Towards Robust Stereo 3-D SELD: A Study of Perceptual Features and Data Augmentation

Jun-Wei Yeow*, Ee-Leng Tan*, Santi Peksi*, Woon-Seng Gan*, and Qirui Huang*
*Smart Nation TRANS Lab, Nanyang Technological University, Singapore
E-mail: junwei004@e.ntu.edu.sg, {etanel, speksi, ewsgan}@ntu.edu.sg, huang.qirui@ieee.org

***Abstract*—Sound Event Localization and Detection (SELD) unifies Sound Event Detection (SED) and Direction-of-Arrival (DOA) estimation to deliver a unified spatiotemporal representation of auditory scenes. By incorporating Sound Distance Estimation (SDE), SELD extends into three dimensions (3-D SELD), enabling systems to estimate not only the directions but also the distances of sound sources. While 3-D SELD has seen considerable progress with ambisonic and multi-microphone arrays, it remains underexplored for stereo formats commonly found in consumer electronics. In this paper, we present the first comprehensive study of stereo-based 3-D SELD. We evaluate a range of perceptually motivated stereo features, alongside proven augmentation methods originally developed for multi-channel and ambisonic audio. From these results, we then distill our insights into clear best-practice guidelines for developing robust 3-D SELD systems for everyday consumer stereo-based applications. Our implementation is publicly available at this repository[1].**

I. Introduction

Sound Event Localization and Detection (SELD) combines the complementary tasks of Sound Event Detection (SED) and Direction-of-Arrival (DOA) estimation to provide a unified spatiotemporal representation of an acoustic scene [1]. In its original 2D form, SELD systems employ either multi-channel microphone arrays (MIC) or First-Order Ambisonics (FOA) recordings to determine both what sounds occur and where they originate [2]. Recent developments explore 3-D SELD by further incorporating Sound Distance Estimation (SDE), enriching spatial intelligence with explicit distance information [3]. These advances have largely been driven by dedicated challenges such as the Detection and Classification of Acoustic Scenes and Events (DCASE) [4].

However, this progress has largely focused on MIC or FOA-based methods. In contrast, stereo-only 3-D SELD has received little systematic attention, even though stereo recordings are commonplace in consumer devices and offer a low-cost pathway to easily deployable real-world configurations. The transition of the DCASE Challenge 2025 Task 3 to a stereo-only format underscores this need and calls for rigorous evaluation of methods that are fully tailored to the perceptually meaningful cues of the stereo audio format.

In this paper, we present a comprehensive study of stereo-based 3-D SELD that addresses these challenges on two fronts. First, we benchmark an extensive set of perceptually motivated stereo features, quantifying how each set of cues affects detection, localization, and distance estimation. Second, we adapt and extend common 2-D SELD augmentation strategies, such as channel swapping and time-frequency masking. From these results, we distill best-practice guidelines for feature selection and augmentation policies, paving the way for robust 3-D SELD for everyday consumer stereo-based applications.

[1]https://github.com/itsjunwei/NTU_SNTL_Task3

II. Input Features

In this section, we describe the various stereo-based features used to capture both spectral and spatial information.

A. Stereo Log-Mel Spectrograms

Let the two-channel stereo input signal be denoted as $x_\ell[n]$, where $\ell \in \{\mathrm{L}, \mathrm{R}\}$ indexes the left and right channels, and n is the discrete-time sample index. The Short-Time Fourier Transform (STFT) of the ℓ-th channel at time frame t and frequency bin f is denoted as $X_\ell(t, f)$. To approximate human auditory perception, each magnitude-squared spectrogram is converted into a K-band Mel representation using a filter bank $\mathbf{W}_{\mathrm{mel}}$ of size $F \times K$. Specifically, the log-Mel spectrogram (MelSpec) of the ℓ-th channel is

$$\mathrm{MelSpec}_\ell(t,k) = \log_{10}\Big(\sum_{f=0}^{F-1} |X_\ell(t,f)|^2\, \mathbf{W}_{\mathrm{mel}}(f,k)\Big). \quad (1)$$

In the subsequent sections, all described spectral and spatial features are also projected onto the same K-band Mel scale using $\mathbf{W}_{\mathrm{mel}}$ to ensure consistent dimensionality.

