EXPLOITING STEREO SPATIAL PROPERTIES WITH RECOOP FRAMEWORK FOR JOINT SOUND EVENT DETECTION AND LOCALIZATION

Banerjee Mohor[1]*, *Nagisetty Srikanth*[2], *Han Boon Teo*[2]

[1] Nanyang Technological University {mohor001@e.ntu.edu.sg}
[2] Panasonic R&D Center Singapore
{srikanth.nagisetty@sg.panasonic.com, hanboon.teo@sg.panasonic.com}

Abstract—**The integration of intelligent systems into daily environments increases the need for a robust understanding of the acoustic scene. Applications such as assistive technologies, audio navigation, and public safety rely on accurate localization and detection of sound events (SELD). Commercially, embedding spatial audio intelligence into smart devices, vehicles, healthcare tools, and surveillance systems, particularly where visual input is limited, has generated significant interest. Traditional signal processing methods struggle to meet the localization and classification demands of compact, microphone-limited devices. As stereo and multichannel audio become prevalent, developing SELD systems capable of joint direction-of-arrival (DoA) estimation and real-world event detection is essential. In response, we propose ReCoOP (ResNet-Conformer with ONE-PEACE), a deep learning framework that combines a ResNet-Conformer backbone with stereo spatial features and contextual embeddings. The system incorporates Interaural Level and Phase Differences, Generalized Cross-Correlation, alongside Mel spectrograms, to model spatial cues, while global semantics are captured through pre-trained ONE-PEACE embeddings. ReCoOP features specialized modules for direction and distance estimation, with outputs fused via a joint head. Evaluated on the DCASE2025 Task 3 dataset, our approach improves performance by approximately 17.8% over the baseline, securing 3rd place in the DCASE 2025 Task 3 challenge.**

Index Terms—**Acoustic scene understanding, Sound event localization, Stereo audio, Spatial audio features, ONE-PEACE embeddings**

1. INTRODUCTION

Sound Event Localization and Detection (SELD) is a fundamental task in Computational Auditory Scene Analysis (CASA), aimed at interpreting acoustic scenes by jointly identifying what sound events occur, when they occur, and where they originate. It comprises three interrelated components: Sound Event Detection (SED) for temporal classification, Direction-of-Arrival (DoA) estimation for spatial localization, and Sound Distance Estimation (SDE) for inferring source proximity. These capabilities enable structured scene understanding across applications such as augmented reality, autonomous navigation, surveillance, and assistive hearing.

The IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) [1] has significantly advanced SELD research by unifying sub-tasks and standardizing evaluation protocols. Earlier editions focused on First-Order Ambisonics (FOA), a 360° spatial audio format, which supported the development of models such as SELDnet [2] (Convolutional Recurrent Neural Network (CRNN)-based joint SED and Cartesian DoA), ACCDOA [3] (regressing 3D DoA vectors for active events), and Multi-ACCDOA [4] (handling overlapping sources from the same class). Event-Independent Network [5], [6] further improved robustness via permutation-invariant training. However, dependence on FOA - requiring specialized microphone arrays - limits practicality for widespread deployment.

In contrast, stereo audio is pervasive in consumer devices such as smartphones, laptops, and webcams, offering a practical alternative for SELD. However, it lacks elevation cues, exhibits front-back ambiguity, and provides limited spatial resolution - particularly for overlapping sources - making localization and distance estimation more challenging. Despite these constraints, the accessibility of stereo microphones renders them well-suited for real-world SELD applications beyond controlled environments.

In response to the growing demand for practical SELD systems, DCASE 2025 Task 3 [7] centers on stereo recordings, emphasizing azimuth-only DoA and SDE. This shift introduces challenges in modeling with limited spatial cues and adapting to two-channel real-world audio. While FOA-based systems like the DCASE 2024 Task 3 Rank 1 system [8] - built on a ResNet-Conformer with multi-branch outputs - achieved strong performance, they were not tailored for stereo input and failed to leverage stereo-specific cues and semantic priors. To address these limitations, we propose ReCoOP, a compact and modular framework purpose-built for stereo SELD.

