ENHANCED UPLAM-MASK R-CNN SYSTEM FOR SEMANTIC ACOUSTIC IMAGING AND SOUND EVENT LOCALIZATION

Technical Report

Yishuo Yang[1], Xinwei Wu[1], Chunrui Zhao[1], Huayang Wang[1], Zheng Wen[1], Jilu Jin[2], Gongping Huang[1]

[1]School of Electronic Information, Wuhan University, Wuhan, China
[2]CIAIC, Northwestern Polytechnical University, Xi'an, China
yang_ys@whu.edu.cn,xinwei_wu@whu.edu.cn,chunruizhao@whu.edu.cn
llwwwhycc@whu.edu.cn,wenzheng104@whu.edu.cn
charles.jilu.jin@gmail.com, gongpinghuang@whu.edu.cn

ABSTRACT

This report presents our UpLAM-Mask R-CNN system for DCASE 2026 Task 3: semantic acoustic imaging (SAI) for sound event localization and detection (SELD). The task requires detecting sound event categories while estimating their spatial acoustic energy distributions and source distances. Based on the official UpLAM-Mask R-CNN baseline, we develop unified frameworks for both the audio-only and audio-visual tracks. Specifically, we introduce modality-adaptive input switching for Track A and Track B, optimize the UpLAM-based acoustic feature extraction pipeline, preserve acoustic energy magnitude cues throughout the model, and enhance RoI-level energy-map prediction, full-frame energy decoding, and distance regression. During inference, temporal tracking, confidence-based filtering, and region-aware energy-map export are further applied to improve the stability and completeness of the predictions.

***Index Terms*—** Semantic acoustic imaging, sound event localization and detection, audio-visual fusion, feature pyramid network, multi-task learning

1. INTRODUCTION

Spatial perception is an essential capability for autonomous systems, including service robots, intelligent cockpits, and security monitoring systems, to interpret complex real-world environments. Sound event localization and detection (SELD) is a multi-task learning problem that jointly addresses sound event recognition and spatial position estimation [1, 2]. Previous SELD frameworks typically adopt a point-source assumption, representing sound sources as discrete coordinates in 3D space or as direction-of-arrival (DoA) vectors in polar coordinates [3]. Although this formulation is computationally efficient, it has inherent limitations in practical acoustic environments [4]. Real-world sound sources often exhibit spatial extent, and their acoustic energy is therefore distributed over space rather than concentrated at a single point. Conventional SELD frameworks are thus insufficient for accurately characterizing such spatial energy distributions.

To address these representational limitations, Task 3 of the 2026 DCASE Challenge aims to explore a semantic acoustic imaging (SAI)-based SELD framework [5]. In this formulation, acoustic scenes are modeled as high-resolution energy fields, where dense semantic masks represent the spatial distributions of sound events. This transforms the original coordinate regression problem into a task analogous to semantic segmentation in computer vision [6].

A major technical challenge of this task lies in high-resolution reconstruction under constrained observation conditions. Since the system is required to estimate high-resolution spatial source distributions from low-channel, e.g., 4-channel, raw audio signals, it must reconstruct fine-grained SAI annotations originally generated from high-channel reference data using inputs with limited spatial information. This process can be regarded as a complex spatial super-resolution problem, requiring the model to establish robust acoustic-semantic associations while recovering detailed sound-field structures from sparse spatial observations.

To this end, we propose a spatio-temporal semantic acoustic imaging framework that leverages universal pretrained localization and mapping (UpLAM) [7] for multi-band acoustic image generation, and employs a residual network-feature pyramid network (ResNet-FPN) backbone [8, 9] with a Transformer neck for robust localization [10]. Precise source-boundary modeling is achieved through RoI-level prediction heads and discriminative loss functions. For Track B, visual semantic priors are further incorporated to enhance audio-visual (AV) fusion and improve overall performance.

