applied system innovation

MDPI

Systematic Review

Real-Time Sound Event Localization and Detection for Edge and Mobile Devices: A Systematic Review

Phathorn Thammasorn and Porawat Visutsak *

Department of Computer and Information Science, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand; s6904062910013@kmutnb.ac.th
* Correspondence: porawatv@kmutnb.ac.th

Abstract

Background: Real-time Sound Event Localization and Detection (SELD) is important for enabling spatial-audio perception in resource-constrained environments. Edge and mobile platforms—including smartphones, wearables, embedded processors, microcontrollers, and IoT devices—impose strict constraints on latency, memory, and power that challenge current SELD methods. Objective: This review evaluates the readiness of current SELD methods for real-time deployment on edge and mobile devices and identifies the principal limitations affecting deployment. Methods: This systematic review was reported in accordance with the PRISMA 2020 statement. A structured search conducted on 30 June 2026 using SciSpace, SciSpace Full Text, Google Scholar, and arXiv retrieved 1078 records before deduplication, resulting in 611 unique records. Eligible studies were English-language publications from January 2018 to 30 June 2026 addressing SELD or directly relevant edge/mobile acoustic perception and deployment, with two foundational pre-2018 exceptions. Study selection and data extraction were performed by the first author, with automated tools used only to assist prioritization and information extraction. No formal risk-of-bias tool was applied. Because of methodological heterogeneity, results were synthesized qualitatively. Following screening and eligibility assessment, 49 publications were included. Results: Significant progress has been made in developing compact systems capable of performing SELD in real time; however, many limitations persist. Trade-offs between model complexity and localization accuracy remain unresolved. Sensitivity to microphone geometry and channel count is a persistent constraint. Robustness under noisy and reverberant conditions has improved, but current systems remain limited in practical deployment. A gap persists between performance on SELD benchmark datasets and real-world deployments. The available evidence was heterogeneous, and deployment-related metrics were reported inconsistently across studies. Conclusions: Several directions are identified that may help close these gaps, including hardware-aware Neural Architecture Search (NAS), self-supervised spatial-audio representation learning, on-device adaptation/personalization, energy-aware duty cycling, multimodal integration, cooperative device fusion, microphone calibration, and privacy-preserving trustworthy edge SELD. This systematic review provides an organized synthesis of the current state of the art in edge/mobile SELD, its limitations, and future opportunities for real-time implementation. The review was not prospectively registered and received no external funding.

Keywords: sound event localization and detection (SELD); edge computing; real-time processing; lightweight deep learning; spatial audio perception

Check for updates

Academic Editor: Daswin De Silva

Received: 23 July 2026
Revised: 10 September 2026
Accepted: 21 September 2026
Published: 23 September 2026

Copyright: © 2026 by the authors. Published by MDPI on behalf of the International Institute of Knowledge Innovation and Invention. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

There are many different types of edge computing devices, mobile platforms, and Internet of Things (IoT) devices being used today, which creates new ways for people to utilize their environment through spatial audio intelligence. Sound Event Localization and Detection (SELD) is primarily concerned with identifying which sound events occur and locating them within three-dimensional space. Traditionally, SELD has been researched using computers with significant amounts of computing power in controlled laboratory settings; however, there is an increasing desire to perform spatial audio intelligence tasks with low latency on resource-constrained devices such as smartphones, wearable audio devices, embedded systems, and edge computing platforms.

There have been attempts to implement SELD in various areas such as assistive technology, smart homes, and IoT. Assistive technology includes assistive listening applications that use spatial audio intelligence to provide hearing-impaired individuals with information about the surrounding environment and to identify and locate important environmental sounds relative to the individual [1–3]. Edge-based acoustic monitoring in smart homes and IoT applications may include low-power detection of specific sound events on microcontrollers [4], cooperative abnormal sound detection for city surveillance [5], and recognition of environmental activities using ambient sounds for elderly care and smart-home automation [6]. Extended Reality (XR) applications also require rapid processing of spatial audio to generate an immersive experience [7]. Acoustic sensing has become increasingly important for automotive safety systems due to its ability to rapidly detect emergency vehicles and hazardous situations [8]. However, cloud-based processing introduces latency, security concerns, and dependency issues associated with network connectivity, which makes it necessary to perform spatial audio intelligence tasks locally.

Performing SELD on mobile and embedded platforms presents several challenges. Although modern deep learning-based SELD systems have demonstrated high performance levels on standard benchmark datasets, they generally require large computational resources and consume excessive amounts of memory bandwidth and energy, which can exceed the capabilities of typical mobile and embedded platforms. In addition, real-time SELD requires sufficiently low latency for the intended application while continuously performing sound event detection and localization. Furthermore, edge devices often possess significantly less complex microphone arrays than the dense arrays utilized in research studies, resulting in lower spatial resolution and accuracy when localizing sound events. Real-world deployments introduce additional challenges such as background noise, reverberation, simultaneous sound events, and domain shifts.

This review evaluates recent developments in real-time SELD for edge and mobile devices. There are two main research questions: How prepared are current SELD methods for real-time deployment on edge and mobile devices? What technological limitations exist with respect to latency, computational cost, microphone constraints, and robustness in real-world deployments? In addition to assessing these advances in terms of efficiency, we also assess recent developments in lightweight model architectures, efficient audio-feature representations, and deployment strategies. We evaluate the distinction between laboratory testing and actual real-world deployment requirements.

The scope of this evaluation covers SELD methodologies that were either designed to be implemented on constrained resources such as mobile phones, hearables/audio wearables (e.g., earbuds), embedded systems (e.g., NVIDIA Jetson platforms (NVIDIA Corporation, Santa Clara, CA, USA) and Raspberry Pi computers (Raspberry Pi Ltd., Cambridge, UK)), microcontrollers (MCUs), and edge computing devices, or are directly relevant to these platforms. These methodologies need to meet the unique constraints of each type of platform, including low-latency inference processing, small memory footprints,

limited microphone-array complexity, and robustness to the wide variety of acoustic environments found in the real world. The scope of this evaluation does not include broad SELD research that has no relevance to the deployment of SELD on mobile or embedded platforms; non-SELD-based sound classification; pure spatial audio generation; music classification; or voice-only systems, unless they directly relate to developing a SELD system for mobile or embedded platforms.

Section 2 provides background on SELD task formulations, evaluation metrics, audio formats, and different families of SELD models. Section 3 examines specific lightweight SELD model architectures, specifically, compact Convolutional Neural Networks (CNNs), Temporal Convolutional Networks (TCNs), and efficient Transformers. Section 4 addresses both the benefits and drawbacks of using different audio-feature representations, and Section 5 examines the challenges of deploying a SELD system into a variety of environments, including latency requirements, the need to reduce computational costs, and limitations imposed by microphones and hardware. Section 6 focuses on robustness under real-world conditions, including noise, reverberation, and shifts in the underlying environment. Section 7 reviews current applications of SELD models in areas such as environmental awareness, assistive devices, smart homes/IoT, and VR/XR systems. Section 8 reviews available datasets and evaluation protocols. Section 9 provides a comparative synthesis of the reviewed systems, Section 10 discusses future research directions, and Section 11 concludes the review.

Systematic Review Protocol and Study Selection

This systematic review was reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement [9]. The review objective was to identify and synthesize the current body of evidence on real-time SELD for edge and mobile deployment, addressing: (i) model architectures for resource-constrained inference; (ii) audio-feature representations for efficient processing; (iii) deployment constraints related to latency, memory, and hardware; (iv) microphone-array limitations; and (v) robustness under real-world acoustic conditions. The review was not prospectively registered, and no separate review protocol was prepared or published.

Literature searches were performed on 30 June 2026 using SciSpace, SciSpace Full Text, Google Scholar, and arXiv. SciSpace and its Full Text search mode were accessed through https://scispace.com, Google Scholar through https://scholar.google.com, and arXiv through https://arxiv.org. The searches were based on the predefined review scope covering real-time SELD, edge and mobile deployment, lightweight models, efficient audio features, microphone-array localization, computational constraints, and robustness. The searches were organized into three thematic search sets: (1) real-time SELD for edge and mobile devices and lightweight models; (2) efficient audio features and low-latency SELD for resource-constrained hardware; and (3) SELD datasets, benchmarks, DCASE/FOA evaluation, and deployment challenges. SciSpace, SciSpace Full Text, Google Scholar, and arXiv were queried using source-specific natural-language, keyword, or Boolean search expressions. The exact search expressions and numbers of records returned by each query are provided in Table 1. Across the 18 search executions, 1078 records were retrieved before deduplication. After removal of 467 duplicate occurrences, 611 unique records remained for screening and prioritization.

The primary systematic eligibility window was January 2018 to 30 June 2026. Two earlier publications were retained as explicit exceptions because they provided foundational technical evidence directly relevant to the development of real-time sound-source localization on mobile and embedded hardware. These studies were included in the synthesis where their contributions remained directly relevant to the review objectives.

Table 1. Search strategy and number of records retrieved on 30 June 2026.

Search Set	Source	Exact Search Expression	Records
1	SciSpace	What are the recent advances in real-time Sound Event Localization and Detection for edge and mobile devices using lightweight models?	100
1	SciSpace Full Text	real-time Sound Event Localization Detection SELD edge mobile devices lightweight models microphone array	100
1	Google Scholar	((“Sound Event Localization” OR “SELD” OR “sound source localization”) AND (“edge device” OR “mobile device” OR “embedded system” OR “lightweight model” OR “model compression”) AND (“microphone array” OR “audio processing” OR “real-time”))	20
1	arXiv	all: sound + event + localization + detection + AND + (mobile + OR + edge + OR + lightweight + OR + embedded) + AND + (microphone + array + OR + real-time)	20
2	SciSpace	What are efficient audio features and methods for low-latency sound event localization and detection on embedded systems and resource-constrained hardware?	100
2	SciSpace Full Text	efficient audio features low-latency SELD embedded systems resource-constrained hardware real-time inference	100
2	Google Scholar	((“sound event localization” OR “SELD” OR “sound event detection”) AND (“low latency” OR “real-time” OR “efficient”) AND (“embedded systems” OR “resource-constrained” OR “edge computing” OR “wearable”))	19
2	arXiv	all: sound + event + localization + detection + AND + (embedded + OR + edge + OR + efficient + OR + low-latency + OR + real-time)	20
3	SciSpace	What are the current datasets, benchmarks, and evaluation methods for sound event localization and detection (SELD) including first-order ambisonics and DCASE challenge?	100
3	SciSpace	What are the deployment challenges and resource constraints for implementing sound event localization and detection systems on embedded and resource-limited hardware?	100
3	SciSpace	What are the recent advances in SELD systems using first-order ambisonics for DCASE challenge tasks?	100
3	SciSpace Full Text	sound event localization detection SELD datasets benchmarks DCASE challenge first-order ambisonics FOA	100
3	SciSpace Full Text	SELD deployment embedded systems resource-limited hardware edge computing real-time audio processing	100
3	Google Scholar	(SELD OR “sound event localization” OR “sound event detection”) AND (dataset OR benchmark OR evaluation) AND (DCASE OR “first-order ambisonics” OR FOA)	20
3	Google Scholar	(SELD OR “sound event localization and detection”) AND (deployment OR embedded OR “resource-limited” OR “edge computing” OR “real-time”) AND hardware	19
3	Google Scholar	“DCASE challenge” AND (“sound event localization” OR SELD) AND (“ambisonics” OR “spatial audio” OR “microphone array”)	20
3	arXiv	all: SELD + OR + all: sound + event + localization + detection + AND + all: DCASE + OR + all: ambisonics + OR + all: benchmark	20
3	arXiv	all: sound + event + localization + AND + all: deployment + OR + all: embedded + OR + all: resource + constrained + OR + all: edge + computing	20
Total			1078

English-language studies published within the primary eligibility window, together with the two foundational pre-2018 studies described above, were eligible for inclusion if they addressed at least one of the following: SELD model architectures and output representations; spatial audio feature-extraction methods; real-time or edge/mobile deployment of acoustic perception systems; microphone-array design constraints; robustness under non-ideal acoustic conditions; SELD datasets and evaluation protocols; or closely related edge acoustic sensing tasks (sound event detection, sound-source localization, speech enhancement, and audio−visual localization) with direct relevance to edge/mobile deployment. Studies were excluded if they were: (i) unrelated to acoustic event detection or spatial-audio localization; (ii) focused only on general SELD or related acoustic tasks with no discernible application to resource-constrained, real-time, or mobile settings; (iii) limited to non-SELD tasks such as music classification, pure spatial audio generation, or voice-only systems unrelated to SELD; or (iv) lacking sufficient technical detail to allow meaningful comparison.