ReCoOP (ResNet-Conformer-OnePeace) integrates low-level interaural cues - Interaural Level Differences (ILD), Interaural Phase Differences (IPD), and Generalized Cross-Correlation with Phase Transform (GCC-PHAT) - with high-level contextual embeddings from ONE-PEACE [9], a 4-billion-parameter multimodal transformer pretrained on large-scale audio-text and vision-text datasets. It adopts a task-decoupled ResNet-Conformer backbone with separate heads for SED, DoA, and SDE, and employs a lightweight joint ensembling strategy to enhance spatial robustness without additional computational overhead. Evaluated on DCASE 2025 Task 3 [7], ReCoOP outperforms the official baseline across all metrics. Ablation studies confirm the complementary value of interaural spatial cues and pretrained semantic context, highlighting their synergy in advancing stereo SELD.

The remainder of this paper is organized as follows: Section 2 provides an overview of the dataset and its characteristics, Section 3 outlines the acoustic and semantic feature extraction strategies, Section 4 describes the data augmentation techniques, Section 5 details the ReCoOP architecture, Section 6 explains the experimental setup including evaluation metrics, Section 7 presents the results of the ablation studies, Section 8 reports the final system performance, and Section 9 concludes the paper with future directions.

2. DATASET

We use the DCASE 2025 Task 3 Stereo SELD Dataset [7], derived from the STARSS23 dataset [10], consisting of 5-second stereo clips simulating realistic indoor acoustic scenes. STARSS23 provides First-Order Ambisonics (FOA) recordings from 16 rooms with varied layouts, participants, acoustic conditions, and spontaneous background noise. For this task, FOA was converted to stereo via mid-side emulation [11], combining the omnidirectional ($W(n)$) and left-right dipole ($Y(n)$) FOA components at time index n to generate left and

*Work was done during an internship at Panasonic R&D Center Singapore

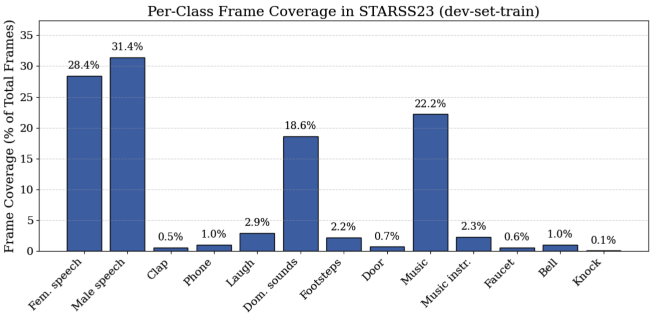

Fig. 1: Per-class frame coverage (%) in the STARSS23 development training set, showing a highly skewed distribution dominated by speech and music classes.

right cardioid microphone signals. The resulting stereo channels $L(n)$ and $R(n)$ are computed as shown in (1):

$$\begin{aligned} L(n) &= W(n) + Y(n) \\ R(n) &= W(n) - Y(n) \end{aligned} \tag{1}$$

Each audio clip is fixed at 5 seconds long and annotated at 100-ms resolution (50 frames) with sound event class, azimuth angle (in degrees), and source-to-microphone distance (in cm). The dataset is challenging, with frequent overlapping events, up to three per frame typically, and occasionally as many as six. To resolve front-back ambiguity, azimuths are folded into the range $[-90°, 90°]$, while elevation is omitted due to top-bottom ambiguity in the stereo setup.

The dataset comprises 30,000 development clips (41.7 hrs) and 10,000 evaluation clips (13.9 hrs), with development audio recorded in Tokyo (24%) and Tampere (76%) and split into training and testing subsets to support generalization. All audio is recorded at 24 kHz, 16-bit resolution. The 13 annotated target classes include speech (female, male), clapping, telephone ringing, laughter, domestic appliances (e.g., vacuum cleaner, boiling water), footsteps, door open/close, music, instruments (e.g., guitar, piano, xylophone), water tap, bell, and knocking. Classes often exhibit high intra-class variability, and speech appears in multiple languages. Sound scenes exhibit significant real-world variability, including non-target interferers such as keyboard typing or clattering dishes, as well as fluctuating background noise levels. Some clips do not contain target events, reflecting the natural sparsity of real-world soundscapes. Stereo recordings are derived from FOA using a length-weighted sampling strategy, producing a frame-level class distribution closely aligned with STARSS23.