B. Mid-Side Conversion and Intensity Vector

To extract explicit inter-channel intensity differences, we convert the stereo waveforms into Mid and Side (M/S) signals, defined sample-wise as

$$m[n] = \frac{x_{\mathrm{L}}[n] + x_{\mathrm{R}}[n]}{2}, \quad s[n] = \frac{x_{\mathrm{L}}[n] - x_{\mathrm{R}}[n]}{2}. \quad (2)$$

The Mid signal $m[n]$ represents the average pressure (omnidirectional component), while the Side signal $s[n]$ captures the horizontal pressure differential. We compute the STFTs $M(t, f)$ and $S(t, f)$ from $m[n]$ and $s[n]$, and each of these is converted to a log-Mel spectrogram exactly as in (1).

An intensity vector (IV) feature analogous to the FOA-based IVs can also be derived [5]. For FOA audio, the active acoustic intensity vector is obtained from the omnidirectional pressure W and the three differential components X, Y, Z. For stereo

979-8-3315-7206-8/25/$31.00

audio, this can be seen as a special 1-D case where only the X-component of the intensity persists. We first compute the real part of the M/S cross-spectrum:

$$I_x(t,f) \;=\; \Re\big\{M(t,f)S^*(t,f)\big\}. \tag{3}$$

The final IV feature is therefore obtained by normalizing over the total instantaneous power:

$$\tilde{I}_x(t,f) \;=\; \frac{I_x(t,f)}{|M(t,f)|^2+|S(t,f)|^2}. \tag{4}$$

C. Spatial Coherence

Distance estimation often benefits from coherence-based features, since magnitude-squared coherence (MSC) values tend to decrease as source distance increases [6]. The cross-power spectral density between the left and right channels is formally defined as

$$\Phi_{\mathrm{L,R}}(t,f) = \mathbb{E}\Big[X_{\mathrm{L}}(t,f)\,X_{\mathrm{R}}^*(t,f)\Big]. \tag{5}$$

In practice, $\Phi_{\mathrm{L,R}}(t,f)$ is estimated using time-recursive averaging [7]:

$$\hat{\Phi}_{\mathrm{L,R}}(t,f) = \lambda\,\hat{\Phi}_{\mathrm{L,R}}(t{-}1,f){+}(1{-}\lambda)\,X_{\mathrm{L}}(t,f)\,X_{\mathrm{R}}^*(t,f), \tag{6}$$

where $\lambda \in [0,1]$ is a smoothing coefficient. In this work, $\lambda = 0.8$ is chosen to effectively balance noise reduction and temporal adaptability [8]. The MSC $\hat{\gamma}(t,f)$ is subsequently calculated as

$$\hat{\gamma}(t,f) = \frac{|\hat{\Phi}_{\mathrm{L,R}}(t,f)|^2}{\hat{\Phi}_{\mathrm{L,L}}(t,f)\,\hat{\Phi}_{\mathrm{R,R}}(t,f)}, \tag{7}$$

where $0 \le \hat{\gamma}(t,f) \le 1$. Generally, MSC values closer to one indicate a more coherent and thereby closer sound source.

D. Inter-channel Phase Relationships

In the context of binaural audio, inter-aural phase differences provide fine-grained localization cues. For stereo audio, the raw inter-channel phase difference (IPD) is given by

$$\mathrm{IPD}(t,f) = \arg\big(X_{\mathrm{L}}(t,f)X_{\mathrm{R}}^*(t,f)\big). \tag{8}$$

To prevent phase-warping, the sines and cosines of the IPD are often further computed [9], yielding a two-channel phase-difference feature set, referred to collectively as SCIPD.

