BASELINE MODELS AND EVALUATION OF SOUND EVENT LOCALIZATION AND DETECTION WITH DISTANCE ESTIMATION IN DCASE2024 CHALLENGE

David Diaz-Guerra[1], *Archontis Politis*[1], *Parthasaarathy Sudarsanam*[1], *Kazuki Shimada*[2], *Daniel A. Krause*[1]
Kengo Uchida[2], *Yuichiro Koyama*[3], *Naoya Takahashi*[4], *Shusuke Takahashi*[3], *Takashi Shibuya*[2]
Yuki Mitsufuji[5,6], *Tuomas Virtanen*[1]

[1] Audio Research Group, Tampere University, Tampere, Finland
[2] Sony AI, Tokyo, Japan [3] Sony Group Corporation, Tokyo, Japan
[4] Sony AI, Zurich, Switzerland [5] Sony AI, NY, USA [6] Sony Group Corporation, NY, USA

ABSTRACT

This technical report presents the objectives, evaluation, and baseline changes for Task 3, Sound Event Localization and Detection (SELD), of the DCASE2024 Challenge. While the development and evaluation dataset, STARSS23, and the division of the task into two tracks, audio-only and audiovisual (AV), remain the same, this year introduces source distance estimation (SDE) along with detection and direction-of-arrival (DOA) estimation of target sound events. Changes in task evaluation metrics and the design and training of the baseline models due to this new SDE subtask are detailed in the report and compared with the previous iteration of the challenge. Further baseline improvements regarding the integration of video information are also presented. Overall, the design of highly effective SELD models evaluated in real scenes with a limited volume of unbalanced training data has proven challenging. The introduction of SDE makes the task even more demanding, as evidenced by the low spatially-thresholded detection scores for both audio-only and AV baselines. While distance estimation error results seem promising, this comes at the expense of lower detection and DOA estimation scores compared to the previous year's baseline models without SDE. Based on the current AV model design, video integration does not bring apparent estimation benefits compared to using only audio input, indicating that more research is required into more effective fusion strategies, model architectures, data augmentation and simulation methods, or training strategies.

Index Terms— Sound event localization and detection, sound source localization, acoustic scene analysis, microphone arrays

1. INTRODUCTION

The sound event localization and detection (SELD) task, detecting the presence of sound events of target classes of interest and tracking their activity and location over time, has seen growing interest from the time of the earliest publications [1]. A large part of the research effort in this topic has been centered around the DCASE challenge[1] and the subsequent workshop, with the task developing every year in terms of data complexity and realism [2–4].

The first three iterations of the task (2019-2021) were based on synthesized spatial recordings including real ambient noise and reverberation. The data were generated with an elaborate synthesis process based on real captured multi-room and multi-point room impulse responses that allowed synthesis of both static and moving reverberant sound events [3]. Some of the task aspects that were considered in these first three SELD challenges were continuous DOA estimation, varying signal-to-noise and direct-to-reverberant ratios, moving sound sources, non-target-class interfering directional sound events, and multiple instances of the same class occurring simultaneously. The top systems of those three first challenges excelled at addressing these problems by employing improved output representations of the SELD objectives [5–7] or advanced data augmentation strategies [8].

However, those synthetic datasets lacked some important aspects of real sound scenes, mainly that of natural temporal and spatial occurences and co-occurences that characterize real sound events and their types as the result of the scene environment and the actions and interactions of the agents in it. To advance SELD research towards that direction, the next iterations up to the current one (DCASE2022-2024) were based on a new dataset of spatial recordings of real scenes [4, 9]. Annotations of sound event activities for 13 sound classes were compiled by human listeners and combined with optical tracking data of the source positions that generated those sound events. 11 hours of such material were collected in multiple rooms of two different sites. Contrary to the fairly balanced earlier synthetic datasets, the presence of classes in the real recordings was highly unbalanced, posing new challenges for the participants. To cope with the increased difficulty of the task and the small amount of training data, participants were allowed to use external data, additional simulations of recordings and pre-trained audio models. Creative use of such resources [10] together with more powerful architectures driven by attention mechanisms [11] allowed the top participants to achieve competitive results with large gains over the baselines.