However, the distribution is highly skewed, as shown in Fig. 1: male and female speech together comprise around 60% of labeled frames, followed by music and domestic sounds. In contrast, classes like clap, phone, door, faucet, bell, and knock each appear in under 2% of frames. This long-tailed distribution poses challenges for both detection and localization, particularly for rare events with limited temporal coverage and fewer overlapping contexts. Overlapping events - including multiple instances of the same or different classes - are represented by repeated frame entries, with each sound source assigned a consistent identifier. Overall, the dataset offers a comprehensive and realistic benchmark for joint SED and localization under acoustically challenging and imbalanced conditions.

3. FEATURE EXTRACTION

This section outlines the acoustic feature extraction strategies used for input representation. We adopt a dual-pronged approach: a physics-inspired extractor for log-mel spectrograms and inter-channel directional cues, and a transformer-based semantic encoder using ONE-PEACE for global contextual representation.

3.1. Acoustic Features Extraction

We extract a compact and spatially-informative set of acoustic features from each stereo waveform. A Short-Time Fourier Transform (STFT) is applied using a Hann window of length 960 samples and hop size of 480, followed by a non-trainable mel-filterbank projection with 64 mel bins. The log-mel spectrograms of the left and right channels, computed by applying a logarithmic scale to the mel-spectrograms, are retained as the first two channels in the feature tensor.

To model spatial cues, we compute three inter-channel directional features [12]. The Interaural Level Difference (ILD) is defined as

$$\mathrm{ILD}[n,m] = \left| \frac{X_{\mathrm{mel},l}[n,m]}{X_{\mathrm{mel},r}[n,m]} \right|, \tag{2}$$

where $X_{\mathrm{mel},l}[n,m]$ and $X_{\mathrm{mel},r}[n,m]$ denote the complex mel spectrogram coefficients at time frame n and mel bin m for the left and right channels, respectively. This ratio captures the difference in magnitude between channels, which is a key spatial cue above 1.5 kHz [12].

Next, we compute the Interaural Phase Difference (IPD) using

$$\mathrm{IPD}[n,m] = \arg\left(X_{\mathrm{mel},l}[n,m]\right) - \arg\left(X_{\mathrm{mel},r}[n,m]\right), \tag{3}$$

where $\arg(\cdot)$ denotes the phase angle of the complex mel spectrogram at time frame n and mel bin m. This difference expresses the relative phase delay between the two channels. To obtain a more robust representation that avoids discontinuities due to phase wrapping [12], we additionally compute the sine and cosine of the IPD:

$$\mathrm{SI}[n,m] = \sin(\mathrm{IPD}[n,m]), \tag{4}$$

$$\mathrm{CI}[n,m] = \cos(\mathrm{IPD}[n,m]), \tag{5}$$

where $\mathrm{SI}[n,m]$ and $\mathrm{CI}[n,m]$ are the sine and cosine of the interaural phase difference at each time-frequency point, enabling continuous encoding of phase.

Finally, we include the Generalized Cross-Correlation with Phase Transform (GCC-PHAT), defined as

$$\mathrm{GCC}[n,d] = \mathcal{F}^{-1}\left(\frac{X_l[n,k] \cdot X_r^*[n,k]}{|X_l[n,k]||X_r[n,k]|} \right), \tag{6}$$

where $X_l[n,k]$ and $X_r[n,k]$ are the complex STFT coefficients for the left and right channels at time frame n and frequency bin k, $X_r^*[n,k]$ is the complex conjugate of the right-channel STFT, $|\cdot|$ indicates complex magnitude, $\mathcal{F}^{-1}$ is the inverse Fourier Transform, and d denotes the discrete time lag. This feature captures inter-channel time-delay cues critical for direction-of-arrival estimation.