For PRISMA reporting, the study-selection counts were reconstructed during revision from the archived search outputs and manuscript-version history. The 611 unique records constituted the screening pool. The first author screened the candidate records by reviewing their abstracts and made the study-selection decisions. The first author also assessed all 52 retrieved reports for eligibility. SciSpace relevance scores and automated reranking were used only to prioritize records for assessment and did not determine inclusion or exclusion; no fixed numerical relevance-score threshold was documented.

Data extraction was performed by the first author with assistance from SciSpace and generative AI tools to identify and summarize relevant technical information. Extracted details were verified by the first author against the original source publications before inclusion in the manuscript; no independent duplicate data extraction was performed.

When a deployment metric or technical characteristic was not reported in the source publication, it was treated as not reported rather than inferred.

No formal study-level risk-of-bias assessment tool was applied because the included literature comprised heterogeneous engineering, benchmark, and system-deployment studies with substantially different experimental designs and evaluation protocols. Methodological and reporting limitations relevant to interpretation were instead considered qualitatively during the synthesis, including whether evaluation was performed on target hardware, whether complete end-to-end deployment metrics were reported, and whether results were obtained from benchmark-only or real-world deployment settings.

Because the included studies used heterogeneous task definitions and evaluation protocols, no common effect measure was applied across studies. Performance and deployment outcomes were synthesized using the measures reported in the original studies, including F-score, Error Rate, Localization Error, Localization Recall, SELD score, inference latency, Real-Time Factor, model size, FLOPs/MACs, memory use, and power consumption where available; no statistical transformation to a common effect size was performed.

Included studies were grouped for synthesis according to their primary relevance to edge/mobile SELD, including model architecture, audio-feature representation, deployment hardware and constraints, microphone configuration, robustness, and evaluation methodology. Results were summarized narratively and in structured comparative tables; no meta-analysis, subgroup analysis, meta-regression, or sensitivity analysis was performed because of the substantial heterogeneity in task definitions, datasets, hardware platforms, and reported metrics.

No formal assessment of reporting bias or publication bias was performed because the review did not include a quantitative meta-analysis and the included studies were highly heterogeneous in their tasks, metrics, and deployment settings. Potential bias arising

from incomplete reporting was considered qualitatively, particularly when studies did not report deployment-relevant measures such as latency, memory use, power consumption, or target-hardware performance.

No formal certainty-of-evidence framework was applied because the included studies were heterogeneous in study design, task formulation, datasets, hardware platforms, and reported outcomes. The strength of deployment evidence was instead considered qualitatively during synthesis, but this was not treated as a formal certainty assessment.

Three reports were excluded at the eligibility stage: Nakadai et al. (2017) [10], Development of Microphone-Array-Embedded UAV for Search and Rescue Task, fell outside the primary publication window and was not retained as a foundational exception; Han et al. (2025) [11], LISTEN: Lightweight Industrial Sound-representable Transformer for Edge Notification, addressed industrial acoustic monitoring without localization; and Wu et al. (2025) [12], Sound-Source Localization Based on Three-Element Microphone Array and RepMobileViT Model, was a localization-only study that did not provide sufficient evidence directly relevant to edge/mobile SELD deployment. The remaining 49 publications were included in the systematic synthesis.

Detailed deployment-oriented characteristics of the included SELD, SELDDE, SELD-like, and related edge/mobile acoustic system studies are provided in Supplementary Tables S1 and S2. Four included publications used primarily for contextual review, benchmark/evaluation, or dataset information [7,13–15] are discussed in the main text rather than tabulated as deployment-system entries. The overall study-selection process is summarized in Figure 1.

For included studies, available information was extracted on study identification (authors, year, venue), system architecture, audio input format, microphone configuration, hardware deployment target, reported performance and deployment metrics (including, where available, detection F-score or Error Rate, Localization Error, inference latency, Real-Time Factor, model size, and FLOPs/MACs), and key findings relevant to edge/mobile deployment. Due to high heterogeneity in task formulation, hardware platforms, and metric reporting, a formal quantitative meta-analysis was not feasible; synthesis was conducted as a structured qualitative narrative organized by deployment-relevant technical category.

Three additional review articles were subsequently added in response to reviewer comments for comparison with prior reviews; these were not part of the original systematic study-selection process.

Several reviews in the literature have examined SELD and acoustic perception tasks related to SELD. However, the primary focus of those reviews differs from that of the present review. Some prior reviews focus on benchmark design and evaluation methodology, some focus on general SELD models and datasets, and others focus on broader sound-source localization or polyphonic sound event detection/localization. Therefore, Table 2 summarizes how the present review relates to these closely related reviews.

As shown in Table 2, the present review is distinguished by its specific focus on the deployment readiness of SELD systems for edge and mobile devices. Prior reviews have addressed SELD benchmarks, model architectures, datasets, spatial features, and broader localization and detection methods, while the present review concentrates on practical deployment issues including latency, real-time factor, feature-extraction cost, model size, hardware platforms, microphone-array configuration, robustness, and end-to-end evaluation.

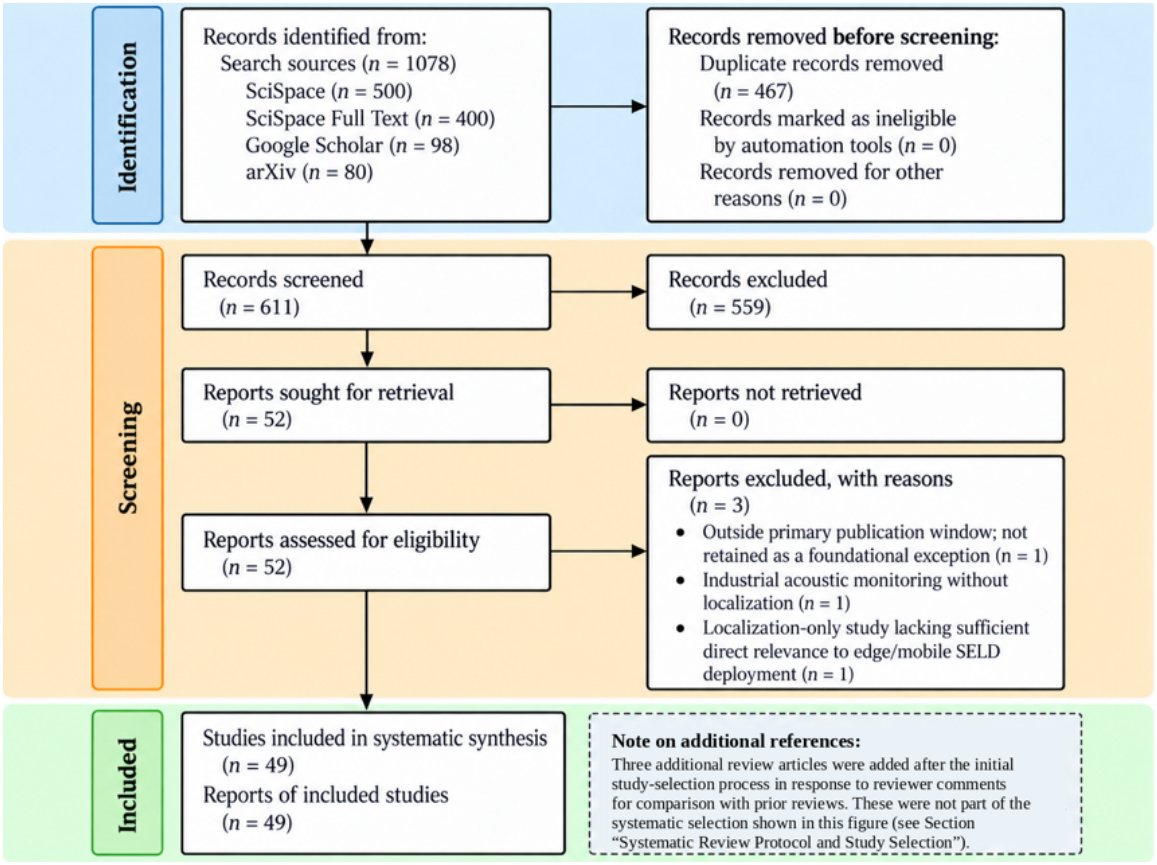


Figure 1. PRISMA 2020 flow diagram of the literature search and study-selection process [9].

Table 2. Comparison between the present review and related review articles.

Review	Main Focus	Difference from the Present Review
Politis et al. [13]	Overview of the DCASE 2019 SELD Challenge, including dataset creation, evaluation methodology, common input features, model architectures, training approaches, and joint SELD evaluation metrics.	Focuses on the DCASE 2019 benchmark and challenge submissions rather than edge/mobile deployment readiness, hardware measurements, or end-to-end real-time evaluation.
Yeow et al. [16]	Modern review of SELD for environmental acoustic intelligence, including model architectures, datasets, spatial features, data augmentation methods, output formats, and future directions.	Discusses deployable SELD systems and edge-related challenges, but the present review provides a more focused discussion of real-time edge/mobile deployment requirements, including latency, real-time factor, feature-extraction cost, hardware platforms, and microphone-array constraints.
Mohmmad and Sanampudi [17]	Parametric survey of polyphonic sound event detection and localization, including datasets, feature extraction, and classification/localization methods.	Focuses mainly on algorithmic representations, classification/localization methods, and the state of the art for polyphonic sound event detection and localization. In contrast, the present review focuses on the practical limitations involved in deploying complete SELD systems on edge/mobile hardware.
Jekateryńczuk and Piotrowski [18]	Broad survey of sound-source localization and detection methods, including classical propagation-based methods, machine learning methods, deep learning methods, and applications.	Provides a broad overview of sound-source detection and localization methods but does not specifically focus on evaluating the feasibility of deploying complete SELD systems on edge/mobile devices.
Present review	Real-time SELD for edge/mobile devices.	Focuses on deployment feasibility for edge/mobile devices, including real-time performance constraints such as latency, real-time factor, end-to-end processing time, feature-extraction cost, model size, hardware platforms, microphone-array constraints, and robustness of deployed systems.

2. Background

Figure 2 illustrates the general processing pipeline of a real-time edge/mobile SELD system, from multichannel audio acquisition and feature extraction to SELD inference and output generation.