Additionally, in the 2023 challenge participants were allowed to use 360° video input in addition to the typical audio input [4]; an effort to foster multimodal analysis and development towards diverse large scale training of SELD systems using video supervision. Submissions of audiovisual systems did not exhibit a clear improvement using this additional modality, with only one method achieving better results than using audio-only input. This first iteration of audiovisual SELD models demonstrated that effective integration of video information was not trivial and further research and experimentation was necessary. In this year's DCASE2024 challenge the task setup remains the same, as well as the development and evaluation dataset, but with some important differences introduced otherwise. In this report, an overview of changes in the SELD task of DCASE2024 challenge is presented in terms of task objectives, baseline models, and task evaluation.

[1] https://dcase.community/challenge2024/

2. DISTANCE ESTIMATION

This year, we introduce a new part of the task, namely sound distance estimation. Research on DNN-based techniques for SDE has been largely confined to the binaural format. These studies typically use a classification method, assigning the source within a very limited set of distances or positions [12, 13]. A study by Kushwaha et al. [14] investigated various loss functions for distance estimation and included an activity detection component for a scenario with a tetrahedral microphone array. Few works have explored the simultaneous estimation of distance and DOA [15–17]. Until recently, there has been no effort to combine distance estimation with event detection and localization. In [18], the authors have investigated a single task and multi-task approach to 3D SELD for the binaural format and Ambisonics. Following that paper, we include some of the solutions in this years' baseline to foster further research in this area.

To employ the distance estimation task within the 3D SELD architecture, we use the multi activity-coupled Cartesian Distance and DOA (**multi-ACCDDOA**) method as described in [18]. The method is basically an extension of the multi-ACCDOA output proposed in [19]. Compared with the former, the 3-element DOA vector is extended to include the distance estimate as well. For N tracks, C classes, and T frames, the output is defined as $y_{nct} = [a_{nct}R_{nct}, D_{nct}]$, where n, c, t indicate the output track number, target class, and time frame, $a_{nct} \in \{0, 1\}$ stands for the detection activity, $R_{nct} \in \langle -1, 1\rangle^3$ is the DOA vector, and $D_{nct} \in \langle 0, \infty)$ corresponds to distance values. The dimensions hold the following characteristics: $\mathbf{a}, \mathbf{D} \in \mathbb{R}^{N\times C\times T}$, $\mathbf{R} \in \mathbb{R}^{3\times N\times C\times T}$, and $||\mathbf{R}_{nct}|| = 1$. We model up to $N = 3$ and $C = 13$. The whole output is linear to contain the range of both DOA and distance values. The multi-ACCDDOA model is trained using Auxiliary Duplicating Permutation Invariant Training (ADPIT) as in [19]. The final loss function is defined as:

$$\mathcal{L}^{ADPIT} = \frac{1}{CT}\sum_{c}^{C}\sum_{t}^{T}\min_{\alpha\in \mathrm{Perm}[ct]} l_{\alpha,ct}^{ACCDDOA}, \quad (1)$$

$$l_{\alpha,ct}^{ACCDDOA} = \frac{1}{N}\sum_{n}^{N}\mathcal{L}(y_{\alpha,nct}, \hat{y}_{\alpha,nct}), \quad (2)$$

where $\mathcal{L}(\cdot)$ is the mean square error loss function, α is one possible track permutation and $\mathrm{Perm}[ct]$ is the set of all possible permutations.