2. DATASET PROCESSING

2.1. Feature Extraction

We use the STARSS23[1] and STAIRS26[2] datasets. STARSS23 contains real-world audio-visual recordings, while STAIRS26 provides high-resolution SAI annotations derived from 32-channel Eigenmike recordings. The objective is to reconstruct dense SAI representations, where sound events are described as dynamic masks encoding class, location, and energy intensity. For Track B, 4-channel audio and synchronized video frames are used to estimate the semantic acoustic maps from STAIRS26.

The multichannel audio signals are divided into 100 ms blocks. Within each block, a visibility matrix $\mathbf{S}_{t,f} \in \mathbb{C}^{4\times4}$ is computed via STFT using a 512-point FFT. Magnitude spectra are then pooled into nine uniformly spaced frequency bands from 1.5 kHz to 4.5 kHz. A pretrained UpLAM model estimates the spatial distribution of acoustic energy across these nine bands. For Track B, the resulting spherical activations are projected onto a 180×360 equirectangular projection (ERP) grid, forming a $9 \times 180 \times 360$ tensor. Finally, a horizontal flip is applied to align the acoustic azimuth

[1]STARSS23, Zenodo record 7880637.
[2]STAIRS26, Zenodo record 18171005.

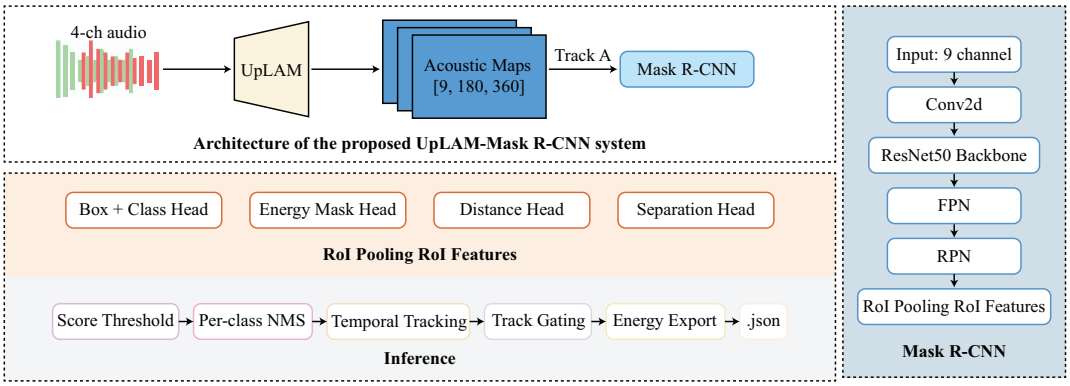

Figure 1: Architecture of the proposed UpLAM-Mask R-CNN system. The audio-only track uses UpLAM acoustic maps as input, while the audio-visual track concatenates acoustic maps with RGB frame features. The Mask R-CNN detector predicts sound event categories, spatial regions, energy maps, and source distances, followed by post-processing to generate the final JSON output.

with the camera perspective, followed by inter-band normalization to preserve relative spatial-frequency patterns while maintaining numerical stability.

2.2. Data Augmentation

Following the baseline [7], we implement a multi-dimensional augmentation suite for SAI features. Random cyclic shifts are applied along the horizontal axis of ERP maps to simulate rotation, with axis-aligned bounding boxes recomputed accordingly. For instances crossing the $360°/0°$ boundary, a splitting algorithm generates fragmented mask regions and removes invalid instances to ensure label accuracy. Inspired by SpecAugment [11], random band masking is performed by zeroing out 0 to 2 frequency bands, thereby reducing spectral overfitting. Additive Gaussian noise ($\sigma = 0.015$) is further applied to the acoustic maps to improve robustness against sensor perturbations.

The augmentation pipeline is modality-aware and supports both audio-only and audio-visual tasks. For Track B, following the baseline augmentation strategy [7], synchronized horizontal flips (prob $= 0.5$) are applied to acoustic images, visual frames, and spatial labels to maintain cross-modal spatial consistency. Visual-specific augmentations[3], including brightness, contrast, and saturation jittering, are also adopted to improve robustness to lighting variations and promote AV fusion.