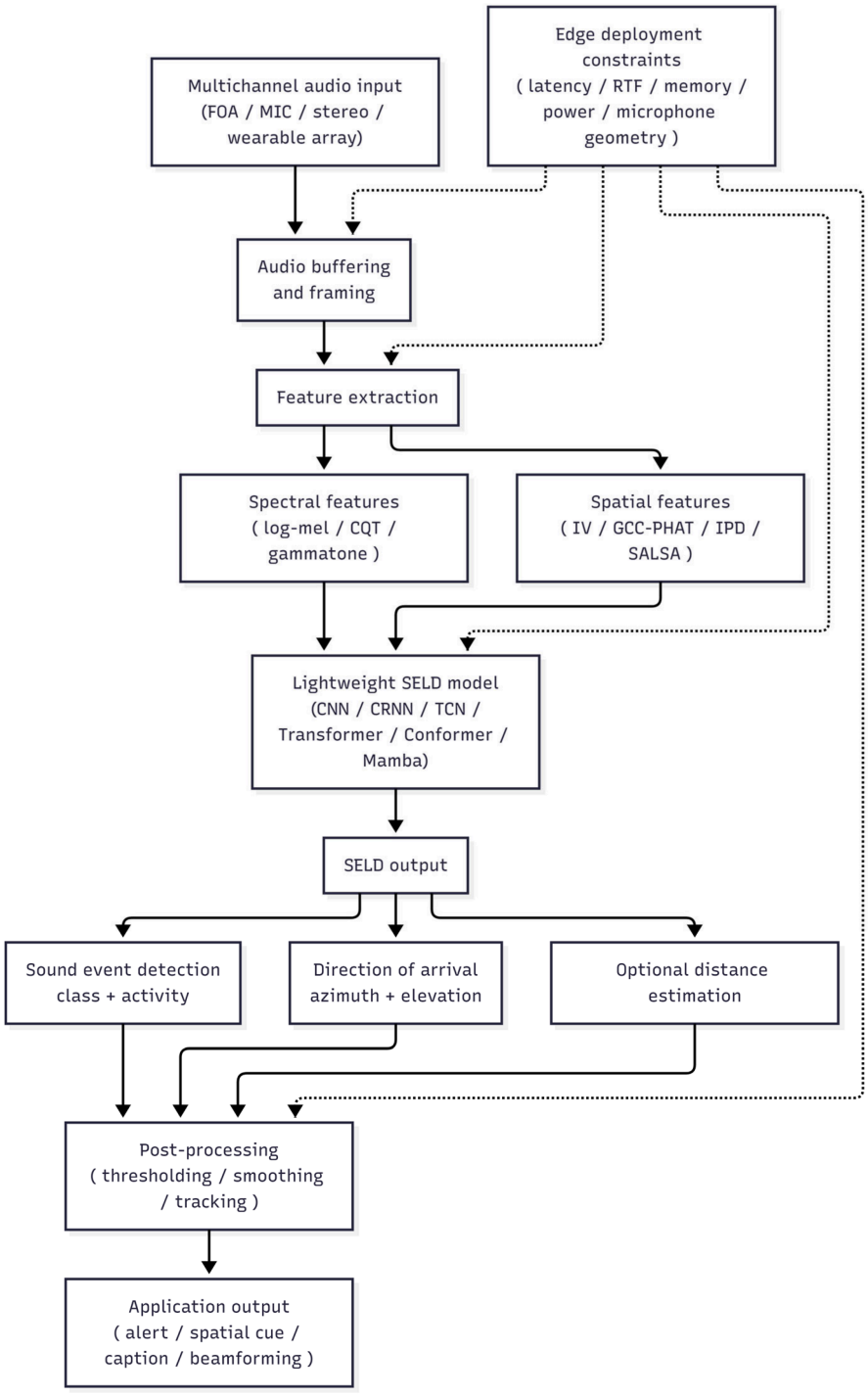


Figure 2. General pipeline of real-time SELD for edge and mobile devices. Multichannel audio is buffered and converted into spectral and spatial features, which are processed by a lightweight SELD model to estimate sound event activity, direction of arrival, and optional source distance. Edge deployment constraints affect multiple stages of the pipeline, including buffering, feature extraction, inference, post-processing, latency, memory use, power consumption, and microphone geometry. Solid arrows indicate the main processing flow, whereas dotted arrows indicate the influence of edge-deployment constraints on the corresponding processing stages [1,2,19–23].

2.1. SELD Task Formulation

SELD stands for Sound Event Localization and Detection, which combines two tasks: (i) Sound Event Detection (SED), identifying whether sound event classes are active in an acoustic environment, and (ii) Direction of Arrival (DOA) estimation, determining where these sound events occur in space using azimuth and elevation angles. A SELD system processes a multichannel audio signal to output frame-by-frame information about whether each sound event class is active and where those active sound events are occurring in space.

In polyphonic conditions, where several audio objects from the same class or different classes occur at once and may overlap in time, the SELD task is very difficult because the model has to identify multiple concurrent events and correctly assign each event to its spatial location. As more overlapping audio sources are added, this problem becomes much harder [23,24]. More recently, researchers have extended traditional SELD by adding source-distance estimation. Source-distance estimation allows the model to perform 3D source localization based on both the DOA and the distance of the source relative to the sensor array [22].

The SELD problem can be addressed using different output formats. Many early CRNN-based approaches to SELD were developed with separate branches for estimating SED and DOA [23]. Some later DCASE 2019 submissions continued this trend by using separate networks and/or two-stage pipelines for SED and localization [13]. Shimada et al. proposed Activity-Coupled Cartesian Direction of Arrival (ACCDOA), which represents event activity and direction in one Cartesian coordinate space by coupling the magnitude of a Cartesian DOA vector with event activity [25]. As reported in ref. [1], ACCDOA has also been successfully applied to lightweight SELD solutions for wearable audio devices. A multi-ACCDOA approach extended the previously reported single-ACCDOA representation by allowing multiple ACCDOA vectors for each class, thus enabling prediction of the spatial locations of overlapping same-class events [24].

2.2. Evaluation Metrics

Standardized evaluation metrics have been developed through the DCASE (Detection and Classification of Acoustic Scenes and Events) Challenge [13] for SELD systems. These metrics evaluate both the detection and localization performance of a SELD system separately before combining the two into an overall SELD score.

DCASE uses these metrics for evaluating Sound Event Detection [13]:

1. Error Rate (ER): ER measures the number of detection errors relative to the number of ground-truth events. It is usually calculated from substitutions, deletions, and insertions, then normalized by the total number of reference events.
2. F-Score (F1): The F-Score measures event-detection performance using the relationship between precision and recall. Precision represents how accurate the detections are, while recall represents how complete the detections are.

These metrics are used to evaluate localization performance [13]:

3. Localization Error (LE): LE calculates the average angular error in degrees between predicted DOAs and ground-truth DOAs for all sound events that were correctly detected.
4. Localization Recall (LR): LR measures the number of sound events or sources that were correctly localized out of the total number of active sound events or sources in the ground truth.

In more recent DCASE-style evaluations, there are also “location-dependent” versions of these metrics, such as ER20° and F20°, which only consider a prediction to be correct if its corresponding event is located within 20° of the ground-truth event [26,27].

Performance metrics beyond those mentioned above, such as latency, inference time or Real-Time Factor (*RTF*), model size, computational cost (i.e., FLOPs and MACs), and power consumption, are critical for edge or mobile deployment:

1. Inference Time: The time required for a neural network to complete a forward pass after all input features have been fed into the network. This excludes the time required to capture audio, buffer audio, extract audio features, and perform any audio post-processing operations [28].
2. Real-Time Factor (*RTF*): This is calculated by dividing the amount of processing time required by the length of the audio track. If the *RTF* is less than 1.0 then the processing took place before the end of the audio track during the test conditions. However, an *RTF* value less than 1.0 does not indicate that the system has low latency, can produce results in real time (causal), or will be able to continuously stream and process new audio at deployment (See Section 5.1) [20].
3. Model Size: Counting the number of parameters in a model, along with its memory requirements, is often necessary to determine whether models can fit onto an edge or mobile device [19,29].
4. Computational Cost: Computational cost measures how much computation is needed for extracting features from audio data or performing inference, expressed in FLOPs or MACs [1,30,31].
5. Power Consumption: Power consumption assesses how much energy is consumed during inference or continuous operation. This is especially important for battery-powered and low-power edge devices [3,32].
6. End-to-End Latency: End-to-end latency measures how long it takes from audio input to full SELD output. It includes processing time, buffering, feature extraction, inference, and post-processing [20].

2.3. Audio Formats and Microphone Arrays

Systems based upon SELD technology use multichannel audio data for processing. There are two primary types of audio formats utilized by researchers who study SELD:

1. First-Order Ambisonics (FOA): This is a type of spatial audio format that uses four channels. It has an omnidirectional W channel, along with three directional X, Y, and Z component channels. Together, these four channels provide a way to capture and describe the first order of the surrounding ambient audio environment. FOA has been widely used in SELD research, especially in DCASE datasets, often alongside MIC-format recordings [13–15].
2. Microphone Array (MIC) Format: Files that utilize the MIC format contain the raw, original recordings of multichannel audio data collected by different microphone-array layouts (e.g., linear, circular, spherical). Unlike FOA, MIC-format recordings are directly tied to the physical microphone layout, which makes them useful for studying compact mobile, wearable, and embedded microphone arrays. Because of these constraints, most edge/mobile SELD systems use either small microphone arrays or spatial formats such as FOA. However, there are some research efforts that focus on developing high-density localization systems that incorporate large numbers of microphones. One example is the 32-channel Eigenmike (mh acoustics LLC, Summit, NJ, USA). Although it provides rich spatial information compared to smaller arrays and/or FOA formats, its use may result in significant increases in computational cost [33].

Both the selection of an appropriate audio format and the design of a microphone layout can greatly affect the performance of a SELD system and the amount of computational resources required to run that system. For example, FOA provides standard spatial

coverage across all four available channels; however, MIC-format arrays provide flexibility in terms of layout, spacing, and number of channels. This flexibility comes at a price: each unique layout requires a customized implementation [2,30]. For edge applications, microphone layouts are typically limited by the physical dimensions of the device. Therefore, most edge applications utilize compact spatial formats or small microphone arrays, such as stereo input, headphone-based arrays, smartphone microphones, or larger custom wearable/embedded arrays [1,2,19,33,34].

2.4. Model Families

Over the years, a variety of neural network models have been used to develop multiple generations of SELD systems. Different models balance accuracy and computational requirements in different ways.

1. Convolutional Neural Networks (CNNs): CNNs are one of the main methods used as core components in many SELD studies. CNNs are primarily responsible for extracting spatial information and spectral characteristics from multichannel time-frequency features. Foundational and recent SELD systems utilize architectures such as ResNet-based front-end networks, CRNN convolutional front-end networks, and quaternion-based convolutional blocks prior to either temporal modeling or contextual modeling in an end-to-end system [23,29,35]. Related literature on edge AI/AV systems indicates that CNNs are capable of supporting real-time low-power processing using compact CNNs, specifically, SqueezeNet [36]; however, this example focuses on active speaker detection and does not attempt full SELD.
2. Recurrent Neural Networks (RNNs): Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) are typical recurrent neural network architectures used for modeling temporal dynamics in sequences of audio feature vectors. The SELDNet system comprises convolutional layers to extract features, along with recurrent layers to perform temporal modeling, generating distinct output channels for sound event detection and DOA estimation [23]. SELD-TCN demonstrated that replacing the recurrent block of SELDNet with temporal convolutional layers can improve hardware-friendliness and substantially reduce inference time [31]. Related edge acoustic detection work also highlights the importance of reducing computation, memory use, and energy consumption for resource-constrained audio models [37].
3. Temporal Convolutional Networks (TCNs): TCNs provide an alternative to RNNs for temporal modeling, since they use dilated temporal convolutions to capture long-range dependencies while allowing more computations to be performed in parallel. A study called SELD-TCN showed that substituting the recurrent block of SELDNet with TCN layers made the architecture more hardware-friendly and greatly reduced inference time [31]. Later studies have used TCN blocks within hybrid architectures to model long-term temporal dependencies, including G-SELD with bio-inspired gammatone features [38]. More recent studies continue this direction with multi-scale and lightweight TCN designs, such as MMT-Net's Modern Multi-Scale TCN block designed for real-time SELD [19].
4. Transformers and Conformers: Attention-based and Conformer architectures have achieved strong results in recent SELD benchmarks by using self-attention modules capable of modeling long-range temporal dependencies as well as spectro-temporal dependencies [35,39]. Conformers combine convolutional and self-attention modules and balance local and global feature extraction [35]. Despite this, the quadratic complexity of self-attention makes it difficult for real-time edge deployment, resulting in various works on hybrid Transformer-TCN architectures and attention alternatives such as selective state-space models [19,40].

