates their directions of arrival (DOA), and predicts their distances [1]. To enhance model generalization, we apply advanced audio data augmentation techniques to generate diverse training samples. Our approach leverages the ResNet-Conformer [2,3] architecture, a powerful deep neural network (DNN) optimized for 3D SELD. Prior studies, such as [4] and [5], adopt multi-task learning with dual branches for sound event detection (SED) and DOA estimation. We advance this framework by exploring two integration strategies for distance estimation [6]: (1) a unified DOA-SDE branch merging direction and distance regression; and (2) a modular pipeline combining a standalone SDE model with DOA predictions. Finally, model ensemble techniques are employed to robustly predict sound categories, directions, and distances. This report details the methodology's core components: data augmentation, network training, and model ensemble, providing a effective solution for 3D SELD.

1.1. Audio Data Augmentation

The official DCASE2025 Task3 Stereo SELD Dataset [7,8] contains 41 hours and 42 minutes of stereo audio recordings, each segmented into 5-second clips. The dataset is partitioned into 16,214 training clips and 13,786 testing clips. Given this distribution, data augmentation becomes essential to improve sample diversity and prevent model overfitting. In this challenge, we employ three augmentation strategies to expand the training set.

First, we utilize stereo channel swap (SCS) spatial augmentation, an improved version of our prior method [5]. This technique increases DOA representation by systematically swapping the left and right audio channels, thereby simulating varied spatial configurations. Second, we generate synthetic multi-channel data by convolving single-channel sound samples from the FSD50K dataset [9] with spatial room impulse responses (SRIRs). This process leverages a newly released simulation library [10], enabling us to expand the training dataset with 41.7 hours (30,000 clips) of acoustically diverse samples. This approach effectively simulates realistic acoustic environments while preserving label consistency for SELD tasks. Third, we apply Mixup augmentation [11], which creates new training samples through linear interpolation of both input features and their corresponding targets. This technique enhances model generalization by encouraging smoother decision boundaries and mitigating overfitting, particularly in scenarios with limited labeled data.

1.2. Network Training

In this challenge, we process stereo-format audio data sampled at 24 kHz to extract spatiotemporal features for 3D SELD. A 1024-point short-term Fourier transform (STFT) is applied to each 40 ms Hanning window with a 20 ms hop length, converting the two-channel audio into log Mel-spectrogram features. This yields a time-frequency representation with a feature shape of $2 \times 250 \times 64$ for each 5-second audio segment, ensuring consistent input dimen-

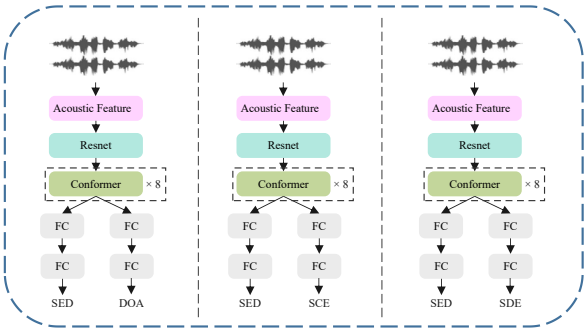


Figure 1: The network architecture of our proposed audio 3D SELD models.

sions for the network. The SCS strategy systematically swaps left and right audio channels, doubling the training data size to about 130 hours. This enhances spatial diversity and robustness in DOA representation.

We adopt ResNet-Conformer [3] as the backbone network, combining the strengths of ResNet and Conformer. This hybrid design captures both local spectro-temporal patterns and long-range dependencies in audio signals, critical for joint sound event detection and localization. The network is optimized for 3D SELD, jointly predicting sound event categories, DOAs, and source distances, leveraging multi-task learning frameworks. In this technical report, we adopt three models with different output formats to handle 3D SELD task, as illustrated in Figure 1. Each model is designed to handle specific aspects of sound event localization and detection, with tailored output formats and loss functions.

Adopting the dual-branch framework proposed by [5], the first SED-DOA model simultaneously performs sound event detection and direction of arrival estimation. The SED branch classifies sound events, while the DOA branch predicts azimuth and elevation angles. To extend localization capabilities to 3D space, the second model integrates source distance into DOA estimation framework. The source coordinate estimation (SCE) branch predicts absolute Cartesian coordinates. The labels of the SCE branch is obtained by multiplying the normalized DOA vectors with the source distance. The mean square error (MSE) loss is applied to the SCE branch, ensuring precise regression of spatial coordinates. This model aims to predict the absolute Cartesian coordinates of the sound source, where the direction of the coordinate vector represents the DOA and the length represents the source distance. This formulation unifies DOA and distance estimation, enabling full 3D localization within a single output branch. Focused on distance-aware applications, the third model decouples distance estimation from directional analysis. The source distance estimation (SDE) branch employs mean square percent error (MSPE) [1] as the loss function for handling the wide dynamic range of distance values. The SED branch remains consistent with the other models, ensuring comparable event detection performance.