The final feature tensor is constructed by concatenating the log-mel spectrograms with ILD, IPD (both sine and cosine components), and GCC-PHAT features, resulting in a multi-channel representation of shape (6, 251, 64) for each audio clip, where 6 is the number of feature channels, 251 is the number of time bins and 64 is the number of mel bins.

3.2. ONE-PEACE Feature Extraction

To supplement the local spectral features with global semantic cues, we extract contextual embeddings from the ONE-PEACE multimodal transformer model [9]. Specifically, we use a 4B vision-audio-language pretrained checkpoint, trained from scratch on LAION-2B [13] (image-text dataset) and open-source environmental audio-text datasets including AudioCaps [14], Clotho [15], AudioSet [16], FreeSound [17], etc using cross-modal and intra-modal contrastive learning.

The raw stereo waveform is segmented into 100 ms chunks with a hop size equal to the segment length. Each segment is passed through the ONE-PEACE model, which applies tokenization, attention-based encoding, and feature projection to yield a fixed-length vector embedding. These embeddings are then stacked temporally to form a dense contextual representation of the entire audio file, resulting in a tensor of shape (1, 50, 1536), where 1 denotes the batch size, 50 the number of segments, and 1536 the embedding dimension.

These embeddings are designed to capture long-range dependencies and event semantics, making them suitable for hybrid architectures where local spectral features are fused with global context for downstream tasks such as sound event detection or localization.

4. DATA AUGMENTATION

To enhance model robustness and generalization, we incorporated the publicly released synthetic SELD mixtures from DCASE 2024 [18], generated by convolving isolated events with real room impulse responses recorded at Tampere University. We downmixed this dataset using the mid-side stereo procedure described in Section 2 to produce 30,000 stereo clips aligned with the DCASE2025 Task 3 format, effectively doubling the development set duration to 83.4 hours. To further introduce spatial diversity, we applied Audio Channel Swapping (ACS) [19] - a lightweight augmentation that permutes FOA channels to simulate directional variation [20] - on the original FOA-format STARSS23 dataset, increasing its size sevenfold. This ACS-augmented data was then similarly downmixed into 30,000 stereo clips, bringing the development set duration to 125.1 hours in total - tripling its original size.

5. RESNET-CONFORMER-ONEPEACE (RECOOP) DEEP LEARNING FRAMEWORK

ReCoOP, a unified framework for joint sound event detection (SED), direction-of-arrival estimation (DOA), and source distance estimation (SDE) was proposed. Inspired by top-performing systems in DCASE 2024 Task 3, ReCoOP is built on a shared architecture comprising a ResNet-18 encoder [21] without final pooling to extract localized spatial features from the input multichannel acoustic feature tensor. These features are then passed through a linear projection layer to map them to a fixed embedding dimension of 256, followed by a stack of eight Conformer blocks [22] that capture long-range temporal dynamics and cross-channel spatial dependencies. A temporal max-pooling layer with kernel size 5 reduces the sequence length by retaining the most salient activations over short temporal windows, facilitating more efficient downstream processing.

From this backbone, ReCoOP instantiates two specialized variants [8]. The first, shown in Fig. 2, is the SED-DOA model, which augments the pooled acoustic representation by concatenating it with contextual embeddings from ONE-PEACE, providing high-level semantic information. The fused features are passed through another projection layer and two additional Conformer blocks with an embedding dimension of 256, which refine the fused representation by aligning semantic and acoustic contexts. The output is then fed into two parallel heads. The SED prediction head comprises two fully connected (FC) layers with a LeakyReLU nonlinearity, followed by a final FC layer with sigmoid activation to predict the probabilities of each of the 13 sound event classes being active. The DOA prediction head follows the same structure but ends with tanh activation to produce azimuthal direction estimates.

The second variant, the SED-SDE model (Fig. 3), shares the same ResNet-Conformer backbone but omits the ONE-PEACE integration. This variant excludes external semantic embeddings and instead focuses entirely on exploiting the full set of spatial and spectral cues present in the audio input. This design ensures that both detection and distance estimation are guided directly by physically grounded acoustic features. It uses two output heads: a SED prediction head, identical in structure to that in the SED-DOA model, and an SDE prediction head, consisting of two FC layers with LeakyReLU, followed by a final FC layer with ReLU activation to estimate class-wise source distances.