5. Hybrid Architectures: Recently, researchers have begun combining multiple architectures to increase both the efficiency and accuracy of their models. For example, CRNN-based systems, which combine convolutional feature extraction with recurrent temporal modeling, have been implemented on edge devices such as Raspberry Pi for real-time SELD analysis [28]. Some hybrid implementations combine CNN, TCN, RNN, and fully connected components to take advantage of both spectral-spatial feature extraction and longer temporal relationships; G-SELD is an example of this type of implementation [38]. More recently, some researchers have developed new systems by combining transformer-based models for global context understanding with multi-scale TCN back ends to improve real-time capability while maintaining competitive sound event detection and localization performance [19].

3. Lightweight SELD Model Architectures

The need to deploy SELD on edge and mobile platforms requires a trade-off between maintaining good localization and detection performance and operating with low computation and limited memory. In this section, we review some recent developments in lightweight architecture designs for SELD models. These include compact CNNs, Temporal Convolutional Networks (TCNs), efficient Transformers, and techniques for compressing models.

3.1. Compact CNN-Based Architectures

Convolutional Neural Networks (CNNs) remain among the most popular models for Sound Event Localization and Detection (SELD) because of their ability to process both spatial and spectral characteristics. To improve the feasibility of deploying CNN-based models in edge computing environments, researchers have created compact versions of CNNs that require fewer model parameters and reduce computational complexity.

Brignone et al. introduced a compact quaternion-based SELD system called QSELD-TCN, which is optimized to run efficiently in the quaternion domain using First-Order Ambisonics (FOA) input [29]. The QSELD-TCN model processes all four Ambisonic channels as a single quaternion-valued signal. Additionally, it utilizes quaternion-valued convolutional, Temporal Convolutional Network (TCN), and fully connected layers to capture inter-channel correlation. As a result of using quaternion values, QSELD-TCN requires significantly fewer trainable parameters than its real-valued counterparts. Brignone et al. demonstrated that QSELD-TCN produces similar or improved SELD performance compared to real-valued counterparts while reducing runtime memory usage, storage requirements, and inference time. Therefore, the QSELD-TCN model appears suitable for deployment in constrained environments [29].

In order to create additional efficiencies in SELD systems, researchers have explored the use of group convolutions and depthwise separable convolutions. Inspired by MobileNet architectures, Shah et al. developed a single-stage CRNN model incorporating group convolutional layers for deployment on small-form-factor devices such as the Raspberry Pi [28]. Using Shah et al.'s proposed architecture, they achieved an F1-score of 90.0% while obtaining Real-Time Factors (*RTF*s) ranging from 0.5 to 0.7, demonstrating that the proposed system provides feasible real-time processing capabilities on severely limited hardware resources [28].

Researchers have employed ResNet-style backbones to extract features from multiple types of sound-related localization tasks, including SELD. ResNet-style backbones are often combined with other network elements such as Conformer blocks, as demonstrated in ResNet-Conformer architectures for SELD and source-distance estimation [35]. Related real-time edge-SELD studies also investigate CNN-based deployment bottlenecks on con-

strained hardware [20]. Marhoff et al. used raw four-channel microphone-array signals to generate beamforming source maps for active audio frames. These acoustic source maps, rather than the raw microphone signals, were then provided as input to a CNN-based source-map recognition stage to estimate source DOA [41]. However, this approach performs localization only and does not constitute complete SELD. This beamforming-derived acoustic-map approach is specific to some localization architectures and is not a general requirement for neural-network-based SELD, since many SELD systems operate directly on multichannel spectral and spatial features rather than precomputed acoustic maps [20,21,23,28].

To further optimize resource consumption in extremely resource-constrained environments, researchers have proposed the development of ultra-minimalist architectures. Yin et al. developed a CNN-based sound-source localization system that employs SRP-PHAT features for deployment on extreme edge devices with severe constraints on both computational and memory resources [30]. Yin et al.'s design focused on the trade-off between accuracy and efficiency and provided low-latency localization at reduced computational cost [30].

3.2. Temporal Convolutional Networks

Temporal Convolutional Networks (TCNs) have become a popular choice for replacing Recurrent Neural Networks (RNNs) when handling temporal dependencies within Sound Event Localization and Detection (SELD). Due to their ability to use dilated temporal convolutions to handle large temporal ranges, they are easier to parallelize than RNN-based architectures. In addition, TCNs offer many advantages for real-time and edge-oriented applications.

Brignone et al. expanded upon TCN-based SELD in the quaternion domain. Instead of treating each channel as a separate signal, they treated all four Ambisonic channels as one quaternion value. They used this approach to represent the feature-extraction process as a series of quaternion convolutional and Q-TCN layers. This method resulted in a lower number of parameters, lower memory usage, and lower computational requirements while still preserving inter-channel spatial relationships [29].

Guirguis et al. presented SELD-TCN. The main difference between SELD-TCN and other SELD networks was that, instead of utilizing an RNN layer, it utilized temporal convolutional layers. The authors found that SELD-TCN had comparable performance to RNN-based architectures. Additionally, due to the parallelized nature of convolutional operations, SELD-TCN had much faster inference times. As a result, Guirguis et al. believed that TCNs were better suited for modern hardware accelerators [31].

Kim et al. developed MMT-Net, a lightweight hybrid network that consists of a transformer-based front end and a multi-scale modern temporal convolutional network (MM-TCN) back end [19]. By developing such a hybrid network, Kim et al. [19] were able to achieve nearly 7× fewer parameters than conformer-based networks. This resulted in an approximate 2-million-parameter count for MMT-Net. Moreover, MMT-Net achieved competitive F1-scores while achieving high processing speeds. For example, MMT-Net was able to operate at speeds of over 20 frames per second (fps) on both NVIDIA Jetson Xavier NX and Orin AGX platforms when utilizing only two microphones [19]. Furthermore, the multi-scale aspect of MM-TCN allows MMT-Net to capture temporal features at multiple scales. These features allow MMT-Net to model sound events with various lengths.

The primary reason that TCNs are efficient is that they utilize dilated convolutions. Dilated convolutions increase the temporal receptive field without increasing the sequential dependency in the architecture. Therefore, TCNs provide many advantages for sound event localization and detection where the lengths of sound events vary greatly, from transients

to sustained sounds. In addition, recent advancements in TCN-based SELD architectures include the utilization of residual connections, multi-scale temporal convolutions, and lightweight convolutional structures in order to enhance both training stability and real-time applicability [19,31].

3.3. Efficient Transformer and Conformer Models

Transformer architectures have reached the top levels of performance for SELD benchmarks due to their capacity to model long-range dependencies through self-attention. However, the high quadratic complexity of self-attention with respect to sequence length creates significant limitations for real-time edge implementation. As such, recent research focuses on creating lower-complexity Transformer variants and hybrid architectures that provide good performance relative to computational cost.

The Conformer model type combines convolutional units with self-attention to create a framework that is particularly well suited for SELD tasks. Convolutional units are capable of capturing local patterns efficiently, while self-attention is able to model global relationships. In addition, Yeow et al. showed that ResNet-Conformer architectures can be used for competitive SELD performance with source-distance estimation [35]. The authors of this paper also developed a Squeeze-and-Excite version of the ResNet-Conformer that includes both channel and spatial attention to enhance feature separation and SELDDE performance [35].

Mu et al. investigated selective state-space models, including SELD-Mamba, as alternatives to transformer/conformer-based architectures for the task of source-distance estimation in SELD [40]. While SELD-Mamba provides linear complexity with respect to sequence length, unlike standard self-attention with quadratic complexity, it retains long-range context modeling capabilities at reduced computational cost, thereby providing a potential solution for deploying systems under limited resources [40].

Another possible future direction is hybrid transformer−TCN architectures. For example, MMT-Net uses a Transformer front end for extracting initial features, followed by a TCN back end for modeling time-series data [19]. By leveraging Transformer models' ability to extract global contextual information during feature extraction, along with TCNs' efficiency for modeling sequences, MMT-Net achieves approximately 2 million parameters and performs in real time on some types of embedded platforms without a large amount of accuracy loss [19].

Shul et al. [39] proposed dividing spectral and temporal attention into separate units. They applied spectral attention independently from the application of temporal multi-head self-attention. Dividing these two attention mechanisms allows the model to first collect features over all frequencies and then use temporal multi-head self-attention to model these collected features, improving overall SELD performance using real acoustic scenes from the STARSS23 development dataset [39].

3.4. Model Compression and Hardware-Oriented Optimization Techniques

Model compression and hardware-oriented optimizations are alternatives to architectural innovation that can be used to optimize both the amount of computation required for a model to perform tasks and the memory that models require for edge-based deployments. Some examples of these alternative methods include the following:

1. Quantization: Using a reduced-precision representation of model weights and activations. Examples of reduced-precision representations are 8-bit integers and/or binary values. For example, Cerutti et al. showed how using binary neural networks could be very beneficial for detecting sound events on resource-constrained devices because they had much smaller memory footprints and were able to save energy [42]. Similar

results have been reported for acoustic event detection. Related compact acoustic event detection work has also shown that reducing model precision, such as using 8-bit representations instead of 32-bit floating point, can reduce memory use and execution cost with only limited performance loss [37].

2. Pruning: Removing "redundant" or "less important" connections in neural networks. Pruning reduces the number of parameters in a network, therefore reducing the size of the trained model, as well as reducing the computational cost associated with each layer. There are two types of pruning: unstructured pruning removes individual weights or connections, whereas structured pruning removes larger units such as filters, channels, or blocks. Structured pruning is especially interesting when deploying models on hardware accelerators, since many accelerators rely on fixed-size kernels. Unfortunately, we found no literature directly demonstrating the effectiveness of pruning for SELD applications.
3. Knowledge Distillation and Model Collaboration: Knowledge distillation uses a larger or stronger teacher model to train a smaller student model. Typically, knowledge distillation is done during training. However, Srinivas et al. recently proposed a method called knowledge boosting, which performs distillation after training and during inference. In their method, a smaller model receives delayed feedback from a larger model during inference. Their work focused on speech separation and enhancement, not SELD [43].
4. Neural Architecture Search (NAS): NAS searches through a defined space of possible model configurations to find the optimal one given certain criteria. Although NAS has seen success in many other areas, we did not find any examples of NAS being applied to SELD problems for edge devices.
5. Hardware−Algorithm Co-Design: Hardware−algorithm co-design is a process that considers both the limitations of hardware, such as latency and memory, and the design of algorithms for those same platforms. The I-SPOT project demonstrated hardware-aware acoustic perception optimization, reporting a model that was approximately 86% smaller and 47% faster than a baseline version [8].

4. Effective Representations of Audio Features

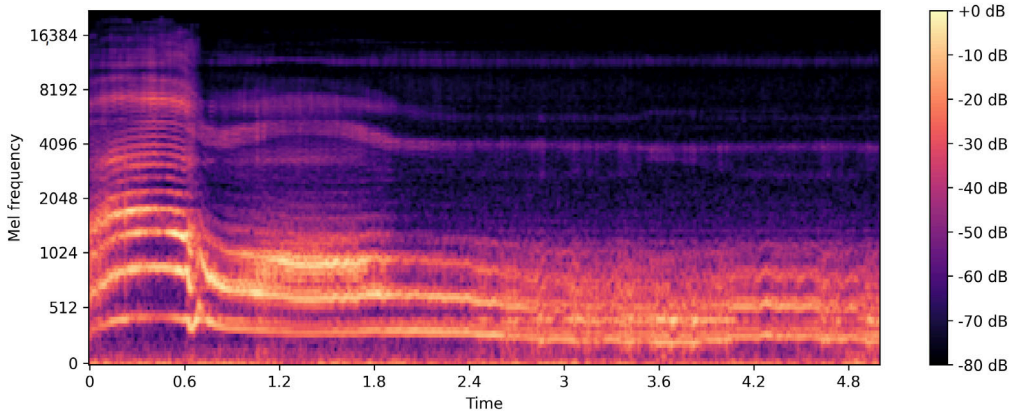
The choice of audio features affects both SELD performance and computational efficiency. In this section, we evaluate spectral features, spatial features, and learned representations by comparing their trade-offs for edge and mobile deployments.