1.3. Model Ensemble

To enhance generalization capability and boost overall performance, we employ a model ensemble approach that combines the outputs of three specialized models: SED-DOA, SED-SDE, and SED-SCE. The SED-DOA model predicts sound event directions in Cartesian coordinates of unit length. The SED-SDE model estimates source distances, addressing a critical limitation of traditional SELD frameworks that ignore distance information. The SED-SCE model predicts absolute source positions in Cartesian coordinates, enabling full 3D localization.

To get more robust SED results, we fuse the posterior probabilities of these three models. The final results for direction and distance come from each respective model. Integrating multiple models improves the generalization and 3D SELD performance. The final prediction is a combination of the SED-DOA, SED-SDE and SED-SCE models.

2. TRACK B: AUDIO-VISUAL INFERENCE

2.1. Video Data Augmentation

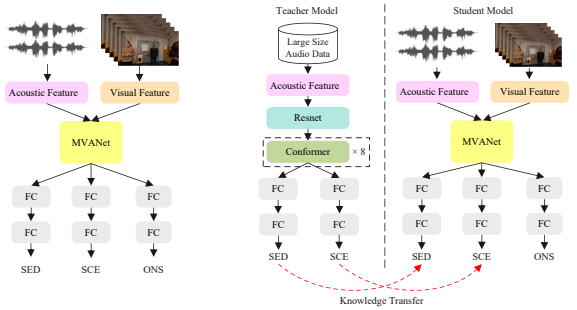
The DCASE2025 Task3 Stereo SELD Dataset contains about 22.5 hours of audio-visual training data [7]. However, a critical limitation arises from the fact that most sound sources are off-screen, which poses a great challenge for audio-visual (AV) 3D SELD methods due to difficulties in modal alignment. To improve the diversity of video data, we apply data augmentation techniques similar to those used for audio data. In audio-only SELD, SCS effectively doubles the training data by swapping left and right audio channels while preserving spatial cues. For the audio-visual dataset, we extend the augmentation method by horizontally flipping video pixels to simulate mirrored perspectives and simultaneously swapping left/right audio channels to maintain spatial correspondence between visual and audio modalities. This augmentation approach yields approximately 45 hours of augmented audio-visual training data.

2.2. Audio-Visual Network Training

The audio-visual Stereo SELD network takes both audio features and visual features as inputs. Audio features are extracted identically to audio-only SELD systems. A pre-trained ResNet-50 [12] backbone extracts frame-level features at 10 fps. Global average pooling is applied to the last convolutional layer, yielding a 7×7 feature map per frame. For the fixed 5-second input segments, this results in a $50 \times 7 \times 7$ feature.

Audio-guided video attention was firstly introduced for audio-visual event localization in [13]. Inspired by this work, we employ the multi-stage video attention network (MVANet) [14] for AV SELD task. Audio embeddings from multiple network layers are used to guide the attention module to focus on spatial information of visual features related to sound events in MVANet, leveraging the complementary characteristics between audio and video modalities.

Our proposed framework extends the SED-SCE model by integrating two key ideas to address the challenges of limited audio-visual data and on-screen estimation. To explicitly model the spatial correspondence between sound events and visual scenes, we augment the SED-SCE architecture with an on-screen (ONS) branch to predict whether the sound event is within the screen. The network structure is shown in Figure 2a. To mitigate data scarcity in the audio-visual domain, we adopt a cross-modal teacher-student learning (TSL) framework, inspired by Jiang [15]. The framework transfers knowledge from a teacher model (trained on large-scale audio-only data with augmentations like SCS and multi-chanel simulation) to a student audio-visual model (trained on limited multi-modal data). The proposed TS-SED-SCE-ONS model is shown in



(a) The SED-SCE-ONS model (b) The TS-SED-SCE-ONS model

Figure 2: The network architecture of our proposed audio-visual 3D SELD model.

Figure 2b. The network structure of the teacher model is ResNet-Conformer, while the network structure of the student model is MVANet.

2.3. Model Ensemble and Post-processing

We train two kinds of audio-visual 3D SELD systems as described in the previous subsection. The submission system is obtained by fusing these two single systems with the model trained on the audio-only track using posterior probability fusion.