In addition, we explore a normalized inter-channel phase difference (NIPD) based on SALSA-Lite SELD features [10]. The NIPD features include an additional term to compensate for frequency dependency and are defined as follows:

$$\mathrm{NIPD}(t,f) = -\frac{c}{2\pi f}\arg\big(X_{\mathrm{L}}(t,f)X_{\mathrm{R}}^*(t,f)\big), \tag{9}$$

where $c = 343\,\mathrm{m/s}$ is the speed of sound.

III. Augmentation Methods

In this section, we describe commonly used augmentation methods that can improve model robustness and generalization.

A. Channel Swapping

Channel swapping (CS) simulates new source directions by interchanging microphone channels while simultaneously adjusting the corresponding directional labels. Originally introduced for FOA recordings [12], this technique has since been extended to MIC inputs to increase the amount of directional data without altering room acoustics [13].

In the stereo setting, CS reduces to a single permutation: the left and right signals are exchanged while the azimuth labels are mirrored about the frontal axis. In this work, CS is performed at the waveform level to preserve both phase and magnitude information.

B. Time-Frequency Masking

Time-Frequency Masking (TFM) enhances resilience to missing or corrupted spectral content by randomly occluding contiguous blocks in the time-frequency (TF) domain. Two common variants are SpecAugment [14], which applies independent masks along the time and/or frequency axes; and Cutout [15], which masks rectangular patches anywhere in the spectrogram. In our pipeline, we randomly choose between SpecAugment and Cutout with equal probability, then apply an identical mask to both L/R channels.

However, standard TFM can inadvertently eliminate inter-channel level differences (ILDs), which are crucial psychoacoustic cues for stereo-based localization and distance estimation. To address this, we propose and introduce *Inter-channel Level-Aware TFM* (I-TFM). Specifically, for each masked region, we first compute the ILDs pre-masking. Once the mask is applied, we restore the original ILDs by adjusting the masked bins in one of the channels. This ensures that the masking process does not erase the relative intensity cues vital for accurate spatial localization.

C. Frequency Manipulation

While TFM removes information outright, Frequency Manipulation (FQM) alters spectral content to encourage robustness to pitch shifts and varied channel responses. We implement two primary FQM strategies: Frequency Shifting (FreqShift) and FilterAugment (FiltAug). In FreqShift, a random number of frequency bands is shifted up or down, simulating pitch variation in the frequency domain [16]. Complementing this, FiltAug applies random gains on the spectrograms to mimic acoustic filtering effects that can occur in real-world environments [17].

Across both FQM methods, we apply identical shifting or spectral gains to both left and right channels to preserve the relative inter-channel differences. Figure 1 illustrates these FQM methods, alongside TFM variants, applied to a single five-second clip from the STARSS23 stereo set.

IV. Experimental Method

A. Dataset

Our experiments utilize the Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) dataset [11], which comprises