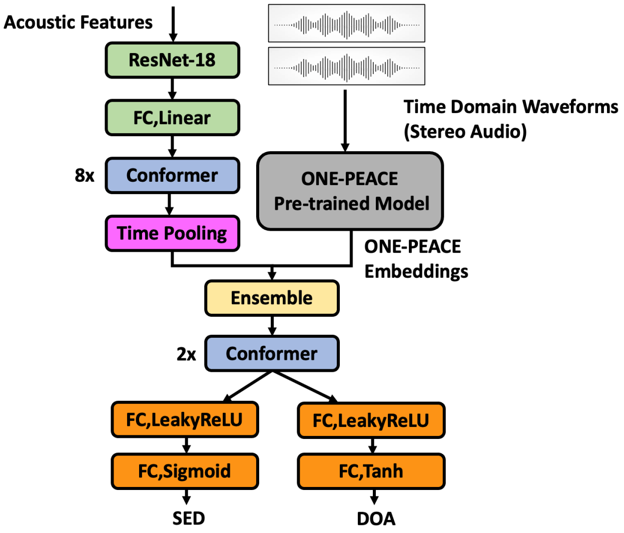

Fig. 2: Architecture for SED-DOA prediction

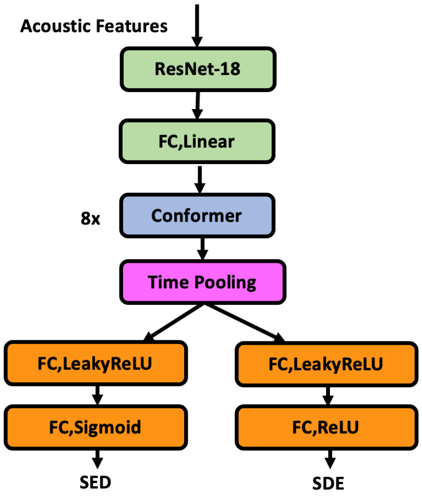

Fig. 3: Architecture for SED-SDE prediction

Final predictions are obtained through task-specific ensembling: SED outputs are averaged across both models to leverage complementary strengths; DOA predictions are taken from the semantically enriched SED-DOA model; and SDE values are obtained exclusively from the physically grounded SED-SDE model. This hybrid design allows ReCoOP to combine semantic context with spatial fidelity, yielding strong performance across all subtasks.

6. EXPERIMENTAL SETUP

The ReCoOP framework was trained using task-specific composite loss functions. For the SED-DOA-OnePeace model, binary cross-entropy loss was used for sound event detection, and mean squared error loss was applied to DOA estimation, gated by SED predictions to focus localization only on active events. For the SED-SDE model, binary cross-entropy loss was similarly used for detection, while the distance regression loss was computed using mean squared percentage error (MSPE) and masked by active event labels to ensure stability. In both cases, the SED and localization losses were weighted with a ratio of 0.1 to 1.

Training was conducted for 300 epochs with a batch size of 32, using the Tri-Stage Learning Rate Scheduler to ensure stable convergence. Final evaluation was based on the macro-averaged location-dependent F1-score (F@20), which captures joint detection and localization performance under thresholds of 20 degrees for azimuth and 1 for relative distance error. In addition, we report DOA error (DOAE) and relative distance error (RDE), both class-dependent metrics that measure unthresholded localization accuracy in azimuth and distance, respectively. Higher F@20 and lower DOAE and RDE values indicate better performance.

During training, the best model checkpoints were selected based on their F@20 performance on the dev-test split, which was used as the validation set throughout the training process.

7. ABLATION STUDY

To assess the individual contributions of each architectural component and training strategy, we conducted an ablation study on the stereo SELD task using the dev-test split of the DCASE2025 Task 3 Stereo SELD dataset. Results are presented in Table 1.