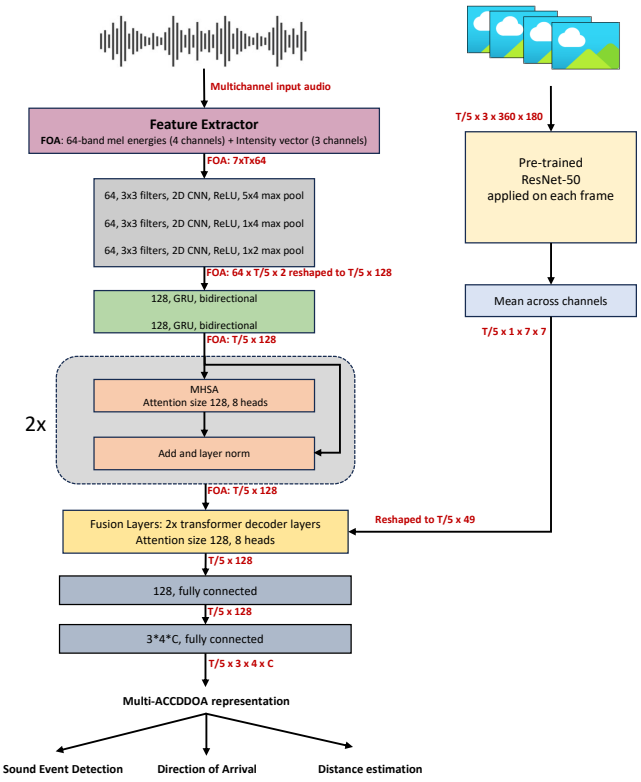

Figure 1: Audiovisual baseline model architecture.

3. BASELINE

For the audio baseline, we retain the same architecture from the previous challenge. It is a modified version of the SELDnet presented in [1]. Last year, we introduced multi-head self-attention blocks in the SELDnet architecture based on the findings in [20].

For the last year's audiovisual baseline [4], an object detector [21] was used to extract visual information. The bounding box outputs were encoded to vectors along with azimuth and elevation [22]. The encoded vectors were treated as visual features in the previous challenge [4]. In this edition of the challenge, the visual pipeline is simplified. Inspired by the work in [23], we use a pre-trained ResNet-50 [24] to extract the visual features from each frame of the video corresponding to the audio input. This visual representation of the input video is combined with the audio representation using audio-visual fusion layers. A transformer decoder block [25] with 2 layers, having an attention size of 128 with 8 heads is used for the fusion of audio and visual features. The new audio-visual baseline architecture used in the challenge is shown in Figure 1.

Differing from previous years, we changed the training procedure to fairly compare the performances of the audio-only model and the audio-visual model. In the previous iterations of the challenge, the audio baseline was trained simultaneously on the synthetic dataset and the train split of the STARSS23 development data. However, it is to be noted that the synthetic data is available only for the audio data and hence direct comparison of the audio-only and the audio-visual models was not possible. To this end, we first trained the audio baseline model on the synthetic dataset and use it for initializing the weights of the audio feature extraction layers for both the audio-only and audio-viusal models. As a second step, we trained both the models on the STARSS23 development dataset. Generation of synthetic data was switched this year from the provided code by the task organizers to the more flexible spatialScaper [26] published recently.

4. EVALUATION METRICS

In previous editions of the challenge, the models were evaluated according to four metrics: localization-dependent F-score (F_{20°) and error rate (ER_{20°) and class-dependent localization error (LE_c) and localization recall (LR_c), all of them computed in one-second non-overlapping segments [2, 27]. One of our goals in this edition was to simplify the evaluation, so we decided to drop the error

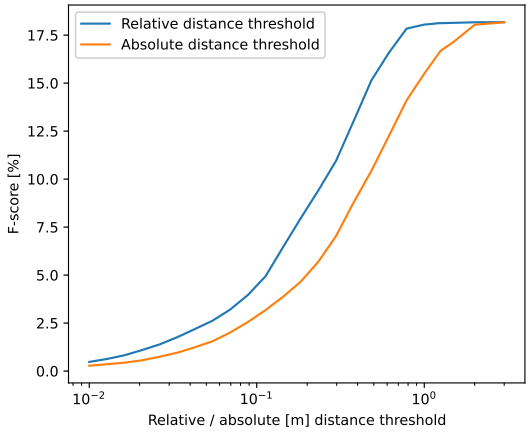

Figure 2: F-score of the 2024 audiovisual baseline system on the evaluation dataset for different values of relative and absolute thresholds. The DOA error threshold was set to 20° in all the experiments.

rate and the localization recall and keep the localization-dependent F-score (which focuses on detection) and the class-dependent localization error (which focuses on the DOA estimation) and to add a new distance estimation metric. In order to make clear that the localization error only evaluates the DOA estimation without taking into account the distance estimation, we renamed it to DOA error ($DOAE_c$).

4.1. Frame-based metrics

The computation of the metrics in one-second non-overlapping segments done in previous challenges [27] was a common practice for evaluating SED systems [28], but not for localization and tracking. It made the metrics and the evaluation code more difficult to interpret and maintain and also prevented them from being extended to more tracking-based metrics in the future, such as measuring the identity-switch ratio, which must be computed at frame level (i.e. for every time output of the system).