3. TRACK A: AUDIO-ONLY INFERENCE

3.1. Network Architecture

Our system is built upon a customized UpLAM-Mask R-CNN framework. The model is designed as a multi-task instance detection network that predicts sound event categories, spatial regions, acoustic energy maps, and source distances from acoustic or audio-visual inputs. As shown in Fig. 1, the input is the $9 \times 180 \times 360$ acoustic ERP feature extracted by the UpLAM frontend. The detection backbone follows the Mask R-CNN framework and consists of a ResNet50 feature extractor with an FPN[4]. The ResNet50 backbone extracts hierarchical spatial features from the acoustic ERP maps. These features are then passed to the FPN, which constructs multi-scale feature maps denoted as $\{P_2, P_3, P_4, P_5\}$. Here, P_2 has the highest spatial resolution and is useful for capturing fine-grained acoustic energy peaks, while deeper pyramid levels such as P_4 and P_5 have lower spatial resolutions but stronger semantic and contextual representations. In this way, the FPN enables the detector to handle sound events with different spatial extents on the ERP plane.

Based on this architecture, we introduce several core enhancements over the baseline system to improve SAI performance:

- **RoI-based Instance Modeling:** Instead of direct full-frame prediction, we use a region proposal network (RPN) and RoIAlign to focus on local event regions. This enables precise instance-level prediction, which is critical for resolving scenes with multiple overlapping sound events.
- **Enhanced SAI Prediction Heads:** We introduce a task-specific RoI energy mask head with a resolution of 28×28 and a parallel distance regression head. The distance head adopts a Softplus activation to ensure physically valid, non-negative radial distance estimation.
- **Global-Local Supervision Strategy:** A full-frame energy decoder is integrated to provide global supervision during training. During inference, the system prioritizes localized RoI masks to prevent over-expansion of energy regions and maintain spatial stability.
- **Auxiliary Instance Separation:** An auxiliary projection branch maps RoI features into a 32-dimensional embedding space. This auxiliary representation is designed to improve instance discrimination in overlapping or multi-source scenarios.

During inference, the network outputs event categories, confidence scores, bounding boxes, RoI energy masks, full-frame energy maps, distance estimates, and optional instance embeddings. These predictions are further processed by score thresholding, per-class non-maximum suppression, temporal tracking, track gating, and region-aware energy-map export. The final output is converted into the official challenge JSON format, including the event category, confidence score, distance, and spatial acoustic energy region for each detected sound event.

3.2. Learning Strategy

We employ a multi-task loss framework in which energy-field reconstruction is supervised by weighted MSE losses, with weights of 10.0 and 5.0 assigned to the global and instance-level energy maps, respectively. A cosine margin loss is used for source separation, while source distances are regressed using Smooth L1 loss. To mitigate class imbalance, a dynamic balanced-sampling strategy is implemented, ensuring that 50% of the samples in each epoch are

[3]Torchvision ColorJitter.
[4]FasterRCNN ResNet50-FPN.

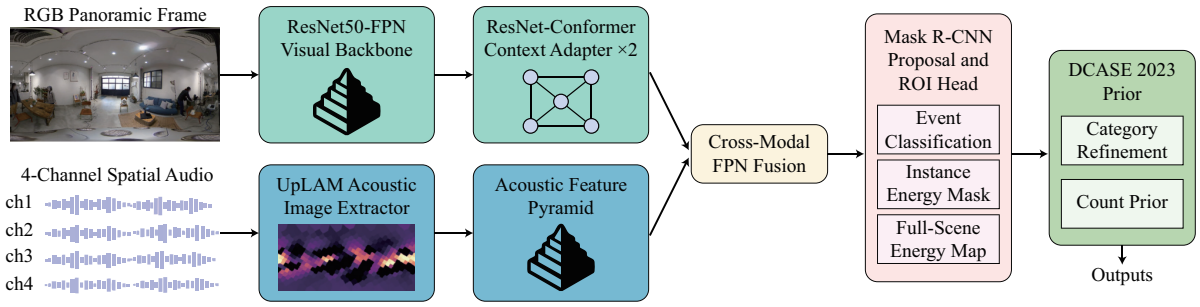

Figure 2: Network architecture of the proposed audio-visual instance SELD model for Track B.

drawn from rare categories.