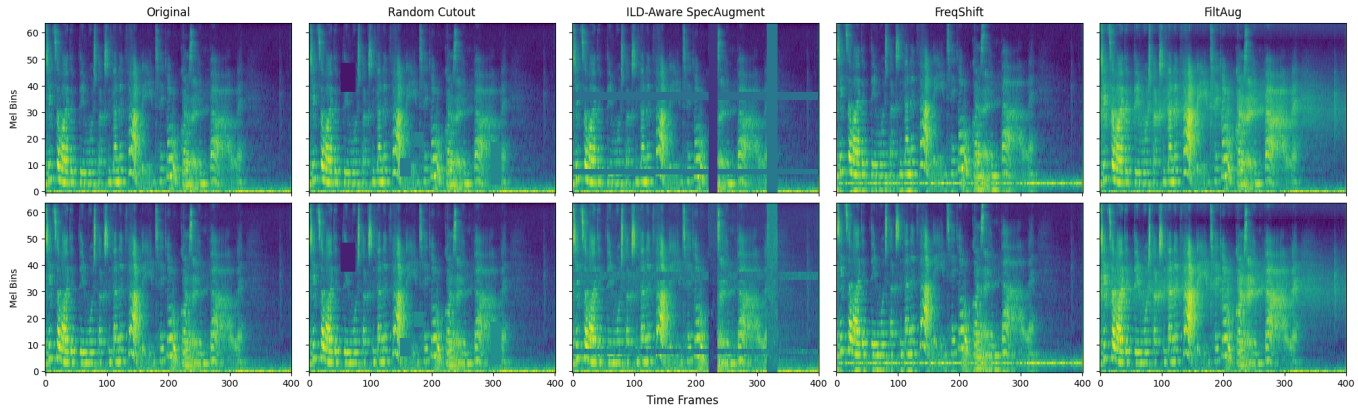

Fig. 1. Augmentation methods on a stereo log-Mel spectrogram pair from a five-second clip from STARSS23 [11]. The top and bottom rows display the spectral plots of the left and right channels, respectively. From left to right, the columns show the original log-Mel spectrograms, standard TFM (Cutout), I-TFM (SpecAugment), FreqShift, and FiltAug. The magnitude of masking or spectral change has been increased for greater visual clarity.

real-world, multi-room recordings with frame-level spatial annotations and serves as the official benchmark for DCASE 2025 Task 3 on stereo 3-D SELD. The original FOA recordings were converted into the stereo format by the task organizers, yielding 22.5 hours of stereo audio. To address class imbalance, we synthesize additional FOA audio using SpatialScaper [18] before converting them into stereo audio using the publicly available DCASE stereo generator[2], resulting in 30,000 additional five-second synthetic stereo clips.

All audio signals are sampled at 24 kHz. For feature extraction, we compute 1024-point STFTs using a Hann window of length 1024 samples and a hop size of 300 samples, yielding 400 time bins for each five-second audio clip. All features are computed using a 64-band Mel filter bank.

B. Baseline System

We adopt SELDNet as the baseline architecture for this work [1]. In particular, we use the same architecture as the baseline model for DCASE 2025 Challenge Task 3[3].

All models are trained for 50 epochs with a batch size of 64, using a tri-stage learning rate scheduler with a peak learning rate of $1\mathrm{e}^{-3}$ and the Adam optimizer with a weight decay of $1\mathrm{e}^{-4}$. All TFM and FQM augmentations are applied on-the-fly and are limited to affect at most 10% of the total TF bins.

C. ACCDDOA Representation

The model predicts a $(C \times 3)$-dimensional Activity-Coupled Cartesian Distance and DOA (ACCDOA) vector per frame, where C is the number of sound classes. For class i, the first two elements $\mathbf{v}_i(t) = [x_i, y_i]$ encode azimuth in Cartesian form, where the Euclidean norm $||\mathbf{v}_i|| \in [0, 1]$ indicates event activity. The third element r_i is a scalar denoting distance. During training, the Mean Squared Error (MSE) is used as the loss function [19].

Because the STARSS23 stereo set exhibits low same-class polyphony, we employ the *single-track* ACCDDOA variant rather than the *multi-track* form [20]. This helps to reduce variation and inefficiencies arising from track ambiguity.

D. Distance Normalization

Raw distance values in STARSS23 range from near-field (0.05 m) to the farthest annotated sound source (7.5 m). Directly regressing these raw values can bias the MSE loss towards distant events [3]. To mitigate these challenges, a Distance Normalization (DN) method can be applied as introduced in [21]. This procedure scales the distribution of distances, d, to a uniform range of $[-1, 1]$ in two steps:

$$d' = \frac{d - \bar{d}}{\sigma_d}, \quad d_{\text{norm}} = \frac{d'}{\max(d')}, \tag{10}$$

where $\bar{d}$ and σ_d represent the mean and standard deviation of all distances, respectively. This normalization ensures that all elements in the ACCDDOA vector lie within the same scale of $[-1, 1]$, preventing larger distances from disproportionately affecting the MSE loss.

