

SPATIAL AND SEMANTIC EMBEDDING INTEGRATION FOR STEREO SOUND EVENT LOCALIZATION AND DETECTION IN REGULAR VIDEOS

Technical Report

*Davide Berghi**

CVSSP, University of Surrey, U.K.
davide.berghi@surrey.ac.uk

Philip J. B. Jackson

CVSSP, University of Surrey, U.K.
p.jackson@surrey.ac.uk

ABSTRACT

This report presents our systems submitted to the audio-only and audio-visual tracks of the DCASE2025 Task 3 Challenge: Stereo Sound Event Localization and Detection (SELD) in Regular Video Content. SELD is a complex task that combines temporal event classification with spatial localization, requiring reasoning across spatial, temporal, and semantic dimensions. The last is arguably the most challenging to model. Traditional SELD architectures rely on multichannel input, which limits their ability to leverage large-scale pre-training due to data constraints. To address this, we enhance standard SELD architectures with semantic information by integrating pre-trained, contrastive language-aligned models: CLAP for audio and OWL-ViT for visual inputs. These embeddings are incorporated into a modified Conformer module tailored for multimodal fusion, which we refer to as the Cross-Modal Conformer. Additionally, we incorporate autocorrelation-based acoustic features to improve distance estimation. We pre-train our models on curated synthetic audio and audio-visual datasets and apply a left-right channel swapping augmentation to further increase the training data. Both our audio-only and audio-visual systems substantially outperform the challenge baselines on the development set, demonstrating the effectiveness of our strategy. Performance is further improved through model ensembling and a visual post-processing step based on human keypoints. Future work will investigate the contribution of each modality and explore architectural variants to further enhance results.

Index Terms— Sound Event Localization and Detection, Stereo Sounds, Audio-Visual Machine Learning, Multimodal Localization, Audio Understanding

1. INTRODUCTION

Sound Event Localization and Detection [1] is a combined task that integrates sound event detection (SED) [2] and sound source localization (SSL) [3]. The goal is to identify active sound events from predefined target classes, track their temporal activity, and estimate their spatial positions. SELD systems are crucial for a wide range of real-world applications, including human-robot interaction [4], security monitoring, and immersive media production [5].

*This research was funded by EPSRC-BBC Prosperity Partnership ‘AI4ME: Future personalised object-based media experiences delivered at scale anywhere’ (EP/V038087/1). For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising. Data supporting this study is available from <https://zenodo.org/records/15559774>

Since 2019, SELD has been a dedicated task in the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge. Over successive editions, the task has evolved to include increasingly complex scenarios, such as detecting moving sound sources [6], ignoring external interfering sounds [7], incorporating visual input to enable multimodal SELD in 360° videos [8], and estimating source distance [9]. In this year’s edition, the challenge has shifted from the traditional 4-channel first-order ambisonics (FOA) and microphone array (MIC) formats to stereo SELD using conventional frontal video content, i.e., leveraging only left and right audio channels and perspective video. This format is better aligned with the requirements of conventional media content. In stereo SELD the DOA prediction is limited to the azimuth angles in the range $[-90^\circ, 90^\circ]$, as elevation is ill-defined when using just two horizontally arranged channels, and distinguishing between front and back sources is inherently ambiguous. Source distance estimation, introduced in the previous edition, remains part of the task. Additionally, in the audio-visual track, a new subtask has been introduced: predicting whether sound sources are onscreen or offscreen.

SELD is a non-trivial task, as it requires reasoning across spatial, temporal, and semantic dimensions: spatial information gives cues for direction and distance estimation; temporal information marks movements and onsets/offsets of sound source activity; semantic information identifies objects, their relations and likely behaviors. Recent advances in large language models (LLMs) [10], vision-language models (VLMs) [11], and audio-language models (ALMs) [12] demonstrate that language is a powerful lens for enabling semantic understanding and complex reasoning in relation to media content. Building on this, we posit that leveraging language-aligned models can enhance a SELD model’s semantic reasoning and hence indirectly benefit spatial and temporal reasoning. Traditionally, spatial localization relies on multichannel audio, allowing models to infer source positions through inter-channel time and level differences. However, this dependence limits the use of large-scale pre-training datasets, as language alignment typically requires extensive pre-training on large and diverse datasets. Therefore, we extend our SELD architecture by integrating two existing language-aligned models: CLAP [13] for the audio modality and OWL-ViT [14] for the visual modality. Specifically, we extract audio and visual embeddings using their respective pretrained encoders and combine them with SELD embeddings obtained from a CNN-Conformer backbone, a widely adopted architecture in SELD research [15, 16, 17]. We argue that the embeddings from CLAP and OWL-ViT are semantically rich and provide complementary information to the task. To fuse these embeddings, we employ an adapted version of the Conformer architecture [18], which we refer to as the Cross-Modal Conformer (CMC). This module is used to