The model is optimized using AdamW with per-layer learning-rate assignment. To promote stable adaptation, the task-specific heads and backbone stem use the base learning rate, whereas the pretrained FPN and deeper ResNet layers are assigned smaller learning-rate scales. A progressive unfreezing schedule activates these layers sequentially at epochs 2 and 10. Training stability is further improved through gradient clipping at 2.0 and automatic mixed precision (AMP). The learning rate is decayed by a factor of 0.5 using a ReduceLROnPlateau scheduler based on the 5-epoch smoothed validation loss.

During inference, a category-constrained IoU tracker performs temporal association using the Hungarian algorithm. Active tracks are retained for up to 5 frames to ensure trajectory continuity and suppress transient localization errors.

4. TRACK B: AUDIO-VISUAL INFERENCE

4.1. Audio-Visual Network Training

For Track B, we develop an AV-SELD system based on an instance-level localization framework, as illustrated in Fig. 2. Instead of directly regressing event-wise directions, the proposed model predicts sound-event instances on an equirectangular azimuth-elevation canvas. Each input frame contains RGB visual information and spatial acoustic representations. The acoustic features are derived from the multichannel microphone signals using UpLAM and projected onto equirectangular acoustic maps, such that the audio and visual modalities are represented in a shared spatial coordinate system.

The backbone of the proposed system follows a Mask R-CNN-style ResNet50-FPN architecture. The visual stream employs a pretrained ResNet50-FPN detector to extract robust frame-level visual representations. In parallel, the acoustic maps are processed by a lightweight acoustic feature pyramid, generating multi-scale audio features aligned with the FPN levels of the visual branch. To exploit the complementary information between the two modalities, we adopt cross-modal attention-based fusion. At the feature-pyramid level, the audio stream provides spatial guidance to the visual feature maps through gated residual fusion. This design preserves the stability of the pretrained visual detector while allowing acoustic cues to guide sound-source localization.

On top of this baseline architecture, we introduce several task-specific modules to improve sound-source localization and event prediction:

- **RoI-CMAF Cross-Modal Fusion:** We introduce a RoI-level cross-modal acoustic-visual fusion module to adaptively combine visual RoI features with acoustic spatial features. This enables each candidate event region to exploit both image context and sound localization cues.
- **ResNet-Conformer Context Adapter:** A lightweight ResNet-Conformer adapter is inserted into the feature extraction pipeline to enhance contextual modeling. The convolution branch captures local spatial patterns, while the self-attention branch improves long-range dependency modeling over panoramic acoustic-visual features.
- **Hull-based Energy Region Export:** During inference, sparse high-energy points are converted into compact energy regions using a hull-based post-processing strategy. This makes the exported JSON predictions more consistent with the official mask rendering and IoU computation.
- **DCASE 2023 Teacher Prior:** We employ a pretrained DCASE 2023 model as an external teacher prior for post-processing. Its predictions are used for category refinement, per-frame class-count constraints, and score re-ranking, improving the reliability of the final event predictions.

Specifically, the ResNet-Conformer adapter operates on compact FPN tokens to strengthen contextual modeling over the panoramic feature space. Its convolutional component captures local spatial patterns, while the self-attention component models long-range dependencies, which is beneficial for sound events with weak, partially visible, or spatially ambiguous visual evidence. In addition, RoI-level audio-visual fusion is performed after proposal generation. For each candidate region, visual and acoustic RoI features are integrated through a cross-modal attention module. The fused representation is primarily used for event classification, whereas the bounding-box regression branch remains visual-dominant. As a result, the acoustic stream serves as a conservative yet effective cue for category prediction and localization refinement.