E. Metrics

The official 3-D SELD metrics used for the DCASE 2025 Task 3 are adopted for this work [4]. These include the location-dependent F-score ($\mathrm{F}_{20^\circ/1}$), the class-dependent localization error ($\mathrm{LE}_{\mathrm{CD}}$), and the class-dependent relative distance error ($\mathrm{RDE}_{\mathrm{CD}}$). The $\mathrm{LE}_{\mathrm{CD}}$ and $\mathrm{RDE}_{\mathrm{CD}}$ are defined as the mean angular and relative distance errors, respectively, of matched class predictions. The $\mathrm{F}_{20^\circ/1}$ score considers a prediction as a true positive if the localization error is within 20° of the ground truth and the relative distance error is below 1.

An SELD error ($\mathcal{E}_{\mathrm{SELD}}$) combines the three metrics to provide a comprehensive evaluation of overall system performance, calculated as follows:

$$\mathcal{E}_{\mathrm{SELD}} = \frac{1}{3}\left[(1 - \frac{\mathrm{F}_{20^\circ/1}}{100}) + \frac{\mathrm{LE}_{\mathrm{CD}}}{180^\circ} + \mathrm{RDE}_{\mathrm{CD}}\right]. \tag{11}$$

An ideal 3-D SELD system, therefore, aims to maximize $\mathrm{F}_{20^\circ/1}$ while minimizing $\mathrm{LE}_{\mathrm{CD}}$, $\mathrm{RDE}_{\mathrm{CD}}$, and $\mathcal{E}_{\mathrm{SELD}}$.

[2]https://github.com/SonyResearch/dcase2025_stereo_seld_data_generator
[3]https://github.com/partha2409/DCASE2025_seld_baseline

TABLE I
3-D SELD PERFORMANCE OF THE BASELINE WHEN CHANNEL-SWAPPING (CS), DISTANCE NORMALIZATION (DN), AND SYNTHETIC DATA (SD) ARE ADDED INCREMENTALLY.

Experiment	Data (h)	$F_{20^\circ/1}\uparrow$	$LE_{CD}\downarrow$	$RDE_{CD}\downarrow$	$\mathcal{E}_{SELD}\downarrow$
Baseline	22.5	22.8	24.5°	0.410	0.439
+ CS	45.0	26.6	19.2°	0.316	0.386
+ CS + DN	45.0	28.9	**15.3°**	**0.288**	0.361
+ CS + DN + SD	86.7	**36.9**	16.5°	0.301	**0.341**

TABLE II
PERFORMANCE OF BASELINE WHEN DIFFERENT FEATURES ARE ADDED SEPARATELY TO THE BASE L/R MELSPEC.

Experiment	$F_{20^\circ/1}\uparrow$	$LE_{CD}\downarrow$	$RDE_{CD}\downarrow$	$\mathcal{E}_{SELD}\downarrow$
Baseline	36.9	16.5°	0.301	0.341
+ MSC	36.4	16.1°	0.299	0.341
+ SCIPD	34.9	16.6°	0.309	0.351
+ NIPD	35.9	16.5°	0.313	0.349
+ M/S	38.2	15.7°	0.299	0.335
+ M/S + IV	**38.5**	**15.4°**	**0.287**	**0.329**

V. RESULTS

This section presents experimental results on the STARSS23 validation set. All reported results are averaged over three runs.

A. General Methods for Baseline Enhancement

To establish a robust and effective baseline for stereo 3-D SELD, we systematically investigate the incremental impact of three fundamental techniques: CS, DN, and the incorporation of synthetic data (SD). Table I summarizes the performance gains achieved by applying these methods.

From Table I, we observe that the integration of both DN and CS improves 3-D SELD performance across all metrics. Notably, DN refines the model's performance without increasing the dataset size, making it an attractive and simple plug-and-play improvement method.