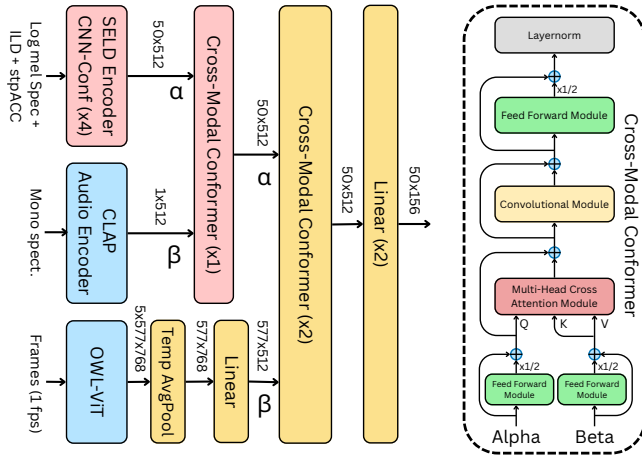


Figure 1: Proposed audio-visual model (left). Blue blocks are frozen, red ones are pre-trained on the audio 5k and audio-visual 2k datasets, yellow on audio-visual 2k dataset only. Cross-Modal Conformer (CMC) architecture adapting the original Conformer [18] to process two generic modalities “Alpha” α and “Beta” β (right).

integrate intra-modal embeddings from different sources (e.g., the SELD audio encoder and CLAP) and inter-modal embeddings (i.e., the combined audio representation and the visual embedding from OWL-ViT). This architecture is outlined in Sec. 2.

This report describes the two audio systems we submitted to track A, and the four audio-visual systems we submitted to track B of the challenge. To tackle the distance estimation subtask, we partnered common SELD input features with short-term power of the autocorrelation stpACC features [19] in Sec. 3. We synthesized and carefully curated large audio and audio-visual datasets used to pre-train our models, described in Sec. 4. Our audio-visual work includes a visual post-processing step based on human keypoint detection and model ensembling, which is described in Sec. 5 on experiments.

2. PROPOSED ARCHITECTURE

The proposed model incorporates a primary SELD encoder that extracts SELD embeddings from multichannel input features. These embeddings are then integrated with the CLAP audio embedding through a cross-modal cross-attention mechanism. We adapted the Conformer architecture to accommodate inputs from different modalities. The resulting audio representation is subsequently fused with a visual embedding extracted using OWL-ViT leveraging a second cross-modal Conformer. Finally, a feed-forward module composed of two linear layers predicts multi-ACCDDOA vectors modeling up to $N=3$ tracks [20], including on- and off-screen activity predictions as in the challenge baseline. The model is trained using class-wise Auxiliary Duplicating Permutation Invariant Training (ADPIT) loss [21, 22, 23].

2.1. SELD Encoder

For the SELD encoder, we adopted a CNN-Conformer architecture, as it is widely adopted in SELD research [15, 16, 17]. It consists of a CNN encoder followed by a Conformer module [18]. The CNN encoder processes stereo input features, with an input of

shape $C_{in} \times T_{in} \times F_{in}$. Where, C_{in} represents the number of feature channels, while T_{in} and F_{in} the number of temporal and frequency bins, respectively. The CNN encoder comprises four convolutional blocks with residual connections, each containing two 3×3 convolutional layers, BN [24], ReLU activation, and Avg pooling with a stride of 2, halving the temporal and frequency dimension at each block. The resulting tensor of shape $512 \times T_{in}/16 \times F_{in}/16$ is reshaped and frequency Avg pooling is applied to achieve a $T_{in}/16 \times 512$ embedding. T_{in} is chosen so that $T_{in}/16$ matches the label frame rate (10 labels/sec). A Conformer module with four layers and eight attention heads processes this embedding, using depthwise convolutions with kernel size of 51.