During training, we optimize the RoI classification branch, the audio-visual fusion modules, the ResNet-Conformer adapter, and the energy-map prediction heads, while freezing the proposal network and most parameters of the pretrained visual backbone. The overall objective combines detection losses with spatial energy-map supervision. The detection terms optimize proposal classification and box/mask prediction, whereas the energy-map losses encourage compact sound-source mask estimation on the equirectangular canvas.

4.2. Model Ensemble and Post-processing

To improve the robustness of Track B inference, we combine the proposed audio-visual instance model with a pretrained DCASE 2023 AV-SELD teacher model. The audio-visual instance model generates sound-source proposals, masks, spatial locations, and confidence scores, whereas the DCASE 2023 model serves as a same-input inference-time prior for category and event-count refinement. For each detected instance, we associate the peak location

Table 1: Development-set results of the proposed systems for Track A.

System	Macro				Micro		Detections
	mAP	AP25	AP50	AP75	mAP	AP50	
FAO	0.0324	0.0721	0.0248	0.0004	0.0222	0.0190	265049
OR-AO	0.0286	0.0598	0.0255	0.0006	0.0220	0.0202	196036

Table 2: Detailed development-set metrics of the proposed clean Track B systems.

System	**Macro**					**Micro**				
	mAP↑	AP25↑	AP50↑	AP75↑	PearsonR↑	mAP↑	AP25↑	AP50↑	AP75↑	PearsonR↑
Baseline	0.0003	0.0010	0.0000	0.0000	0.1358	0.0011	0.0032	0.0000	0.0000	0.4429
TPAV	0.0789	0.1516	0.0838	0.0014	0.5366	**0.0741**	**0.1365**	**0.0845**	0.0013	0.5828
ERAV	0.0718	0.1375	0.0765	0.0012	0.5315	0.0709	0.1322	0.0794	0.0011	0.5826
CFAV	**0.1508**	**0.2525**	**0.1880**	**0.0118**	**0.5948**	0.0417	0.0645	0.0508	**0.0099**	**0.5948**

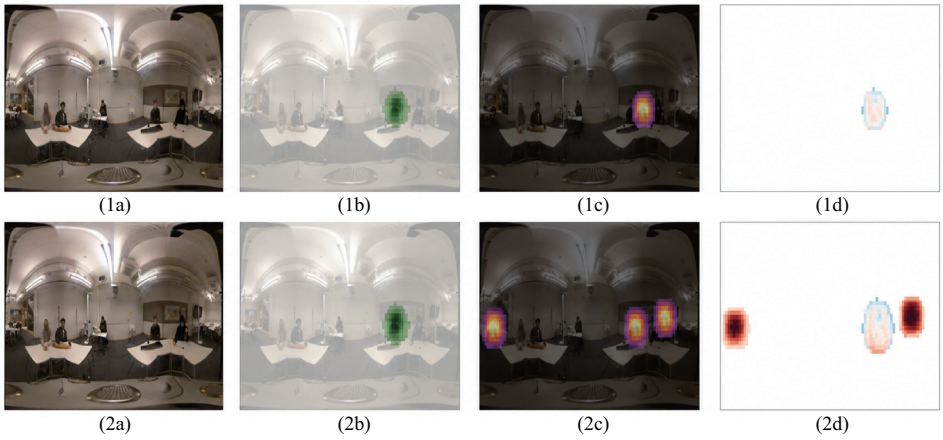

Figure 3: Demo of processing result of Track B. Rows show our model (1) and the baseline (2); columns from (a) to (d) respectively show RGB, ground truth, prediction, and residual.

of its predicted mask with the nearest teacher output in the same frame according to angular distance. When a reliable match is identified, the spatial mask is preserved, while the event label is replaced by the corresponding teacher-predicted category.