The incorporation of SD expands the total training set size to 86.7 hours, resulting in the largest $F_{20^\circ/1}$ of 36.9. However, both LE_{CD} and RDE_{CD} increase slightly to 16.7° and 0.302, respectively. This trade-off suggests that while synthetic data may significantly improve event detection, it can lead to domain-shift challenges under real-world conditions. The discrepancies between synthetic and real-world acoustic conditions, such as reverberation characteristics, might not be fully captured by the synthetic generation process [18].

Considering that 3-D SELD requires balanced performance across detection, localization, and distance estimation, all subsequent experiments adopt all three methods (CS, DN, and SD) as the foundational configuration.

B. Input Feature Analysis

This section evaluates the effectiveness of various perceptually motivated stereo features when incorporated individually into the baseline system. Table II quantifies how each feature type contributes to overall 3-D SELD performance.

First, we observe that integrating the coherence-based MSC feature yields only marginal differences in 3-D SELD performance. This suggests that for stereo-based 3-D SELD, the MSC provides negligible additional information beyond what is already captured by the base L/R Mel spectrograms. This could be due to the inherent limitations of coherence alone as a distance cue in complex, reverberant environments [22].

In contrast, incorporating phase-based features detrimentally impacts 3-D SELD performance. Specifically, using SCIPD and NIPD increases the $\mathcal{E}_{SELD}$ by 2.93% and 2.35%, respectively. This degradation can be attributed to two primary factors. Firstly, phase-based features are inherently sensitive to noise [10], making them less reliable in real-world environments. Secondly, and crucially for this dataset, the stereo signals are derived from FOA audio rather than recordings from two physically separated microphones. Consequently, the IPDs in these pseudo-stereo signals are negligible or synthetically generated, lacking the true physical cues that would normally be exploited for accurate binaural localization [23].

Conversely, converting the L/R to M/S spectrograms proves highly beneficial. With only M/S added, $F_{20^\circ/1}$ increases to 38.2 and both LE_{CD} and RDE_{CD} decrease to 15.7° and 0.299, respectively. The most substantial gains are realized when further incorporating the IV feature derived from the M/S components. This combined feature set produces a $F_{20^\circ/1}$ of 38.5, LE_{CD} of 15.4°, and RDE_{CD} of 0.287. Consequently, $\mathcal{E}_{SELD}$ decreases by 3.52% to 0.329, the lowest among all features considered. This conversion process effectively reconstructs the underlying FOA-based representation, providing the network with a more complete and robust set of spatial cues.

TABLE III
PERFORMANCE OF BASELINE WHEN CONSIDERING DIFFERENT AUGMENTATION METHODS APPLIED TO THE BASE L/R MELSPEC.

Experiment	$F_{20^\circ/1}\uparrow$	$LE_{CD}\downarrow$	$RDE_{CD}\downarrow$	$\mathcal{E}_{SELD}\downarrow$
Baseline	36.9	16.5°	0.301	0.341
+ TFM	36.9	15.1°	0.312	0.343
+ ITFM	**38.0**	**14.6°**	0.301	**0.334**
+ FiltAug	37.0	16.8°	**0.287**	0.337
+ FreqShift	**38.0**	16.2°	0.299	0.336

C. Feature-Level Augmentation Strategies

This section investigates the impact of various augmentation methods on 3-D SELD performance. Table III summarizes the results of each augmentation method applied individually.

Firstly, applying standard TFM demonstrates a notable improvement in localization performance, reducing LE_{CD} by 1.4°. However, this comes at the cost of significantly worsened distance estimation accuracy, with RDE_{CD} increasing from 0.301 to 0.312. This adverse effect is likely due to the masking of both stereo channels, which can inadvertently eliminate the ILDs needed for accurate distance estimation.

TABLE IV
PERFORMANCE OF SELECTED STATE-OF-THE-ART DCASE CHALLENGE SUBMISSIONS. DATA USAGE (H), MODEL PARAMETERS (M), AND OFFICIAL DCASE RANKING ARE PROVIDED WITHIN THE BRACKETS.