2.2. Cross-Modal Conformer

The CMC is an adapted version of the Conformer architecture proposed by Gulati *et al.* [18]. A schematic representation is shown on the left side of Figure 1, where a generic modality “Alpha” $\in \mathbb{R}^{T_{\alpha} \times d_k}$ is combined with another modality “Beta” $\in \mathbb{R}^{T_{\beta} \times d_k}$ [4]. Each sub-module within the architecture retains the structure of the original Conformer modules, with the key modifications being two initial feed-forward layers that independently process the two modalities in parallel. Additionally, the standard multi-head self-attention mechanism is replaced by a multi-head cross-attention module, where queries are derived from modality Alpha and keys/values from modality Beta. We utilize two CMC blocks in our model: the first fuses SELD embeddings with semantically rich CLAP embeddings; the second combines the resulting audio representation with semantically and spatially rich visual embeddings extracted by OWL-ViT. We found that a single layer sufficed for audio fusion, while two layers yielded the best performance for audio-visual fusion. Consistent with the Conformer module of the SELD encoder, each CMC employs eight attention heads and depthwise convolutions with a kernel size of 51.

2.3. Video Embedding Extraction

Since SELD involves spatial, semantic, and temporal reasoning, our previous works [16, 25] employed ResNet50 [26] to extract visual features from each individual video frames, capturing the temporal dimension across time. To obtain per-frame visual embeddings, a 7×7 average pooling operation was applied, resulting in a single feature vector for each frame. However, we now argue that this approach degrades spatial resolution, limiting the model’s ability to utilize fine-grained spatial cues. In contrast, other works, such as [27] or this year’s baseline system, apply average pooling across the channel dimension of the ResNet50 output, preserving the 7×7 spatial layout. Yet, we believe that this alternative sacrifices semantic richness, as pooling across channels degrades the learned feature representations.

Effectively managing time, space, and channel information in a unified framework is non-trivial. Nevertheless, we argue that temporal dynamics can be effectively captured by the SELD encoder itself. As such, the visual processing branch should be optimized to better leverage semantic and spatial information. To this end, we sacrifice temporal granularity in the visual stream. We replaced ResNet50 with OWL-ViT [14], a contrastive, language-aligned model, like CLAP, but specifically trained for visual grounding tasks like object detection. As a result, OWL-ViT produces visual embeddings that are both semantically and spatially rich. To retain some temporal context without incurring high computational costs, we sample

video frames at 1 fps. The resulting embeddings are then aggregated via temporal average pooling. All individual ViT token embeddings are maintained to preserve spatial and semantic information.

OWL-ViT requires square input frames of size 768×768 . To avoid distorting the original rectangular frames in the dataset, we explored two pre-processing strategies during preliminary experiments: letterboxing the frames with black bars to preserve the aspect ratio, and applying a non-linear spatial transformation that preserves object proportions in the central region while progressively stretching the top and bottom areas. This second approach is based on the assumption that most sound events occur near the center of the frame. The non-linear re-framing method yielded slightly better performance, and we therefore integrated it into our pre-processing pipeline. Visual tokens are extracted from 32×32 pixel patches, resulting in 576 tokens per frame, plus an additional classification token. Each token embedding has a dimensionality of 768. To align with the model’s architecture, we apply a linear projection to reduce this dimensionality to $d_k=512$. These OWL-ViT embeddings serve as key/value pairs (i.e., modality “Beta”) in the second CMC.

3. ACOUSTIC INPUT FEATURES

Since the left and right audio channels in the dataset are arithmetically derived from FOA signals, rather than captured by two physically separated microphones, they should not present inter-channel time or phase differences. So, we adopted the inter-channel level difference (ILD) as the primary spatial feature for the SELD encoder, alongside log mel spectrograms computed independently from each channel. ILD features are calculated as the ratio of the squared magnitudes of the short-time Fourier transforms (STFTs) of the two channels, and subsequently mapped into the log mel domain:

$$\text{ILD}(m, t) = \log \left(\frac{|\mathbf{L}(f, t)|^2 + \epsilon}{|\mathbf{R}(f, t)|^2 + \epsilon} \right), \quad (1)$$

where $\mathbf{L}(f, t)$ and $\mathbf{R}(f, t)$ are the STFTs of the left and right channels, respectively, \mathbf{H}_{mel} is the mel filter bank, ϵ is a small constant to avoid division by zero and instability, and m in the mel frequency index. To further support the model in the distance estimation sub-task, we include short-term power of the autocorrelation (stpACC) features, as proposed in [19]. The final concatenation of acoustic input features has a channel dimension of $C_{\text{in}}=4$.