Additionally, we adopt a two-stage visual post-processing strategy. In the first stage of visual post-processing, we refine the ONS and DOA predictions of sound sources using the keypoint detection results to generate more accurate DOA and ONS results [15]. In the second stage, we use Grounding DINO [16] to detect potential sound sources in video frames. We design specific prompts for different sound source categories to get more accurate DOA and ONS results for each category.

3. RESULTS ON DEVELOPMENT DATASET

3.1. Results on Track A

We conducted the evaluation of our proposed method using the DCASE2025 Task3 Stereo SELD Dataset. To address data scarcity in Track A, we expand the training set through the data augmentation techniques described earlier. The performance of our 3D SELD systems is compared to the baseline method on the audio-track development dataset, with the results summarized in Table 1. The table presents three variants of our proposed model. In the Table 1, "SED-DOA" denotes the modeling method based on the SED-DOA output format, "SED-SDE" denotes the modeling method based on the SED-SDE output format, and "SED-SCE" denotes the modeling method based on the SED-SCE output format. "Model Ensemble" represents using model ensemble for joint prediction of these three models. Our proposed 3D SELD systems significantly outperform the baseline system.

3.2. Results on Track B

For Track B, we conducted the experiments using approximately 45 hours of audio-visual training data, focusing on fine-tuning the

Table 1: Experimental results of the audio-only 3D SELD systems on the development dataset using stereo format data.

System	$F_{20^\circ,1}$	DOAE	RDE
Baseline-A	0.23	24.50°	0.41
SED-DOA	0.50	12.50°	-
SED-SDE	0.53	-	0.26
SED-SCE	0.50	12.80°	0.27
Model Ensemble	0.54	11.80°	0.26

AV SELD models initialized with audio pre-trained parameters. Table 2 presents the experimental results of the proposed AV SELD methods on the development dataset. All systems in Table 2 adopt the MVANet [14]. In the Table 2, "AV SED-SCE-ONS" represents a single audio-visual model, while "AV TS-SED-SCE-ONS" is an improved audio-visual model trained via the TSL framework. Both of these two models can simultaneously predict the class, azimuth angle, distance of the sound event, and whether it is within the screen. "AV Model Ensemble" represents the fusion of several AV SELD systems with a single audio system, leveraging complementary strengths across modalities. "+PP" indicates the use of a two-stage post-processing method. Our proposed audio-visual system demonstrates significant improvement over the baseline system. The AV TS-SED-SCE-ONS model outperforms its non-TSL counterpart in F-score and DOAE metrics. The ensemble system achieves a 23% improvement in F1-score over the baseline, highlighting the benefits of audio-visual data augmentation and effective multi-model fusion. Post-processing (+PP) further improves F1-score, DOAE, and OSA metrics, underscoring the effectiveness of late-stage visual refinement.

Table 2: Experimental results of the audio-visual 3D SELD systems on the development dataset using stereo format data.

System	$F_{20^\circ,1,on}$	DOAE	RDE	OSA
Baseline-AV	0.20	23.80°	0.40	0.80
AV SED-SCE-ONS	0.38	12.60°	0.28	0.81
AV TS-SED-SCE-ONS	0.39	12.50°	0.29	0.80
AV Model Ensemble	0.43	11.70°	0.26	0.80
+ PP	0.47	11.80°	0.26	0.86

4. REFERENCES

[1] D. A. Krause, A. Politis, and A. Mesaros, "Sound event detection and localization with distance estimation," *arXiv*, 2024.

[2] Q. Wang, Y. Dong, H. Hong, R. Wei, M. Hu, S. Cheng, Y. Jiang, M. Cai, X. Fang, and J. Du, "The nerc-slip system for sound event localization and detection with source distance estimation of dcase 2024 challenge," DCASE2024 Challenge, Tech. Rep., June 2024.

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

[4] 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.

[5] 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.

[6] 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)*, April 2025, pp. 1–5.

[7] K. Shimada, A. Politis, P. Sudarsanam, D. A. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, *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, 2024.

[8] 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.

[9] 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, 2021.

[10] 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)*, Seoul, South Korea, April 2024.

[11] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[13] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 252–268.

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

[15] Y. Jiang, Q. Wang, J. Du, M. Hu, P. Hu, Z. Liu, S. Cheng, Z. Nian, Y. Dong, M. Cai, X. Fang, and C.-H. Lee, "Exploring audio-visual information fusion for sound event localization and detection in low-resource realistic scenarios," *Accepted by International Conference on Multimedia and Expo (ICME)*, 2024.

[16] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, pp. 38–55.