4.1. Spectral Features

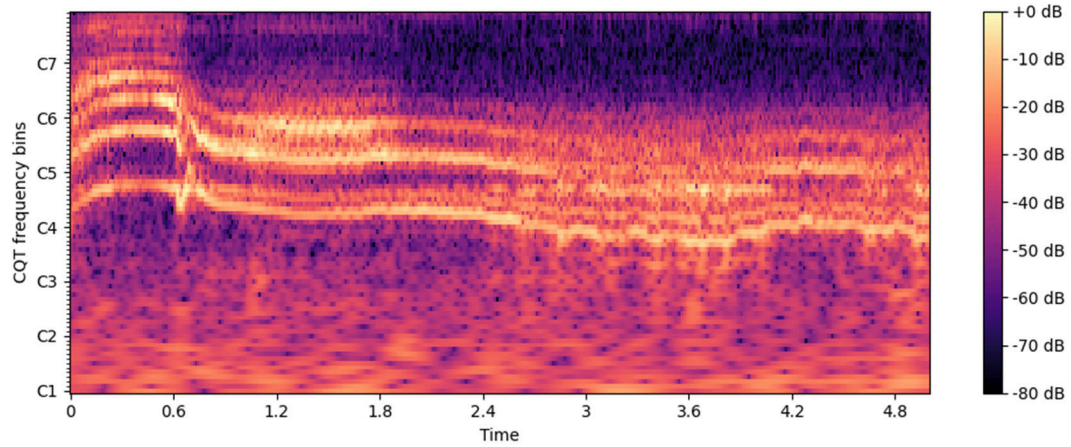
Spectral features obtained from time-frequency analysis form the basis of most SELD systems. A typical way to represent spectral information is through the use of a log-mel spectrogram. The process of creating a log-mel spectrogram typically begins by taking the logarithm of the power spectrum or magnitude after applying a mel-scale filterbank. Through the application of these two transformations together, the resulting representation is similar to what humans perceive through hearing. Log-mel spectrograms provide a time-frequency representation that is compact enough to fit into a CNN's memory space. Often, log-mel spectrograms are combined with spatial features such as GCC-PHAT for SELD [20,28]. Some other ways to obtain effective SELD feature representations involve combining spectral features with spatial cues, such as normalized inter-channel phase differences, as in SALSA-Lite [21].

Figure 3 illustrates examples of common spectral feature representations used in audio learning systems. Although all three representations describe how signal energy changes over time and frequency, they differ in frequency scaling and filterbank design. Log-mel

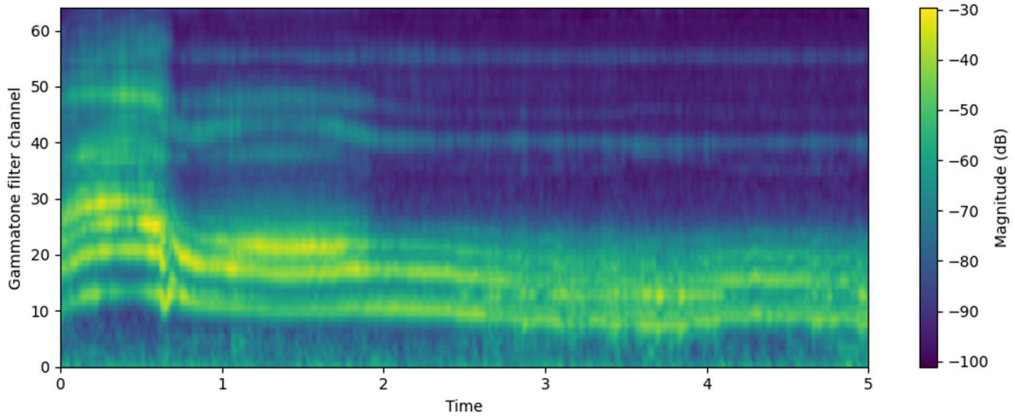
spectrograms are widely used because they provide a compact perceptual representation, CQT features use logarithmically spaced frequency bins, and gammatone features approximate aspects of cochlear frequency analysis.



(**a**) Log-mel



(**b**) CQT



(**c**) Gammatone

Figure 3. Examples of spectral feature representations for audio analysis: (**a**) log-mel spectrogram, (**b**) CQT spectrogram, and (**c**) gammatone spectrogram. Each representation shows time on the horizontal axis, frequency or filter channels on the vertical axis, and signal energy through color intensity.

The computational cost of mel-spectrogram extraction depends on parameters such as FFT size, hop length, and the number of mel bins. As part of evaluating various MIC-format feature combinations, Shah et al. tested log-mel energy with intensity vectors, CQT with intensity vectors, log-mel energy with GCC-PHAT, and log-mel energy with both

intensity vectors and GCC-PHAT [28]; they also reported that they chose 64 mel bins as the mel-filter setting for extracting features. Reducing FFT size reduces computation but may compromise frequency resolution; reducing hop length increases the number of frames processed per second, which improves temporal resolution but increases computation. Therefore, real-time SELD researchers must pick feature-extraction settings based on the target hardware. For example, Shah et al. selected 1024-sample windows and 320-sample hops at 32 kHz for their experiments [28], while Yeow et al. selected 512-sample windows and 300-sample hops at 24 kHz for their experiments [20].

Other spectral representations exist, such as the Constant-Q Transform (CQT), which, similarly to log-mel features, employs a nonlinear frequency scale and has been studied as a spectral feature for SELD in conjunction with spatial cues such as intensity vectors [28]. However, the CQT is somewhat less frequently employed as a spectral feature in SELD and edge-SELD pipelines than log-mel-based features. Rather, real-time SELD research papers tend to focus on log-mel, MelGCC, SALSA, and SALSA-Lite features.

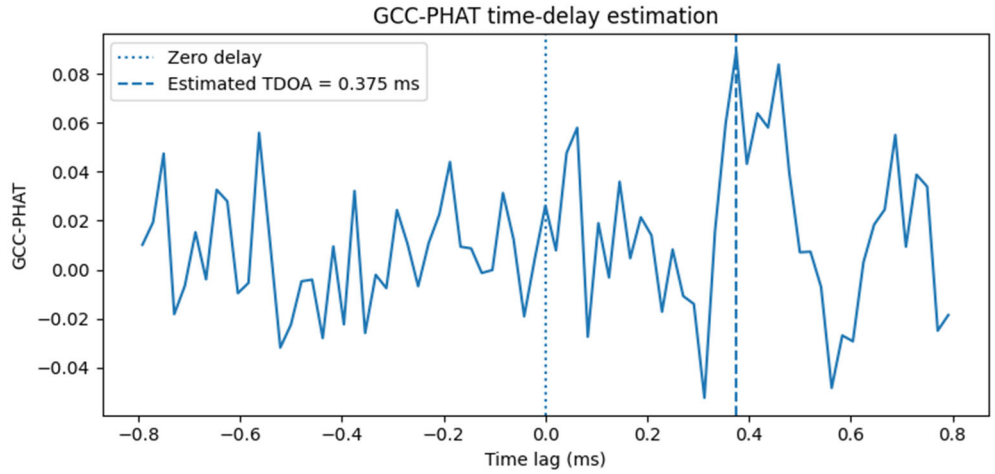
Auditory model-inspired Gammatone filterbanks have been examined for SELD. Unlike mel-scale filterbanks, Gammatone filterbanks are designed to be closer approximations to the tonotopic frequency distribution found in the cochlea. Rosero et al. developed G-SELD, which combines log-gammatone spectrograms, intensity vectors, and a neural network containing CNNs, TCNs, RNNs, and FCNs [38]. Results from G-SELD demonstrate that substituting mel filterbanks with Gammatone filterbanks can lead to improved SELD metric values and that, overall, the full G-SELD system operates effectively across the tested conditions with varying degrees of polyphony, reverberation, and interference [38]. However, because adding another preprocessing step can be detrimental to deployability on resource-constrained edge devices, the suitability of Gammatone filtering should be thoroughly evaluated.

For applications requiring very low levels of power consumption, event-driven acoustic sensors could potentially serve as alternatives to conventional frame-based spectral analysis. Ceolini et al. demonstrated a low-latency keyword-spotting/verification system employing a binaural spiking silicon cochlea sensor; additionally, they achieved >93% keyword-spotting accuracy and >88% speaker-verification accuracy, along with a lowest possible system latency of 5 ms [44]. Although Ceolini et al. [44]'s system contains localization and source-separation components, it is not a complete SELD system; however, it does present one potential approach to realize ultra-low-power acoustic perception.

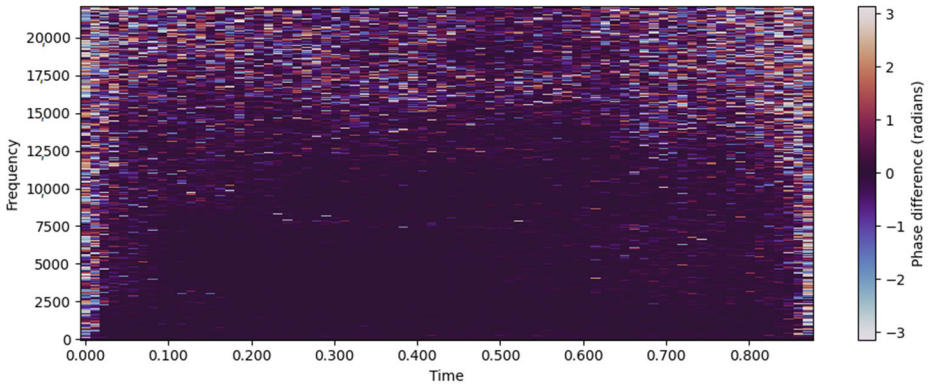
4.2. Spatial Features

To capture the directionality of sound in multichannel recordings, spatial features encode directional information. The best type of spatial feature will vary depending on the specific characteristics of the microphone-array configuration, as well as the trade-off between localization accuracy and processing cost. Figure 4 illustrates representative spatial feature representations, including GCC-PHAT, inter-channel phase difference (IPD), and intensity-vector (IV) features.

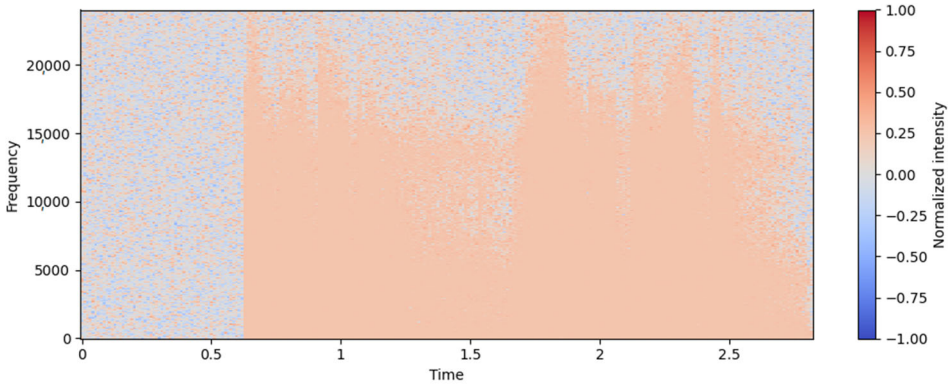
1. Intensity Vectors (IVs): IVs have been most frequently used as spatial features in FOA-based SELD systems. For First-Order Ambisonic data, intensity-vector features are created from the relationship between the omnidirectional channel and the directional channels. This produces three spatial components that provide directional information about each event. Intensity-vector features are often paired with log-mel spectrograms as FOA input features to SELD systems [21,22,39]. While IVs use spatial energy relationships in the sound field to create localization cues, these cues may degrade under background noise, reverberation, and diffuse reflections in realistic acoustic settings.