4. PRE-PROCESSING AND DATA AUGMENTATION

We pre-trained our model on synthetic data while keeping the CLAP audio encoder and OWL-ViT weights frozen. Synthetic FOA audio was generated using SpatialScaper [28], which convolves FSD50K sounds [29] with RIRs from various datasets [6, 30, 31, 32, 33, 34, 35]. We created 5,000 one-minute FOA clips averaging 18 events per clip (std=6), and 500 additional clips for validation. Background noise levels were sampled from a normal distribution (mean=-65 dB, std=15) to encourage robustness. We overrepresented “Knock” and “Bell” sounds due to their detection difficulty. Using the provided stereo SELD generator, we split each FOA clip into twelve 5-second segments (i.e., with hop size=5s), filtered out silent ones, and applied four random FOA rotations to each segment before extracting the stereo sounds, resulting in $\sim 150\text{k}$ training clips. We’ll refer to this synthetic dataset as “audio 5k”, as it is derived from 5,000 FOA files. Models were then fine-tuned on the real content of the stereo STARSS dataset.

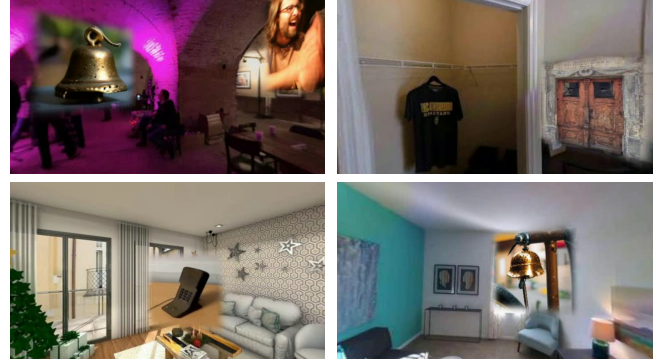


Figure 2: Examples of synthetic scenes (clockwise from top left) in a restaurant, a walk-in wardrobe, a hotel room, a first-floor flat. The foreground tiles (bell*, laughter, door*, bell*, telephone) are applied with a soft cross-fade to avoid strong edges. * Generated with NitroFusion [39].

For the audio-visual model, we generated an additional 2,000 FOA clips with SpatialScaper and synthesized corresponding videos using SELDVisualSynth [36]. To enrich the quality of the visual output, we manually curated and extracted class-relevant images from Flickr30k [37]. We also selected images of doors from DoorDetect [38], which have been employed for “door open/close” and “knock” sounds. Furthermore, we created additional synthetic foreground images for each sound class using NitroFusion [39]. Instead of employing the background canvases provided with SELDVisualSynth which consists primarily of outdoor environments, we adopted the +2,000 environments of the 360-Indoor dataset [40]. We also implemented a soft cross-fade between background canvases and foreground object tiles to remove strong artificial edges from the pre-training dataset. Figure 2 shows example frames from the dataset. We initialized the SELD encoder and audio CMC with weights from the audio-only pre-training, and continued training on the +60,000 audio-visual clips generated following this approach. We refer to this dataset here as “audio-visual 2k”.

To further increase the size of the training dataset and model robustness, we adapted augmentations to the stereo scenario, the audio-channel swap (ACS) [15] and video pixel swap (VPS) [16, 41, 42]. Thus, we swapped the left and right audio channels, and flipped the corresponding video frames to double the size of the training data. The synthetic data augmented with such methods gave over 410h of stereo audio data and an additional 168h of audio-visual data.

5. EXPERIMENTS

5.1. Implementation Details

Audio spectrograms were generated via STFT with a 512-point Hann window and 150-sample hop size. With a sampling rate of 24 kHz, this produces 800 temporal bins for the 5-second input clips. We used 64 mel bins for our spectrograms. For stpACC features, we applied an STFT with a 1014-point Hann window. This ensures that the autocorrelation covers delays up to approximately 20 ms after the direct sound. We then downsample the time-lag dimension by a factor of 8 to achieve 64 bins and allow concatenation with the other features. More details about stpACC can be found in [19]. The SELD encoder and the first CMC are pre-trained on the

synthetic audio 5k dataset for 100 epochs. Subsequently, the full audio-visual model is further trained on the synthetic audio-visual 2k dataset for other 80 epochs. Finally, the model is fine-tuned on the training partition of the real development set for 80 epochs, selecting the best based on the highest F1 score on the test partition. Models are trained with batch size 32 using Adam optimizer. After the first 30 epochs, the learning rate is reduced by 5% per epoch.