Table 1: Comparison of SELD performance across different experimental configurations

Experiment	F@20 (%)	DOAE (deg)	RDE
Baseline	22.78	24.5	0.41
Baseline + Directional Features	24.70	19.5	0.32
ResNet-Conformer	46.30	13.0	0.37
ReCoOP [submitted system]	48.20	13.3	0.36
ReCoOP + ACS	47.50	13.7	0.36

The baseline system employed a lightweight convolutional-recurrent-attention architecture using only log-Mel spectrograms, achieving an F@20 of 22.78%, though with high localization and relative distance errors. Incorporating spatial acoustic features - specifically interaural level and phase differences (ILD, IPD) and GCC-PHAT - led to measurable improvements: the F@20 increased to 24.70%, and localization performance improved, highlighting the importance of spatial cues.

A substantial performance gain was observed upon adopting the ResNet-Conformer framework, which had shown strong results in previous DCASE challenges. This configuration reached an F@20 of 46.30%, with further reduction in localization error.

Building on this, we integrated pretrained ONE-PEACE embeddings into the ResNet-Conformer setup to form the proposed ReCoOP system. This resulted in the highest detection accuracy, with an F@20 of 48.20% and improved distance estimation - a strong result given its modest 26M parameter size.

Finally, we applied Audio Channel Swapping (ACS) augmentation to ReCoOP, achieving a comparable F@20 of 47.50%. This demonstrated that ACS acts as an effective regularizer, preserving both detection and localization performance.

All experiments included synthetic audio data augmentation, with ACS uniquely applied in the final configuration.

8. RESULTS

This section presents the official evaluation results of our proposed ReCoOP framework on the hidden test set of DCASE 2025 Task 3A, with comparisons to other top-performing submissions. System rankings were based on F@20, and a summary of key metrics is provided in Table 2.

Table 2: Official DCASE2025 Task 3a challenge results. The submitted system Banerjee_NTU_task3a_1 corresponds to the ReCoOP system in Table 1.

System	System Info		Evaluation Metrics		
	Rank	Size	F@20↑	DOAE↓	RDE↓
Du_NERCSLIP_task3a_4	1	58M	50.4	12.2	26.9
He_HIT_task3a_1	2	104M	47.0	13.3	38.6
Banerjee_NTU_task3a_1	3	26M	43.9	14.0	35.2
AO_Baseline	13	734k	26.1	23.0	33.2

Our submission, Banerjee_NTU_task3a_1, ranked third overall, achieving an F@20 of 43.9%, a DOAE of 14.0°, and an RDE of 35.2%. Importantly, ReCoOP accomplished this using only 26 million parameters - less than half the size of the top-ranked system and just one-quarter of the second-ranked - demonstrating a highly favorable F@20-to-model-size trade-off.

The top-ranked Du_NERCSLIP_task3a_4 [23] attained an F@20 of 50.4% using a 58M-parameter ResNet-Conformer ensemble trained on log-mel spectrograms with extensive augmentation. The second-ranked He_HIT_task3a_1 [24] reached 47.0% with a much larger 104M-parameter ensemble and synthetic audio strategies. While both systems demonstrated higher detection accuracy, their substantially larger footprints highlight the efficiency advantage of ReCoOP.

Relative to the official AO Baseline system, which reported an F@20 of 26.1%, ReCoOP achieved a substantial 17.8-point improvement in detection accuracy. This highlights the strength of our approach in bridging the performance gap through architectural and feature-level innovations.

Thus, by integrating directional acoustic cues with contextual ONE-PEACE embeddings , ReCoOP demonstrates that high SELD performance can be achieved within a compact and efficient architecture. This underscores its potential for real-world deployment where computational efficiency is paramount.

9. CONCLUSION

In this work, we proposed ReCoOP, a compact two-model SELD ensemble that combines a rich acoustic feature set - including log-mel spectrograms and interaural directional cues (ILD, IPD, GCC-PHAT) - with contextual embeddings from ONE-PEACE, using a lightweight ResNet-Conformer backbone.

ReCoOP achieved a 25.42% F@20 improvement over the baseline on the validation set and ranked third overall on the official test set, with a 17.8-point F@20 gain. With just 26 million parameters, it achieves competitive performance, demonstrating an efficient F@20-to-model-size trade-off. Future work includes better temporal integration of semantic features, cross-dataset generalization, and self-supervised spatial learning.