During post-processing, we further impose a teacher-count prior to constrain the number of retained detections for each frame and class. This procedure reduces duplicated predictions and improves category consistency. Overall, the strategy combines the fine-grained spatial localization capability of the audio-visual instance model with the more stable category estimates provided by the DCASE 2023 teacher, leading to improved inference robustness.

5. RESULTS ON DEVELOPMENT DATASET

5.1. Results on Track A

We evaluate the proposed Track A **audio-only (AO)** systems on the development set under the official evaluation protocol. The reported metrics include macro mAP, AP at different localization thresholds, micro mAP, micro AP50, and the total number of detections. As shown in Table 1, we compare two audio-only systems: **Final Audio-Only (FAO)** and **Objectness-ranked AO (OR-AO)**.

FAO is our submitted system for Track A. It employs light acoustic augmentation and disables the auxiliary classification loss, allowing the model to focus on instance-level detection, RoI-level energy-mask prediction, full-frame energy supervision, and distance regression. This configuration achieves the best overall macro mAP of 0.0324 and macro AP25 of 0.0721, and is therefore selected as the final Track A submission.

OR-AO is a score-calibration variant that ranks candidate detections using class-agnostic detector objectness scores while retaining the category predictions from FAO. Compared with FAO, OR-AO improves macro AP50 from 0.0248 to 0.0255 and micro AP50 from 0.0190 to 0.0202. It also reduces the number of detections from 265,049 to 196,036, suggesting that objectness-based ranking can suppress redundant predictions and improve precision under stricter localization thresholds.

However, OR-AO decreases the overall macro mAP from 0.0324 to 0.0286 and macro AP25 from 0.0721 to 0.0598. This indicates that the more conservative ranking strategy also removes useful detections under relaxed localization criteria. Therefore, OR-AO is treated as a precision-oriented analysis variant, whereas FAO remains the submitted Track A system.

We additionally conducted several diagnostic experiments, including a ResNet-Conformer variant, class-wise score selection, independent router-based classification, and coverage-based refinement. The ResNet-Conformer configuration did not yield consistent improvements on the development set. The remaining experiments indicate that score calibration can be beneficial but is sensitive to the ranking criterion, while simple router decoupling or per-class coverage refinement is insufficient for reliable category assignment.

Overall, FAO provides the best balance between detection coverage and localization accuracy under the official macro mAP criterion. Although objectness-based ranking improves AP50 and reduces redundant detections, its lower macro mAP suggests that con-

fidence calibration must be carefully balanced against recall.

5.2. Results on Track B

We evaluate the proposed Track B audio-visual systems on the development set, using the official **Baseline** as a reference. The baseline adopts an early-fusion strategy, where the 4-channel microphone signals are first transformed by UpLAM into a 9-band acoustic ERP representation. The resulting acoustic maps are then concatenated with the RGB frame, forming a 12-channel input to a modified Mask R-CNN-style instance segmentation model.

In contrast, our system employs a more structured audio-visual fusion pipeline. Specifically, the RGB and acoustic inputs are processed by modality-specific feature extractors, and the acoustic representation is further projected into a feature pyramid. The acoustic and visual pyramids are subsequently integrated through cross-modal FPN fusion, followed by a ResNet-Conformer context adapter and RoI-level audio-visual fusion for event classification, localization, and semantic acoustic energy prediction. This design avoids imposing heterogeneous modalities on the same early convolutional filters and enables more explicit cross-modal interaction at both feature-pyramid and instance levels.

For clarity, the submitted systems are denoted as **Teacher-Prior AV (TPAV)**, **Energy-Ranked AV (ERAV)**, and **Class-Focused AV (CFAV)**. TPAV applies DCASE 2023 teacher-prior calibration to the proposed audio-visual model. ERAV further introduces an energy-based tie-breaking rule for detections with identical confidence scores. CFAV adopts a conservative category-level filtering strategy, retaining only the most reliable category during the final prediction stage.