Therefore, we decided to compute the metrics at frame level this year. In table 1 we can see the metrics of the top-5 systems resulting from re-evaluating the audio-only systems from the previous challenge at frame level and the comparison with the original segment-based evaluation. We can see how there are no changes in the leaderboard and the metrics slightly degrade but without dramatic changes.

4.2. Distance estimation evaluation

The main novelty of this year's challenge was introducing distance estimation into the SELD task. Since we are now estimating both DOA and distance, we could have combined both into a 3D position estimation and evaluated it just as the Euclidean distance in meters to the actual source position. However, distance estimation is a more difficult task than DOA estimation when working with compact arrays due to the geometrical and physical principles of the problem, so we could expect the errors of the distance estimation to be quite larger than the ones of the DOA estimation. Hence, we preferred to keep the evaluation of both estimations separately.

Also due to the geometrical principles of the problem, distance estimation with compact arrays becomes harder when distance increases (the impact of distance in the phase differences between microphones reduces) so we decided to evaluate the distance in terms of relative distance (i.e. the ratio of the difference between the estimated and actual distance and the actual distance) instead of in absolute terms. This also fits most applications, where an absolute error of a few centimeters is more important if the source is closer to the microphones than if it is several meters away.

We did not want poor distance estimations to penalize the F-score too much this year, so we chose a relative error threshold of 1 so only really large errors have an impact on it. Figure 2 shows how the F-score of the baseline degrades when the distance estimation error threshold is reduced. In the following editions of the challenge, we will adjust the threshold according to the performance of the systems submitted this year.

4.3. Estimate-reference assignment

When we have several estimated and/or reference events of the same class simultaneously, we need to assign the estimates to the references before computing the evaluation metrics. In previous editions of the challenge, we did this by using the Hungarian algorithm [29] to find the assignment that minimized the DOA error. As previously explained, since this year we also have distance estimation, we can compute the localization error (LE) defined as the Euclidean distance between the estimate and the reference position, so we could use the Hungarian algorithm to minimize this metric instead of the DOAE. However, since we are not using this LE as an evaluation metric, we decided to maintain the estimate-reference assignment as in previous editions of the challenge.

Table 2 compares the results of the audio-only baseline model when the assignment is done to optimize the DOAE and the LE. We can see how the differences of both approaches are minimal since this only affects to the situations where there are several concurrent events of the same class, which is not very frequent in the STARSS23 dataset.

	Frame-based					Segment-based				
Rank	Submission	ER_{20°	F_{20°	$DOAE_c$	LR_c	Submission	ER_{20°	F_{20°	$DOAE_c$	LR_c
1	Du_NERCSLIP_task3a_1	0.34	59.8%	12.9°	67.5%	Du_NERCSLIP_task3a_1	0.33	62.7%	12.9°	72.1%
2	Liu_CQUPT_task3a_2	0.37	54.0%	13.7°	61.5%	Liu_CQUPT_task3a_2	0.35	58.5%	13.5°	65.7%
3	Yang_IACAS_task3a_2	0.36	50.2%	16.3°	61.0%	Yang_IACAS_task3a_2	0.35	54.5%	15.8°	66.7%
4	Kang_KT_task3a_2	0.41	48.0%	15.3°	60.7%	Kang_KT_task3a_2	0.40	51.4%	15.0°	63.8%
5	Kim_KU_task3a_4	0.46	46.1%	14.9°	58.1%	Kim_KU_task3a_4	0.45	49.0%	15.0°	62.5%

Table 1: Comparison of the frame-based and segment-based metrics applied to the system of the challenge 2023.

Assignation	F_{20°	$DOAE_c$	RDE_c	LE_c [cm]
DOAE	18.0%	29.6°	0.31	137.6
LE	17.9%	29.7°	0.31	137.4

Table 2: 2024 audio-only baseline results when the assignment between estimates and references of concurrent events of the same class are done to minimize the DOAE or the LE.

5. RESULTS

Incorporating all the changes, Table 3 summarizes the results of the baseline models on the STARSS23 evaluation dataset trained for the SELD task along with distance estimation with the new frame-based metrics using the Multi-ACCDDOA loss. The performance of the models on both 4-channel ambisonic (FOA) and tetrahedral microphone array (MIC) audio formats are presented for comparison.