10. ACKNOWLEDGMENT

We thank Hock Tong Seow and Harish Vasanth for introducing the ONE-PEACE paper during the early stages of our audio framework development. Their suggestions were instrumental in guiding the integration of contextual embeddings into our model design.

REFERENCES

[1] A. Mesaros, R. Serizel, T. Heittola, T. Virtanen, and M. D. Plumbley, “A decade of dcase: Achievements, practices, evaluations and future challenges,” 2024. [Online]. Available: https://arxiv.org/abs/2410.04951

[2] 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, p. 34–48, Mar. 2019. [Online]. Available: http://dx.doi.org/10.1109/JSTSP.2018.2885636

[3] 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,” 2021. [Online]. Available: https://arxiv.org/abs/2010.15306

[4] 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,” 2022. [Online]. Available: https://arxiv.org/abs/2110.07124

[5] Y. Cao, T. Iqbal, Q. Kong, Y. Zhong, W. Wang, and M. D. Plumbley, “Event-independent network for polyphonic sound event localization and detection,” 2020. [Online]. Available: https://arxiv.org/abs/2010.00140

[6] 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,” 2021. [Online]. Available: https://arxiv.org/abs/2010.13092

[7] D. Diaz-Guerra, A. Politis, P. Sudarsanam, K. Shimada, D. A. Krause, K. Uchida, Y. Koyama, N. Takahashi, S. Takahashi, T. Shibuya, Y. Mitsufuji, and T. Virtanen, “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.

[8] 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.

[9] P. Wang, S. Wang, J. Lin, S. Bai, X. Zhou, J. Zhou, X. Wang, and C. Zhou, “One-peace: Exploring one general representation model toward unlimited modalities,” *arXiv preprint arXiv:2305.11172*, 2023.

[10] K. Shimada, A. Politis, P. Sudarsanam, D. A. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, T. Virtanen, and Y. Mitsufuji, “Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events,” in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 72 931–72 957. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/e6c9671ed3b3106b71cafda3ba225c1a-Paper-Datasets_and_Benchmarks.pdf

[11] J. Wilkins, M. Fuentes, L. Bondi, S. Ghaffarzadegan, A. Abavisani, and J. P. Bello, “Two vs. four-channel sound event localization and detection,” in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, Tampere, Finland, September 2023, pp. 216–220.

[12] D. A. Krause and A. Mesaros, “Binaural signal representations for joint sound event detection and acoustic scene classification,” 2022. [Online]. Available: https://arxiv.org/abs/2209.05900

[13] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, “Laion-5b: An open large-scale dataset for training next generation image-text models,” 2022. [Online]. Available: https://arxiv.org/abs/2210.08402

[14] C. D. Kim, B. Kim, H. Lee, and G. Kim, “AudioCaps: Generating captions for audios in the wild,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 119–132. [Online]. Available: https://aclanthology.org/N19-1011/

[15] K. Drossos, S. Lipping, and T. Virtanen, “Clotho: An audio captioning dataset,” 2019. [Online]. Available: https://arxiv.org/abs/1910.09387

[16] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[17] F. Font, G. Roma, and X. Serra, “Freesound technical demo,” in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM ’13. New York, NY, USA: Association for Computing Machinery, 2013, p. 411–412. [Online]. Available: https://doi.org/10.1145/2502081.2502245

[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, April 2024.

[19] 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,” 2023. [Online]. Available: https://arxiv.org/abs/2101.02919

[20] A. S. Roman, B. Balamurugan, and R. Pothuganti, “Enhanced sound event localization and detection in real 360-degree audio-visual soundscapes,” *arXiv preprint arXiv:2401.17129*, 2024.

[21] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[22] A. Gulati, J. Qin, C.-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,” 2020. [Online]. Available: https://arxiv.org/abs/2005.08100

[23] Q. Wang, H. Hong, R. Wei, L. Li, Y. Dong, M. Cai, X. Fang, J. Wu, and J. Du, “The nerc-slip system for stereo sound event localization and detection in regular video content of dcase 2025 challenge,” DCASE2025 Challenge, Tech. Rep., June 2025.

[24] 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., June 2025.