System	$F_{20^\circ/1}\uparrow$	$LE_{CD}\downarrow$	$RDE_{CD}\downarrow$	$\mathcal{E}_{SELD}\downarrow$
Wang [24] (130h, 58M, #1)	54.3	11.8°	0.260	0.261
Yeow [25] (87h, 4M, #6)	45.3	13.2°	0.262	0.294
Berghi [23] (410h, 30M, #4)	46.0	15.2°	0.308	0.311
He [26] (134h, 104M, #2)	50.0	13.1°	0.360	0.311
Yeow-base (87h, 0.7M, -)	40.1	14.9°	0.286	0.322

Using our proposed ITFM strategy preserves the ILDs within masked regions, yielding a superior trade-off across the metrics: $F_{20^\circ/1}$ increases to 38.0, LE_{CD} drops further to 14.6°, and $\mathcal{E}_{SELD}$ falls to 0.334. Crucially, ITFM manages to maintain the RDE_{CD} at 0.301, effectively avoiding the distance estimation degradation observed with standard TFM. These gains highlight the importance of preserving inter-channel information for robust stereo-based 3-D SELD performance.

For FQM techniques, FreqShift demonstrates performance comparable to ITFM, increasing $F_{20^\circ/1}$ to 38.0 while showing marginal improvements in both LE_{CD} and RDE_{CD}. In contrast, FiltAug achieves the lowest RDE_{CD} of 0.287 among all methods, indicating its specific effectiveness in improving distance estimation accuracy. This suggests that random spectral filtering, which mimics real-world acoustic filtering, can help the network learn robust distance representations.

Overall, the experimental results reveal that no single augmentation method in isolation provides consistent and tangible improvements across all three facets of 3-D SELD. From our analysis, ITFM offers balanced performance improvements for event detection and localization, while FiltAug provides the largest gain for distance estimation. This suggests a potential for synergistic benefits when combining these techniques. We leave this investigation for future work.

D. Benchmarking Against Top DCASE 2025 Systems

To contextualize the performance of our proposed framework and the insights derived from this study, we benchmark our results against top-performing submissions from the recent DCASE 2025 Challenge on stereo-based 3-D SELD. Table IV details the top-ranking systems based on the calculated $\mathcal{E}_{SELD}$.

Our submitted system [25], which ranked 6th on the challenge rankings, achieved an impressive $\mathcal{E}_{SELD}$ of 0.294, which is the second-lowest among all submissions. Our submission utilized a combination of M/S and IV as input features, along with both FiltAug and FreqShift as data augmentation methods, consistent with the findings presented in this paper. To evaluate the isolated effectiveness of our proposed framework, we re-implemented our DCASE submission using the simpler SELDNet baseline architecture, termed *Yeow-base*.

As shown in Table IV, the *Yeow-base* system, despite its modest parameter count of 0.7 M, achieves a respectable $\mathcal{E}_{SELD}$ of 0.322, which would make it the 7th best out of 16 considered systems. This lightweight system also produces competitive LE_{CD} and a remarkably low RDE_{CD} score, surpassing some higher-ranked submissions. These results strongly suggest that our proposed framework provides a powerful foundation for 3-D SELD, and that the benefits are directly transferable; more complex model architectures can better extract fine-grained spatial information from our enriched input features and robust data augmentation.

VI. CONCLUSION

This study provided a comprehensive evaluation of stereo-based 3-D SELD, focusing on critical aspects of feature design and data augmentation. We demonstrated that a straightforward yet effective preparation pipeline of distance normalization, channel swapping, and synthetic data generation elevates the baseline $\mathcal{E}_{SELD}$ by over 20%. Our analysis further revealed that intensity-based cues are markedly more useful than phase-based cues in stereo audio for this task. Crucially, we also highlighted that data augmentation must respect stereo-specific spatial structure. In this regard, our proposed ITFM method delivered the best balance of detection and localization accuracy. Finally, we illustrated how our robust foundational framework can be integrated with more sophisticated model architectures to achieve competitive performance.