To illustrate the spatial quality of the predicted semantic acoustic energy maps, we visualize a representative development-set frame generated by TPAV. As shown in Fig. 3, the proposed ResNet-enhanced FPN model produces a compact energy response around the target sound source, whereas the baseline yields multiple dispersed high-energy regions. The residual maps further indicate that our prediction is more consistent with the ground-truth acoustic annotation. These observations suggest that the improved feature extraction and fusion architecture can effectively suppress false-positive acoustic regions and enhance spatial alignment.

Table 2 reports detailed metrics for the three clean variants of our system, including both macro- and micro-averaged results. Macro metrics evaluate the average performance across categories and are therefore more sensitive to class-level precision. In contrast, micro metrics aggregate detections over all evaluated instances, reflecting the overall instance-level coverage of the system.

The detailed results reveal distinct effects of the two post-processing variants. Compared with TPAV, ERAV slightly reduces both macro mAP and micro mAP, from 0.0789 to 0.0718 and from 0.0741 to 0.0709, respectively. This degradation suggests that, after teacher-prior calibration, many matched detections already have saturated confidence scores, and the additional energy-based tie-breaking rule does not provide a reliable global ranking cue.

CFAV exhibits a different trend. It improves macro mAP from 0.0789 to 0.1508 and macro AP50 from 0.0838 to 0.1880, indicating that category-level filtering can effectively suppress false-positive regions and improve precision for the selected reliable category. However, its micro mAP decreases from 0.0741 to 0.0417, since most other categories are removed during the final prediction stage. Therefore, CFAV should be regarded as a precision-oriented variant rather than a balanced multi-class solution.

Overall, the Track B results show that system performance is strongly influenced by calibration and category-level post-processing. Among the clean systems, TPAV achieves the best micro mAP and micro AP50, suggesting stronger multi-class coverage. In contrast, CFAV obtains the best macro metrics, demonstrating that class-focused filtering can improve category-level precision at the expense of overall coverage. The legacy post-processed AV variant achieves the highest numerical scores in the compact comparison, particularly in terms of micro mAP. Nevertheless, since it does not follow the final clean protocol, it is reported only as a non-clean diagnostic reference.

6. REFERENCES

[1] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 34–48, 2019.

[2] Archontis Politis, Annamaria Mesaros, Sharath Adavanne, Toni Heittola, and Tuomas Virtanen, "Overview and evaluation of sound event localization and detection in DCASE 2019," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 684–698, 2021.

[3] Yin Cao, Turab Iqbal, Qiuqiang Kong, Fengyan An, Wenwu Wang, and Mark D. Plumbley, "An improved event-independent network for polyphonic sound event localization and detection," in *Proc. ICASSP*, 2021, pp. 885–889.

[4] Chia-Chuan Liu, Chia-Ping Chen, Chung-Li Lu, Bo-Cheng Chan, Yu-Han Cheng, Hsiang-Feng Chuang, and Wei-Yu Chen, "Regression-based sound event detection with semi-supervised learning," in *Proc. APSIPA ASC*, 2023, pp. 2336–2342.

[5] David Diaz-Guerra Aparicio, Archontis Politis, Parthasaarathy Ariyakulam Sudarsanam, Kazuki Shimada, Daniel Krause, Kengo Uchida, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Takashi Shibuya, et al., "Baseline models and evaluation of sound event localization and detection with distance estimation in DCASE 2024 Challenge," in *Proc. Workshop DCASE*, 2024, pp. 41–45.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[7] Adrian S. Roman, Iran R. Roman, and Juan P. Bello, "Latent acoustic mapping for direction of arrival estimation: A self-supervised approach," in *Proc. WASPAA*, 2025, pp. 1–5.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[9] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *Proc. CVPR*, 2017, pp. 936–944.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, 2017, vol. 30.

[11] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: a simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.