(**a**) GCC-PHAT



(**b**) IPD



(**c**) IV

Figure 4. Examples of spatial feature representations used in multichannel audio analysis. (**a**) GCC-PHAT response, where the peak indicates the estimated time delay between two microphone channels. (**b**) Inter-channel phase difference (IPD) heatmap, where color represents the wrapped phase difference between channels over time and frequency. (**c**) Intensity-vector component representation, where color indicates normalized directional acoustic energy over time and frequency.

2. Generalized Cross-Correlation with Phase Transform (GCC-PHAT): GCC-PHAT is another common type of spatial feature that uses time-difference-of-arrival (TDOA) estimation to determine where a sound originates between microphone channels. GCC-PHAT utilizes the phase aspect of the signal while reducing the magnitude aspect, thus allowing it to be applied effectively for analyzing spatially related signals within reverberant environments. GCC-PHAT has been successfully combined with log-mel features for real-time SELD systems [28]. Additionally, Yin et al.'s CNN-based

SRP-PHAT system showed that related PHAT-based spatial correlation features can support localization on extreme-edge devices [30].

3. SALSA (Spatial Cue-Augmented Log-Spectrogram Features): SALSA combines multichannel log spectrograms with spatial cues obtained from the principal eigenvectors of spatial covariance matrices. In SALSA, the spatial component uses eigenvector-derived phase information. SALSA-Lite replaced this with Normalized Inter-channel Phase Difference (NIPD) features calculated directly from multichannel complex spectrograms [21]. Due to this substitution, SALSA-Lite greatly reduced feature-extraction costs compared to full SALSA while maintaining effective spatial cues for microphone-array configurations in polyphonic SELD [21].
4. Inter-channel Phase Difference (IPD): IPD features measure the phase relationships between microphone channels. Since IPD measures phase differences between microphone channels, it provides spatial information without needing to estimate TDOA via cross-correlations. Wang et al. developed IPDnet2, an efficient network for direct-path IPD estimation in sound-source localization, utilizing frequency-time pooling to reduce computing costs while maintaining localization capabilities [45]. Although IPD-based features are beneficial for microphone arrays with known geometries, their spatial cues can be affected by noise, reverberation, phase mismatches, and spatial aliasing [21,45].

The computational cost of extracting spatial features can differ dramatically. As a result, Intensity Vectors for FOA are generally lightweight since they are derived from the FOA channels and align spatial cues with time-frequency energy. Because GCC-PHAT requires pairwise processing between microphone channels and does not align its time-lag dimension with the spectrogram frequency dimension, GCC-PHAT typically results in higher computational costs and can be less desirable when attempting to resolve overlapping events [21]. Overall, SALSA is the most computationally expensive feature extractor since it requires eigenvector computations. However, SALSA-Lite reduces this cost by utilizing NIPD features and has been shown to be many times faster than both full SALSA and MELSPECGCC [21]. Therefore, when considering deployment on edge devices, a key factor in choosing the appropriate spatial feature should be balancing localization accuracy, feature-extraction cost, and the microphone configuration being used [20,21].

4.3. Learned and Hybrid Features

Recently, there has been increasing interest in using learned features in SELD and related spatial-audio tasks. In SELD, learned features allow the neural network to generate task-specific representations from either raw audio or very-low-level processing. This method can generate features that are specific to the task and reduce reliance on predefined hand-crafted representations. However, it may also increase model complexity and deployment cost.

1. End-to-end learning from raw waveforms: In addition to eliminating the requirement for extracting hand-crafted feature representations prior to training, end-to-end learning from raw waveform audio allows the network to learn representations that are specific to the task from the raw audio itself. He et al. presented SoundDet, an end-to-end framework for detecting and localizing polyphonic moving sound events in real time using raw multichannel waveforms [46]. SoundDet used phase-sensitive and frequency-selective filters to preserve DOA (Direction of Arrival) information while minimizing the cost associated with processing raw waveform inputs. However, because raw audio produces long input sequences, real-time edge deployment requires careful optimization of both the model architecture and the learned features.

2. Hybrid Approaches: Many hybrid methods utilize traditional engineered time-frequency and spatial features as inputs to CNNs, CRNNs, or attention-based models, e.g., log-mel spectrograms, intensity vectors, GCC-PHAT, or SALSA-style features [21,22,28,39]. This approach uses the computational efficiency and interpretability of traditional feature-extraction techniques while also allowing the neural network to learn higher-level task-specific representations.
3. Learned Spatial Features: Another hybrid direction is learned spatial features. As opposed to only utilizing fixed spatial feature-extraction algorithms, such as GCC-PHAT or intensity vectors, some hybrid methods use neural network layers to learn or estimate spatial representations. An example of this type of hybrid system is Wang et al.'s IPDnet2 [45], which estimated direct-path inter-channel phase differences for sound-source localization and utilized frequency-time pooling to minimize computational requirements. Although IPDnet2 was designed specifically for SSL rather than full SELD, it illustrates the potential for learned spatial representations to support localization under realistic conditions.
4. Self-supervised and pre-trained features: It is also possible that self-supervised and pre-trained features could become a viable means of enhancing SELD performance when high-quality labeled spatial audio datasets are scarce. For example, if large-scale audio data can be used to train pre-trained models, then it is possible that these models can provide a good starting point for learning general acoustic representations that transfer to SELD tasks, as demonstrated in recent self-supervised SELD works such as w2v-SELD [47]. However, large pre-trained models may be prohibitively expensive to deploy at the edge due to their size and memory usage. Therefore, compression, distillation, or lightweight pre-training methods would likely be required.

5. Deployment Challenges

SELD deployment onto edge and mobile devices presents multiple challenges, including latency constraints, hardware constraints, microphone-array constraints, and hardware-specific considerations. The purpose of this section is to review these deployment challenges and strategies for addressing each challenge. Figure 5 summarizes the end-to-end edge-SELD processing pipeline and highlights the major deployment constraints that can affect each processing stage.

5.1. Latency and Real-Time Requirements

Real-time SELD deployment involves several different timing-related metrics, and this section is intended to disambiguate and clarify them.

Inference time is the time required by the neural network to produce an output after all pre-computed input features have been provided. Therefore, inference time is only one component of the total time required for SELD deployment. It does not include audio capture time, buffering time, feature-extraction time, or any post-processing time.

The real-time factor (*RTF*) is defined as follows:

$$RTF = \frac{T_{processing}}{T_{audio}}$$

$T_{processing}$ = The processing time included in the reported measurement T_{audio} = the duration of the evaluated audio.

If $RTF < 1.0$, then the system was able to process the audio faster than the duration of the audio under the tested setup. However, many factors can influence whether this translates into a system with low latency, causal operation, use of only current and previous audio, and support for continuous streaming deployment.

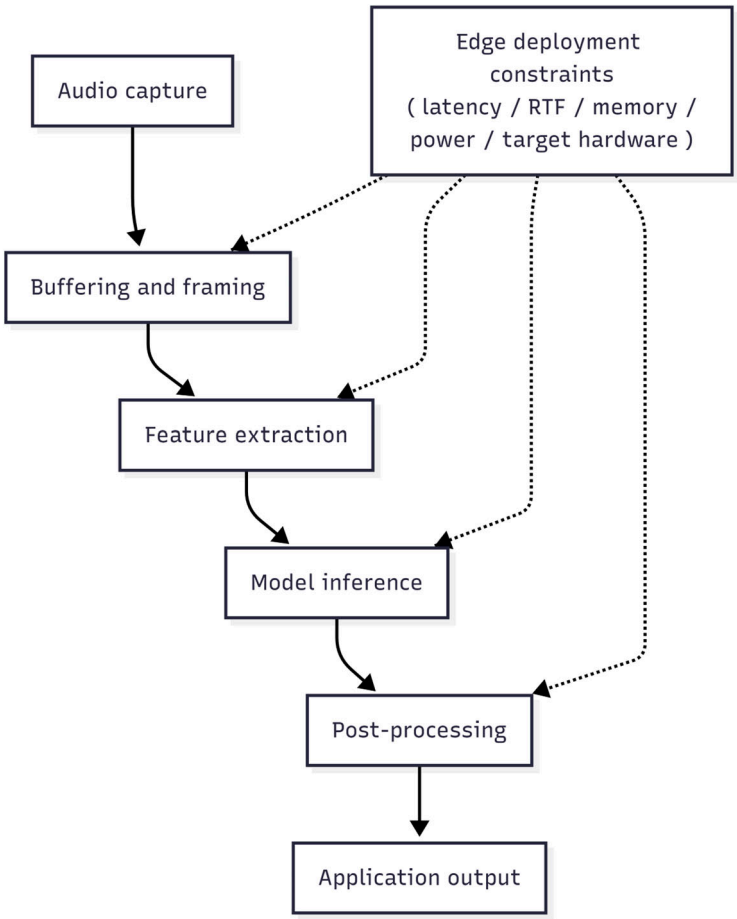


Figure 5. End-to-end edge-SELD deployment pipeline and major deployment constraints. Real-time performance depends not only on model inference, but also on audio buffering, feature extraction, post-processing, and application output latency. Edge deployment is therefore constrained by latency, real-time factor (*RTF*), memory use, power consumption, and target hardware characteristics [20,48]. Solid arrows indicate the main processing flow, whereas dotted arrows indicate the influence of edge-deployment constraints on the corresponding processing stages.

Algorithmic latency is the delay introduced by signal processing and the design of the model, such as window length, hop size, look-ahead, future context, and post-processing delay. Any model that requires future context or look-ahead must have a mandatory minimum latency equal to the duration of the look-ahead period, and this cannot be reduced by using faster hardware or a lower *RTF*.

End-to-end latency is the total time for a SELD system to operate from audio capture to the final SELD output. It includes the delays caused by buffering, feature extraction, model inference, post-processing, and output generation. Therefore, deployment-oriented SELD evaluations should report which processing steps are included in the timing measurement. These timing-related measures and their limitations are summarized in Table 3.

Latency from capture to localization output must be low enough to be acceptable for the targeted application. There will always be some minimum amount of time before the result is shown. Depending upon the application or use case, there could be significant differences in acceptable latency. Active noise control, hear-through processing, and virtual-sound rendering are examples of possible use cases. On top of that, different processing blocks may have varying degrees of acceptable latency based upon the requirements of the application. AR/MR hearables typically have very tight real-sound control timing requirements and therefore must operate at very low latency. Virtual-sound rendering may tolerate larger total latency than real-sound control tasks, depending on the application [7].

Table 3. Timing-related measures for real-time edge-SELD deployment.

Measure	Meaning	What It Includes	Main Limitation
Inference time	The time taken by the forward pass through the neural network.	Normally model inference only, once features have been generated.	Buffering, feature extraction, post-processing, and audio capture are not included.
Real-time factor (*RTF*)	The ratio of the time used by the processor to the duration of the audio.	Depends on which parts of the process the study includes in its metrics.	If $RTF < 1.0$, then, on average, the system can keep up with the testing audio; however, this may not mean that the system has low latency, operates causally, or is capable of continuous real-time streaming.
Algorithmic latency	The amount of time added by the design of the algorithm before an output can be produced.	Window length, hop size, look-ahead, future context, smoothing, and/or tracking.	Even with low *RTF*, there could still be significant algorithmic latency. There is also a minimum latency caused by any use of look-ahead, regardless of how quickly the hardware runs.
End-to-end latency	Total delay from audio input to final SELD output.	Audio capture, buffering, frame generation, feature extraction, inference, post-processing, and output generation.	It is the most important timing measure during deployment, but it is rarely completely reported.