Dataset	Format	F_{20°	$DOAE_c$	RDE_c
Audio	FOA	18.0%	29.6°	0.31
Audio-visual	FOA	15.5%	34.7°	0.31
Audio	MIC-GCC	16.0%	34.2°	0.30
Audio-visual	MIC-GCC	15.8%	36.0°	0.30

Table 3: Baseline results on STARSS23 evaluation dataset.

Compared with the baselines of the previous edition of the challenge, we can observe a reduction in the performance of the audio-only system. This is due to 1. the addition of the distance estimation task, which makes the problem harder, and 2. the changes in the training pipeline, where synthetic data was only used to pre-train the audio feature extraction layers as done in the audio-visual system. On the other hand, the performance of the audio-visual system has clearly improved compared to the previous edition of the challenge thanks to the changes done in the visual feature extraction, so we are narrowing the gap between both systems. However, further research is still needed to really exploit the visual information of the 360° video input.

6. CONCLUSIONS

This report highlights the changes introduced in the SELD task of DCASE2024 challenge. Most of the changes on baseline models, and task evaluation are associated to the newly-introduced distance estimation objective of the challenge. Distance estimation with a single compact array makes the task significantly more challenging as can be observed from the low baseline results for both audio-only and audiovisual tracks. Training losses and metrics are adapted in order to accommodate the new objective effectively. Audiovisual processing for the currently proposed baseline remains inferior to the baseline using only audio input.

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

[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, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9306885

[3] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 165–169.

[4] K. Shimada, A. Politis, P. Sudarsanam, D. 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," in *Proc. of NeurIPS*, 2023.

[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," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 11–15.

[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)*. IEEE, 2021, pp. 915–919.

[7] T. N. T. Nguyen, D. L. Jones, and W.-S. Gan, "A sequence matching network for polyphonic sound event localization and detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 71–75.

[8] 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.

[9] 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," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

[10] 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.

[11] Y. Shul and J.-W. Choi, "Cst-former: Transformer with channel-spectro-temporal attention for sound event localization and detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8686–8690.

[12] M. Yiwere and E. J. Rhee, "Sound source distance estimation using deep learning: An image classification approach," *Sensors*, vol. 20, no. 1, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/1/172

[13] A. Sobhdel, R. Razavi-Far, and S. Shahrivari, "Few-shot sound source distance estimation using relation networks," 2021. [Online]. Available: https://arxiv.org/abs/2109.10561

[14] S. S. Kushwaha, I. R. Román, M. Fuentes, and J. P. Bello, "Sound source distance estimation in diverse and dynamic acoustic conditions," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2023*. IEEE, pp. 1–5.

[15] M. Yiwere and E. J. Rhee, "Distance estimation and localization of sound sources in reverberant conditions using deep neural networks," in *2017 International Journal of Applied Engineering Research*, 2017.

[16] D. A. Krause, A. Politis, and A. Mesaros, "Joint direction and proximity classification of overlapping sound events from binaural audio," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 331–335.

[17] D. A. Krause, G. García-Barrios, A. Politis, and A. Mesaros, "Binaural sound source distance estimation and localization for a moving listener," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 996–1011, 2024.

[18] D. A. Krause, A. Politis, and A. Mesaros, "Sound event detection and localization with distance estimation," in *32nd European Signal Processing Conference (EUSIPCO)*. (accepted - preprint arXiv:2403.11827), 2024.

[19] 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)*, 2022, pp. 316–320.

[20] P. Sudarsanam, A. Politis, and K. Drossos, "Assessment of self-attention on learned features for sound event localization and detection," in *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2021.

[21] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[22] X. Qian, Z. Wang, J. Wang, G. Guan, and H. Li, "Audio-visual cross-attention network for robotic speaker tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 550–562, 2022.

[23] D. Berghi, P. Wu, J. Zhao, W. Wang, and P. J. Jackson, "Fusion of audio and visual embeddings for sound event localization and detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8816–8820.

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

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[26] 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)*. IEEE, 2024, pp. 1221–1225.

[27] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 333–337.

[28] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, 2016. [Online]. Available: https://www.mdpi.com/2076-3417/6/6/162

[29] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.