Our audio-only model is similar to the audio-visual one, but without the OWL-ViT branch. Additionally, the second CMC is replaced by a classic Conformer unit to match the depth of the audio-visual model. Since the stereo viewing angle in the dataset is randomly sampled from the full 360° frames of the STARSS23 dataset [8], the number of off-screen sound events is significantly higher than that of on-screen events. As a result, we observed that the model tends to consistently predict “off-screen”, effectively suppressing on-screen predictions. This behavior allows the model to reach a stable $\sim 80\%$ accuracy, reflecting a bias toward the dominant class. To address this imbalance, we train a second model using a modified loss function, in which the binary cross-entropy (BCE) loss for on/off-screen predictions is multiplied by a factor of 4.0 whenever the ground truth label is on-screen.

5.2. Visual Post-Processing

Inspired by the work of Jiang *et al.* [41], we implemented a visual post-processing step based on human keypoint detection. Specifically, we process the video frames using YOLOv11-Pose [43] to extract keypoints. We then map specific sound event classes to corresponding human keypoints: female speech, male speech, and laughter are associated with the nose keypoint; clapping with the average position of the left and right wrists; and footsteps with the average position of the left and right ankles. If the predicted direction of arrival (DOA) for any of these classes falls within 20° of the corresponding keypoint’s direction, we replace the predicted position with that of the keypoint. We found that this method resulted in a slight decrease in spatial accuracy. This drop may be attributed to events being mistakenly linked to nearby individuals who are actually located in the opposite hemisphere but with similar lateral position, or distortions introduced by the perspective projection, which causes a non-linear mapping between pixel coordinates and DOA values. For this reason, we ultimately chose not to apply this spatial refinement technique to the DOA predictions. However, we do use it to set the on/off-screen label to “on-screen” for all sound event predictions associated with a visible human. As a result, this post-processing step only affects the on/off-screen accuracy and the $F_{\leq 20^\circ/1/on}$ metrics.

5.3. Ensemble Strategy

To further improve performance, we employ an ensemble strategy that combines predictions from multiple systems. At each time frame, a sound event is considered active only if it is detected by at least two systems, and their predicted DOAs are within 20° of each other. For such events, the ensemble DOA is computed as the average of the DOAs from all systems that detected the event. There are two exceptions to this rule: for the “Bell” and “Knock” classes, a detection from even a single system is sufficient to consider the event active, as in [25]. Similarly, for the on/off-screen classification, if at least one system predicts the event as on-screen, the ensemble output is also considered on-screen.

Table 1: Results on the development set of stereo STARSS23 [8]: $F_{\leq 20^\circ/1}$ (F1) [%], $F_{\leq 20^\circ/1/on}$ (F1o) [%], Direction Of Arrival Error (DOAE) [°], On/Off Accuracy (Acc) [%]. The models audio-only (AO) or audio-visual (AV); variants with weighted loss (+W), visual post-processing (+P); Ensemble (Ens.) combines (1),(2),(3) and (4). Submitted challenge models are denoted as ^{a1,a2,b1,b2,b3,b4}.

Model	F1 ↑	F1o ↑	DOAE ↓	RDE ↓	Acc ↑
Baseline AO	22.8	-	24.5	41.0	-
Baseline AV	26.8	20.0	23.8	40.0	80.0
(1) AO ^{a1}	45.7	-	15.0	31.0	-
(2) AO ^{a2}	46.0	-	15.2	30.8	-
(3) AV	44.4	34.0	15.6	30.4	80.5
(3.1) AV+P ^{b1}	44.4	34.4	15.6	30.4	80.5
(4) AV+W	45.5	35.4	15.2	32.2	80.8
(4.1) AV+W+P ^{b2}	45.5	35.7	15.2	32.2	81.0
(5) Ens. ^{b3}	48.0	37.3	14.0	29.3	80.8
(5.1) Ens.+P ^{b4}	48.0	37.5	14.0	29.3	80.8

5.4. Results

The results for our audio-only (AO) and audio-visual (AV) systems are presented in Table 1. The two AO systems are identical but were trained with slightly different initial learning rates. We chose to report and submit both, as they both contribute to improving the ensemble performance. System (4) uses the same AV model as system (3), but with the loss function that weights the on/off-screen predictions more heavily when the event is on-screen. The ensemble model (5) combines predictions from systems (1), (2), (3) and (4). Finally, applying visual post-processing to systems (3), (4) and (5) results in the enhanced versions: (3.1), (4.1) and (5.1), respectively.