For SELD edge-device deployment, latency needs to be measured across all parts of the SELD pipeline instead of only measuring neural network inference [20]. Yeow et al. conducted an in-depth examination of the various latency components involved in deploying real-time SELD on edge devices [20]. In addition to audio buffering and post-processing, they identified other causes of latency, including neural network inference and feature extraction. In order for the system to meet real-time requirements, the processing time of each block must not exceed the capture-block duration [20]. On edge devices, feature extraction has been found to be one of the largest computational bottlenecks [20]. Both feature extraction and neural network inference are primary causes of latency in real-time SELD deployments [20]. As described in the I-SPOT project, co-designing hardware and algorithms is critical to developing embedded acoustic perception capabilities. The I-SPOT team's edge-device versions were approximately 86% smaller and 47% faster than their baseline implementation [8]. Additionally, Kim et al.'s MMT-Net was capable of achieving greater than 20 fps on NVIDIA Jetson Xavier NX and Orin AGX platforms, indicating that with a lightweight hybrid Transformer-TCN architecture, it is possible to deploy real-time SELD on edge devices [19].

Shah et al. reported *RTF*s between 0.50 and 0.70 for their single-stage CRNN deployed on Raspberry Pi [28]. Brignone et al.'s quaternion-domain version of QSELD-TCN demonstrates reduced memory usage, lower storage costs, and faster inference than its real-valued counterparts; however, the reported efficiency comparison was not a Jetson Nano *RTF* test [29].

As previously mentioned, algorithm selection impacts latency. TCN-based SELD architectures tend to be more hardware-friendly than RNN-based SELD architectures because convolutional computations are much easier to parallelize than RNN computations, which inherently have sequential dependency limitations that restrict parallelism [31].

However, SELD-TCN utilizes non-causal convolutions in order to simulate bidirectional RNN context, so causal TCNs need to be considered separately when evaluating the feasibility of streaming deployment [31].

There is additional low-latency hearable-related research showing how severe audio latency constraints can be mitigated using custom models and custom signal-processing techniques. Dementyev et al. showed that speech enhancement is feasible with sub-millisecond algorithmic latency and approximately 3.35 milliseconds of end-to-end latency on low-power DSP hardware; however, speech enhancement is not SELD [49].

5.2. Computational Resource Consumption (Processing Power and Memory)

Unlike cloud servers, edge and mobile devices are severely limited by their processing capabilities, memory availability, and total available power. These hardware differences are summarized in Section 5.4.1, while the present section focuses on how processing capability, memory availability, and power consumption affect SELD deployment. For instance, Chen et al. were able to implement SED on a CC1350 microcontroller (Texas Instruments Incorporated, Dallas, TX, USA) using only 33.4 kB flash and 9.94 kB SRAM [4], whereas Yin and Verhelst's "balanced extreme-edge" SSL model required 127.1 MFLOPs and 0.821 MB on-chip memory [30].

The size of a machine learning model determines how much memory it uses during inference as well as how fast the inference process completes. By reducing the model size to approximately 2–2.4 million parameters, Kim et al. were able to enable real-time SELD inference on both NVIDIA Jetson Xavier NX and Orin AGX platforms [19]. Also related is the fact that hardware−algorithm design at the edge has been important in the development of embedded acoustic perception, as highlighted by the I-SPOT project [8]. Additionally, Küçük showed that mobile sound-source localization on smartphone/tablet devices is possible, although this should be distinguished from full SELD [3].

Computational cost, typically measured in FLOPs or MACs, influences inference time and power consumption. Architectural decisions such as the use of depthwise separable convolution can significantly decrease the computational resources needed to run a model. A good example of this is shown in Cross3D-Edge, where extreme-edge sound-source localization was achieved [30]. Also, Brignone et al. used quaternion-domain SELD, where they decreased the number of parameters and memory usage by utilizing quaternion-valued signals to represent FOA components and quaternion neural layers [29].

Memory bandwidth and buffering requirements also contribute to resource consumption in real-time SELD systems; their effects on end-to-end latency are discussed in Section 5.1 [20]. For FPGA-based localization, Silva et al. showed that beamforming memory requirements depend heavily on sampling frequency, beamforming-stage placement, and microphone-array configuration [32].

Finally, power consumption is extremely important because most edge and mobile devices are battery-powered and always-on. Cerutti et al. implemented binary neural networks for sound event detection on tightly constrained energy-sensitive IoT devices and found significant energy-efficiency gains over Cortex-M4 execution due to their highly compact implementation [42]. Küçük described his prototype sound-source localization platform as requiring approximately 1.15 W [3], whereas Silva et al. described their FPGA-based sound-source localization architectures as having an average power consumption of approximately 17.8–23.7 mW [32].

When selecting a machine learning model for deployment at the edge, the primary trade-off being made is typically between model capacity and computational cost. Increasing the size of a model can lead to higher accuracy; however, larger models typically require more memory, compute, and energy to operate. The ideal operating point for a

given model deployed at the edge depends on many factors, including the target device, microphone configuration, latency requirement, and acceptable accuracy. End-to-end timing considerations are discussed separately in Section 5.1.

5.3. Microphone Array Constraints

The configuration of the microphone array (i.e., number of channels, geometric layout, and distance between elements) significantly constrains both localization performance and, consequently, SELD performance. For instance, edge and mobile devices typically rely on compact microphone configurations, including stereo microphones, smartphone-based microphone arrays, or headphone-mounted microphones [1,2,19]. Conversely, larger customized arrays and spherical microphone arrays have also been used in research systems, such as GPU-accelerated multi-microphone localization and 32-capsule Eigenmike systems [33,50]. Figure 6 illustrates representative microphone-array geometries relevant to edge and mobile SELD, ranging from compact smartphone and wearable configurations to larger research-oriented arrays.

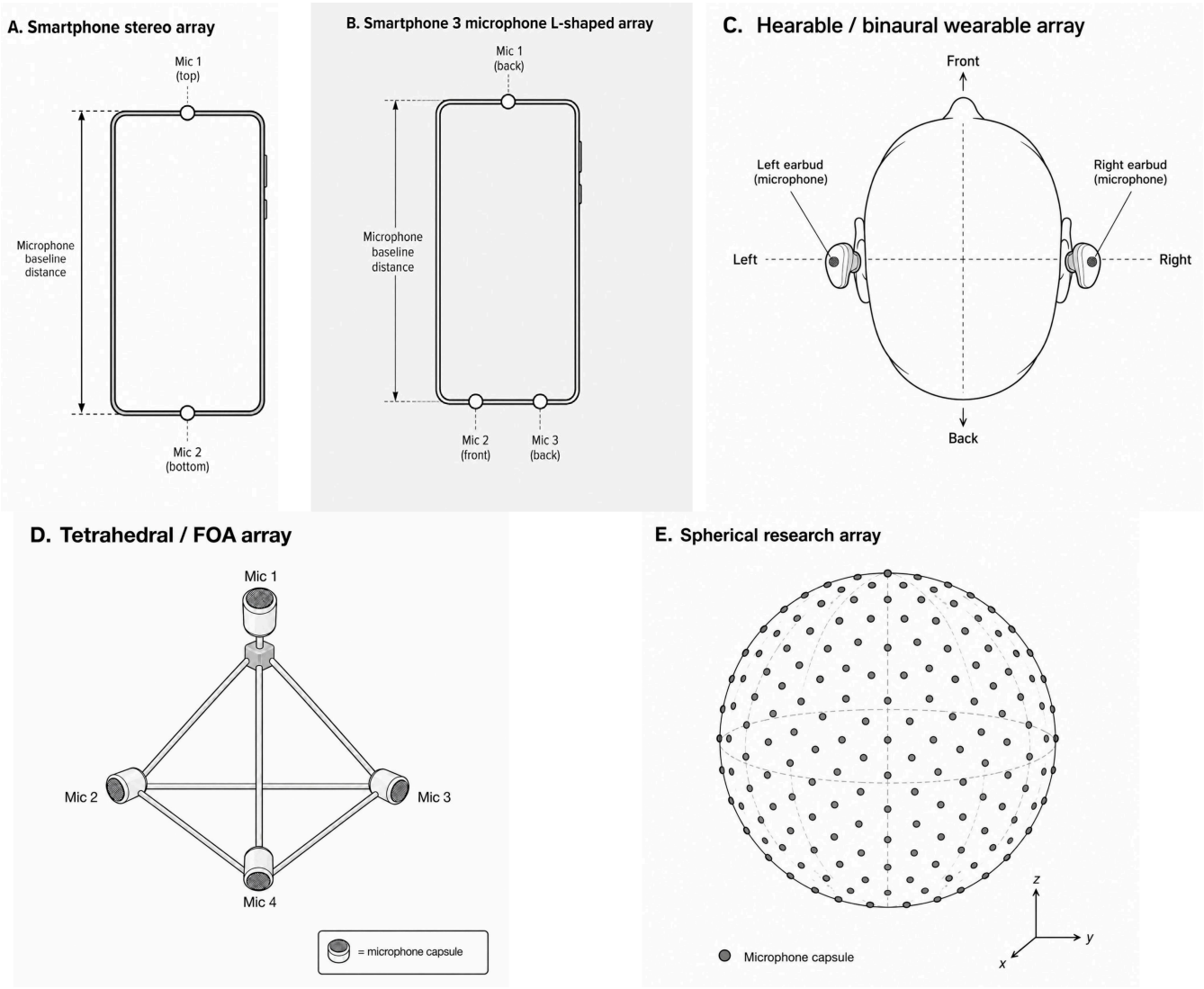


Figure 6. Simplified comparison of microphone-array geometries relevant to edge and mobile SELD. (**A**) Smartphone stereo array with a top-bottom microphone baseline, (**B**) non-uniform three-microphone smartphone array, (**C**) hearable/binaural wearable array, (**D**) tetrahedral array commonly used for first-order Ambisonics (FOA) capture, and (**E**) dense spherical research array.

1. Number of Channels: with increasing numbers of compact SELD systems being considered, smaller microphone counts are becoming more common. The motivation behind Kim et al.'s MMT-Net stereo SELD model utilizing only two microphones was that stereo input may be the only viable modality in many practical settings (e.g., mobile robots, consumer devices) [19]. Stereo setups, however, are much more complicated than those employing four-channel FOA or tetrahedral microphone formats, since elevation cues are either unavailable or reduced [19]. Yeow et al. demonstrated the application of wearable SELD utilizing simulated headphone data generated from four-channel headphone RIRs and highlighted that hearable-oriented SELD needs to take into consideration limited computational resources and wearable microphone configurations [1]. Similarly, Ganguly et al. evaluated the feasibility of a three-element non-uniform, nonlinear microphone array for smartphone-assisted hearing-aid applications, demonstrating that a compact smartphone array can provide robust speech-source localization under low-SNR conditions [2].
2. Array Geometry: The most common types of microphone geometry include linear, non-uniform linear, L-shaped or nonlinear, circular, and spherical configurations. Furthermore, SELD datasets may represent spatial information in FOA format. Ganguly et al. showed that an L-shaped three-microphone smartphone array can reduce spatial ambiguity and aliasing compared to uniform linear arrays and still allow for 360° DOA scanning capability [2]. Large customized systems can utilize spherical arrays; for instance, Belloch et al. employed a 32-capsule Eigenmike spherical microphone array for real-time 3D sound-source localization on an embedded GPU [33]. DCASE-style SELD systems often employ four-channel FOA or microphone-array formats, while 3D SELD systems such as Dong et al. utilize four-channel FOA input with log-mel and intensity-vector features [13,22].
3. Inter-Element Spacing: Spacing defines the frequency band over which an array can accurately localize sources without causing spatial aliasing. Small spacings help avoid aliasing at high frequencies; conversely, larger spacings can enhance sensitivity to temporal/phase differences at low frequencies. Ganguly et al. studied this trade-off for a Nexus 6 smartphone (Motorola Mobility LLC, Chicago, IL, USA) array with spacings of 1.2 cm and 12 cm and illustrated how non-uniform geometry can decrease spatial ambiguity and aliasing [2]. In GPU-accelerated SRP-PHAT localization, computational cost increases with the number of microphone pairs and candidate directions, especially when using larger arrays and finer spatial grids [50]. Additionally, dense spherical-array systems such as a 32-capsule Eigenmike illustrate that increased channel counts can require considerable processing resources for real-time 3D localization [33].
4. Form-Factor Constraints: The device's physical form factor imposes significant constraints upon the design of the array. Smartphone arrays are designed based upon their hardware design, and thus their microphone positions are fixed. Hearables have limited space for the placement of microphones and processors, while embedded IoT devices must weigh their array size against their costs and power consumption. Yarkan's edge-AI acoustic localization system exemplifies another constraint by using only two low-cost microphones on an STM32 microcontroller platform, prioritizing lightweight real-time direction-finding over accuracy from a dense-array setup [51]. Belloch et al. further demonstrate that mobile GPUs can support more microphone channels for source-localization purposes than previously thought; however, such systems must still operate within reasonable computational bounds [50].