For future work, we aim to investigate the optimal combination strategies for these effective input features and data augmentation methods to further enhance robustness, particularly for resource-constrained, online, and real-world 3-D SELD applications.

REFERENCES

[1] 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, 2018.

[2] J. W. Yeow, E.-L. Tan, J. Bai, S. Peksi, and W.-S. Gan, "Real-time sound event localization and detection: Deployment challenges on edge devices," *arXiv preprint arXiv:2409.11700*, 2024.

[3] D. A. Krause, A. Politis, and A. Mesaros, "Sound event detection and localization with distance estimation," *arXiv preprint arXiv:2403.11827*, 2024.

[4] D. Diaz-Guerra, A. Politis, P. Sudarsanam, *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, 2024, pp. 41–45.

[5] 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.

[6] K. Zhagyparova, R. Zhagypar, A. Zollanvari, and M. T. Akhtar, "Supervised learning-based sound source distance estimation using multivariate features," in *2021 IEEE Region 10 Symposium (TENSYMP)*, IEEE, 2021, pp. 1–5.

[7] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1006–1018, 2015.

[8] M. Jeub, C. Nelke, C. Beaugeant, and P. Vary, "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals," in *2011 19th European Signal Processing Conference*, IEEE, 2011, pp. 1347–1351.

[9] D. A. Krause and A. Mesaros, "Binaural signal representations for joint sound event detection and acoustic scene classification," in *2022 30th European Signal Processing Conference (EUSIPCO)*, IEEE, 2022, pp. 399–403.

[10] T. N. T. Nguyen, D. L. Jones, K. N. Watcharasupat, H. Phan, and W.-S. Gan, "Salsa-lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 716–720.

[11] K. Shimada, A. Politis, P. Sudarsanam, *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.

[12] L. Mazzon, Y. Koizumi, M. Yasuda, and N. Harada, "First order ambisonics domain spatial augmentation for dnn-based direction of arrival estimation," *arXiv preprint arXiv:1910.04388*, 2019.

[13] 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.

[14] D. S. Park, W. Chan, Y. Zhang, *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[15] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

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

[17] H. Nam, S.-H. Kim, and Y.-H. Park, "Filteraugment: An acoustic environmental data augmentation method," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 4308–4312.

[18] 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, 2024.

[19] 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)*, IEEE, 2021, pp. 915–919.

[20] 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.

[21] J. W. Yeow, E.-L. Tan, J. Bai, S. Peksi, and W.-S. Gan, "Squeeze-and-excite resnet-conformers for sound event localization, detection, and distance estimation for dcase2024 challenge," DCASE2024 Challenge, Tech. Rep., 2024.

[22] J.-W. Yeow, E.-L. Tan, J. Bai, S. Peksi, and W.-S. Gan, "Enhancing 3d sound event localization and detection with distance estimation using reverberation and spatial coherence features," *IEEE Sensors Journal*, pp. 1–1, 2025.

[23] D. Berghi and P. J. B. Jackson, "Spatial and semantic embedding integration for stereo sound event localization and detection in regular videos," DCASE2025 Challenge, Tech. Rep., 2025.

[24] Q. Wang, H. Hong, R. Wei, *et al.*, "The nerc-slip system for stereo sound event localization and detection in regular video content of dcase 2025 challenge," DCASE2025 Challenge, Tech. Rep., 2025.

[25] J.-W. Yeow, E.-L. Tan, S. Peksi, and W.-S. Gan, "Improving stereo 3d sound event localization and detection: Perceptual features, stereo-specific data augmentation, and distance normalization," DCASE2025 Challenge, Tech. Rep., 2025.

[26] C. He, J. Chen, S. Cheng, J. Bao, and J. Liu, "Stereo sound event localization and detection with source distance estimation using data-driven resnet-conformer ensemble," DCASE2025 Challenge, Tech. Rep., 2025.