Both AO and AV systems substantially outperform the challenge baselines. The AO systems achieved slightly better $F_{\leq 20^\circ/1}$ and DOAE scores compared to the base AV systems (3) and (4), which also need to predict on/off-screen labels, adding an extra layer of complexity. Using a weighted loss function for on/off-screen classification gave only a minor gain in overall on/off accuracy. However, it led to a 1.4 percentage point gain in $F_{\leq 20^\circ/1/on}$. The visual post-processing step, which only modifies on-screen predictions, brought small gains in $F_{\leq 20^\circ/1/on}$. $F_{\leq 20^\circ/1}$, DOAE and RDE remained unchanged as post-processing has no effect on localization and class predictions. Finally, we observed RDE improvements of approximately 10 percentage points compared to the baselines. These gains may be attributed in part to the inclusion of stpACC features, which enhance the model’s ability to estimate distance.

6. CONCLUSION

This technical report describes the systems submitted to Task 3 of the DCASE 2025 Challenge. Our approach leverages a model that integrates semantically rich feature embeddings from CLAP and OWL-ViT, fused through an adapted Conformer architecture. The model is pre-trained on large, curated synthetic audio and audio-visual datasets. All submitted systems outperform both the audio-only and audio-visual baselines on the test partition of the development set. Additional performance gains are achieved through model ensembling and visual post-processing.

7. REFERENCES

- [1] S. Adavanne et al., “Sound event localization and detection of overlapping sources using convolutional recurrent neural networks,” *IEEE J. of Selected Topics in Sig. Proc.*, vol. 13, pp. 34–48, 2019.
- [2] S. Adavanne et al., “Sound event detection using spatial features and convolutional recurrent neural network,” in *ICASSP*, 2017.
- [3] S. Adavanne, A. Politis, and T. Virtanen, “Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network,” in *EUSIPCO*, 2018, pp. 1462–1466.
- [4] X. Qian et al., “Audio-visual cross-attention network for robotic speaker tracking,” *IEEE/ACM Trans. Audio, Speech & Lang. Proc.*, vol. 31, pp. 550–562, 2023.
- [5] D. Berghi et al., “ForecasterFlexOBM: A multi-view audio-visual dataset for flexible object-based media production,” in *ICME*, 2024.
- [6] A. Politis et al., “A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection,” in *DCASE Workshop*, 2020.
- [7] A. Politis et al., “A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection,” in *DCASE Workshop*, 2021.
- [8] K. Shimada et al., “STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events,” in *NeurIPS*, 2023.
- [9] D. Diaz-Guerra et al., “Baseline models and evaluation of sound event localization and detection with distance estimation in DCASE 2024 Challenge,” in *DCASE Workshop*, 2024, pp. 41–45.
- [10] T. Brown et al., “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, 2020.
- [11] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Int. Conf. on Machine Learning*, 2021.
- [12] R. Huang et al., “AudioGPT: understanding and generating speech, music, sound, and talking head,” in *AAAI Conf. on Artificial Intelligence*, 2024.
- [13] Y. Wu et al., “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,” in *ICASSP*, 2023.
- [14] M. Minderer et al., “Simple open-vocabulary object detection,” in *ECCV*, 2022.
- [15] Q. Wang et al., “A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection,” *IEEE/ACM Trans. Audio, Speech & Lang. Proc.*, vol. 31, pp. 1251–1264, 2023.
- [16] D. Berghi et al., “Fusion of audio and visual embeddings for sound event localization and detection,” in *ICASSP*, 2024.
- [17] L. Xue et al., “Resnet-conformer network using multi-scale channel attention for sound event localization and detection in real scenes,” in *Int. Conf. on Wireless Comms. & Sig. Proc.*, 2023.
- [18] A. Gulati et al., “Conformer: Convolution-augmented transformer for speech recognition,” in *Interspeech*, 2020, pp. 5036–5040.
- [19] D. Berghi and P. J. B. Jackson, “Reverberation-based features for sound event localization and detection with distance estimation,” *ArXiv*, vol. abs/2504.08644, 2025.
- [20] D. A. Krause et al., “Sound event detection and localization with distance estimation,” in *EUSIPCO*, 2024, pp. 286–290.
- [21] K. Shimada et al., “Multi-ACCDOA: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training,” in *ICASSP*, 2022, pp. 316–320.
- [22] Y. Cao et al., “Event-independent network for polyphonic sound event localization and detection,” in *DCASE Workshop*, 2020.
- [23] Y. Cao et al., “An improved event-independent network for polyphonic sound event localization and detection,” in *ICASSP*, 2021.
- [24] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International Conference on Machine Learning*, vol. 37, 2015, pp. 448–456.
- [25] D. Berghi and P. J. B. Jackson, “Leveraging reverberation and visual depth cues for sound event localization and detection with distance estimation,” in *Technical Report of DCASE Challenge*, 2024.
- [26] K. He et al., “Deep residual learning for image recognition,” in *IEEE Conf. on Comp. Vis. & Pattern Recogn. (CVPR)*, 2016, pp. 770–778.
- [27] H. Hong et al., “MVANet: Multi-stage video attention network for sound event localization and detection with source distance estimation,” *ArXiv*, vol. abs/2411.14153, 2024.
- [28] I. R. Roman et al., “Spatial Scaper: A library to simulate and augment soundscapes for sound event localization and detection in realistic rooms,” in *ICASSP*, 2024.
- [29] E. Fonseca et al., “FSD50K: an open dataset of human-labeled sound events,” *IEEE/ACM Trans. Audio, Speech & Lang. Proc.*, vol. 30, pp. 829–852, 2022.
- [30] O. Olgun and H. Hacıhabiboglu, “METU SPARG Eigenmike em32 Acoustic Impulse Response Dataset v0.1.0,” Apr. 2019. [Online]. Available: <https://doi.org/10.5281/zenodo.2635758>
- [31] T. McKenzie, L. McCormack, and C. Hold, “Dataset of spatial room impulse responses in a variable acoustics room for six degrees-of-freedom rendering and analysis,” *ArXiv*, vol. 2111.11882, 2021.
- [32] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, “FMA: A dataset for music analysis,” *ArXiv*, vol. abs/1612.01840, 2016.
- [33] G. Götz et al., “A dataset of higher-order ambisonic room impulse responses and 3d models measured in a room with varying furniture,” in *Immersive and 3D Audio (I3DA)*, 2021.
- [34] G. Chesworth, A. Bastine, and T. Abhayapala, “Room impulse response dataset of a recording studio with variable wall paneling measured using a 32-channel spherical microphone array and a B-format microphone array,” *Applied Sciences*, vol. 14, no. 5, p. 2095, 2024.
- [35] C. Schneiderwind et al., “Data set: Eigenmike-DRIRs, KEMAR 45BA-BRIRs, RIRs and 360° pictures captured at five positions of a small conference room,” in *Proc. 45th DAGA*, 2019.
- [36] A. S. Roman, A. Chang, G. Meza, and I. R. Roman, “Generating diverse audio-visual 360 soundscapes for sound event localization and detection,” *ArXiv*, vol. abs/2504.02988, 2025.
- [37] B. A. Plummer et al., “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models,” in *ICCV*, 2015.
- [38] M. Arduengo et al., “Robust and adaptive door operation with a mobile robot,” *Intelligent Service Robotics*, p. 409–425, 2021.
- [39] D.-Y. Chen et al., “NitroFusion: High-fidelity single-step diffusion through dynamic adversarial training,” *ArXiv*, vol. abs/2412.02030, 2024.
- [40] S.-H. Chou et al., “360-Indoor: Towards learning real-world objects in 360° indoor equirectangular images,” in *IEEE Winter Conf. on Applications of Comp. Vis.*, 2020.
- [41] Y. Jiang et al., “Exploring audio-visual information fusion for sound event localization and detection in low-resource realistic scenarios,” in *ICME*, 2024, pp. 1–6.
- [42] A. S. Roman et al., “Enhanced sound event localization and detection in real 360-degree audio-visual soundscapes,” *ArXiv*, vol. abs/2401.17129, 2024.
- [43] G. Jocher, Q. Jing, and A. Chaurasia, “Ultralytics YOLO,” Accessed Jun. 21, 2025, 2023, <https://github.com/ultralytics/ultralytics>.