In summary, limited microphone arrays on edge devices produce less spatial information than larger customized arrays [2,19,33], thereby reducing SELD/localization

ability, particularly for elevation estimation, overlapping sources, noise, and reverberation [14,15,19]. Therefore, edge SELD requires algorithms and features capable of extracting spatial cues from limited microphone configurations instead of relying on assumptions about densely calibrated or large microphone arrays [1,2,51].

5.4. Hardware Limitations

5.4.1. Summary of Device-Level Constraints for Edge/Mobile SELD

Various edge and mobile devices differ in terms of processing speed, memory capacity, available power and battery life, microphone inputs, and hardware acceleration capabilities. As illustrated in Figure 5, the differences between devices can be expected to have a direct impact on all phases of the entire SELD system deployment pipeline, including audio capture, buffering, feature extraction, model inference, post-processing, and output generation. Furthermore, as illustrated in Figure 6, there are also geometrical differences in how the microphones are arranged, including array size and number of channels, with respect to each type of device, which will influence the amount of spatial information available to the SELD system. The main device-level constraints affecting edge/mobile SELD deployment are summarized in Table 4.

Table 4. Summary of device-level constraints for edge/mobile SELD deployment.

Device Type	Computing Power	Memory/Power	Numerical Format	Microphone Input	Main SELD Limitation
Microcontrollers/ MCUs	Very low	Tiny memory; very low power [4]	Integer/fixed-point	1–2 low-cost microphones	Full SELD is hard to implement due to limitations in model size, feature-extraction time/energy cost, memory capacity, and microphone-array size. As such, these platforms are more appropriate for simple event-detection or basic localization applications [4,52].
Wearables/ hearables	Very low to low	Strict size, battery, heat, and latency limits	Low-precision/DSP-supported	Earbud, headphone, glasses-mounted, or body-worn microphones	SELD systems on wearables are limited by microphone geometry, reduced channel count, head/body-related acoustic effects, and extremely low-power conditions, all of which can hinder robust localization and detection [1].
Single-board computers	Moderate	More than MCUs; still power/heat limited	FP32; INT8 possible with add-ons	USB microphones, small arrays, or audio hats	While useful for prototyping, the full SELD pipeline can still bottleneck because of feature-extraction cost, buffering requirements, and/or model-inference processing [28]. Reported *RTF* values will vary depending upon both the model-implementation details and the overall pipeline implemented on a given device [28].
Smartphones/ mobile SoCs	Medium to high	Battery and thermal limits	FP32/FP16; INT8 depends on backend	Built-in fixed microphones	Smartphone SELD implementations suffer from fixed microphone geometry, limited spatial resolution, battery life, thermal throttling, OS constraints, and mobile-inference performance limits [2,3,50].
FPGAs/ accelerators	High (parallel computing)	Efficient but resource-planning needed	Fixed-point/custom precision	Multi-microphone support if designed	FPGA-based implementations can be highly optimized for specific hardware; however, FPGA designs require hardware-design expertise and offer less flexibility compared to software-based deployments. The designs should match the number of channels, sampling rate, and beamforming requirements [32].

Table 4. *Cont.*

Device Type	Computing Power	Memory/Power	Numerical Format	Microphone Input	Main SELD Limitation
Embedded GPU/NPU devices	High (edge computing)	More capable, but heat/power-limited	FP32/FP16; INT8 possible	Stereo, USB arrays, embedded arrays, or multi-mic inputs	These devices are more likely to support real-time SELD implementations; however, deployment of these devices still relies upon limiting model size, feature-extraction cost, memory bandwidth, heat, and power consumption [19].

As a result, there is no single "best" edge computing platform for every implementation of SELD algorithms. Microcontrollers and wearable devices are among the most energy-efficient systems, but despite their low energy use, each type of system has inherent limitations in memory, processing capability, computational capability, and microphone configuration. In contrast, embedded GPU/NPU-based platforms can support much higher levels of real-time inference than microcontrollers or wearable devices. However, several factors still need to be considered when designing applications for these platforms, including model size, feature-extraction cost, memory bandwidth, and total power use.

5.4.2. Hardware-Specific Limitations and Examples

Hardware limitations for SELD and other acoustic perception tasks include constraints across the full processing pipeline, including microphone input, audio buffering, feature extraction, spatial processing, memory movement, post-processing, and power consumption [20].

1. Mobile SoCs: Current mobile system-on-chips (SoCs) include integrated heterogeneous compute units such as CPUs, GPUs, DSPs, and neural accelerators. Belloch et al. showed that mobile SoCs with integrated GPUs may be able to accelerate the SRP-PHAT sound-source localization algorithm and achieve real-time localization with high-resolution spatial grids and up to 12 microphones [50]. Belloch et al. further discussed the trade-offs between real-time performance, energy efficiency, and exploiting parallelism for IoT-oriented acoustic source localization using embedded SoCs [48]. Küçük's work relates to mobile acoustic localization and not full SELD; he showed real-time deep-learning-based sound-source localization on Android smartphones for hearing-aid-related applications with an average accuracy of 94.24% and a root mean square error (RMSE) of 14.75 degrees for smartphone DOA estimation [3].
2. Single-Board Computers and Embedded GPU/NPU Devices: Raspberry Pi and NVIDIA Jetson platforms have been popularly used in edge AI experimentation due to their ability to provide better computation and memory than microcontrollers while maintaining a small footprint. Shah et al. deployed a single-stage Convolutional Recurrent Neural Network (CRNN) for real-time SELD on a Raspberry Pi and achieved a Real-Time Factor (*RTF*) between 0.5 and 0.7 [28]. Kim et al.'s MMT-Net ran at over 20 FPS on both the NVIDIA Jetson Xavier NX and Orin AGX, demonstrating that lightweight Transformer-TCNs can support real-time embedded SELD [19]. The I-SPOT project has also highlighted the need for hardware−algorithm co-design for embedded acoustic perception by reporting optimized edge versions that were substantially smaller and faster than baseline implementations [8].
3. Microcontrollers: Only highly optimized acoustic tasks will run effectively on ultra-low-power MCUs. Munirathinam et al. demonstrated emergency vehicle siren localization and classification using an STM32F411RE microcontroller (STMicroelectronics, Plan-les-Ouates, Switzerland), which shows that resource-constrained acoustic aware-

ness systems are feasible [52]. Chen et al. also showed sound event detection entirely on a CC1350 MCU via a low-power design [4]. Related compact recurrent models for acoustic event detection have also been studied for low-energy and low-complexity platforms [37].

4. FPGAs: Field-Programmable Gate Arrays (FPGAs) are capable of accelerating signal-processing workloads through flexible hardware acceleration. Silva et al. conducted a design exploration of FPGA-based sound-source localization, resulting in power-efficient architectures targeted toward IoT-style acoustic sensing [32]. FPGAs are particularly well suited to perform parallel operations involved in signal processing, such as beamforming; however, specialized development expertise is typically needed to develop an FPGA solution [32].
5. Hearables/Wearable Devices: Wearable audio devices are subject to extreme size, power, and latency constraints. Yeow et al. proposed a lightweight ASC-conditioned framework for wearable audio devices to support situational awareness via simulated headphone data and lightweight SELD processing [1]. Dementyev et al. presented sub-millisecond algorithmic latencies for real-time speech enhancement on hearable-style hardware, illustrating that extremely low-latency audio processing is possible; however, their task was speech enhancement rather than SELD [49].

In general, hardware acceleration and platform optimization can improve edge acoustic perception; however, these approaches only benefit the edge acoustic perception application if the model and feature pipeline are designed specifically for the target hardware. Therefore, hardware-aware design is essential for practical real-time deployment of edge SELD since it requires consideration of the entire pipeline from microphone configuration to buffering, feature extraction, neural inference, spatial post-processing, memory usage, and power consumption [8,20,48].

6. Robustness in Real-World Environments

Real-world deployments often involve additional issues beyond those experienced in laboratory evaluations, such as background noise, reverberation, and simultaneous events. This subsection examines the challenges associated with robustness in real-world environments and provides potential strategies for addressing them.

6.1. Noise and Interference

Background noise and other forms of interference negatively affect the performance of Sound Event Localization and Detection (SELD) models. Most existing models were designed based upon clean or mildly degraded data and do not generalize well to the wide range of noise found in everyday environments. Many types of background noise exist, including traffic, heating, ventilation, and air conditioning (HVAC), people walking by, weather-related noises, and electronic devices. Additionally, as described earlier, DCASE-style datasets are increasingly incorporating more realistic acoustic properties, including reverberation, multiple background noises at varying levels, moving sources, overlapping events, and various types of directional interference [14,15], all of which create significant difficulties for SELD in uncontrolled acoustic environments. Robustness under these uncontrolled acoustic conditions is particularly important for edge and mobile deployment, where the hardware and microphone constraints discussed in Section 5 also apply [1,19,20].

Robust speech-source localization was demonstrated by Ganguly et al. for use within smartphone-assisted hearing aids via a non-uniform three-microphone array along with a dictionary-based Singular Value Decomposition (SVD) method [2]. Their results are pertinent to this study since they represent a form of robust smartphone-based localization at low signal-to-noise ratios (SNRs) for individual speech sources. However, their

approach does not account for localized SELD of multiple simultaneously occurring events. Munirathinam et al. also created a resource-constrained emergency vehicle siren localization and classification system on an STM32F411RE microcontroller, using GCC-PHAT for localization and Mel-Frequency Cepstral Coefficients (MFCCs) with a neural network classifier for siren identification [52]. While this example demonstrates that embedded acoustic awareness can be achieved on constrained embedded devices, the scope of this effort is much smaller than the overall SELD problem space.

The I-SPOT project focuses on developing acoustic perception for automotive applications which experience strong background noise, dynamically changing acoustic scenes, and rigid placement constraints on available microphones [8]. The I-SPOT project places high priority on robust detection/localization of critical events, low-latency real-time operation, and hardware−algorithm co-design for embedded implementation [8].

Phase-based spatial cues, typically derived through GCC-PHAT- and/or SRP-PHAT-based features, are frequently used in localization tasks because they can be more robust than raw magnitude-based cues in certain noisy or reverberant conditions [28,30]. Yin and Verhelst utilized SRP-PHAT features along with a compact Convolutional Neural Network (CNN)-based model for robust sound-source localization on extremely constrained edge devices and demonstrated that localization can become significantly more hardware-efficient while maintaining comparable performance in noisy and reverberant environments [30]. Nevertheless, the approaches mentioned above remain strictly localization techniques, not full SELD systems, and suffer degradation when noise, reverberation, or low-quality phase estimates corrupt the spatial information contained in the phase-based cues.

Data augmentation during training represents another typical strategy for achieving greater robustness. Shah et al. applied SpecMix data augmentation in a single-stage CRNN architecture to support real-time SELD on Raspberry Pi platforms and reported improved performance [28]. Despite numerous attempts to augment training data to better simulate real-world noise, real-world noise continues to pose significant challenges since the possible sound sources encountered in deployment environments are generally unknown beforehand, their SNRs vary, multiple events can occur concurrently, reverberation effects may be present, and the characteristics of each device's microphone can differ [14,15].

Another challenge for SELD is directional interference, i.e., nearby or surrounding events from non-target classes interfering with events from target classes. Slizovskaia et al. proposed a class-conditioned sound event Direction of Arrival (DOA) estimation model that utilizes a target-class one-hot vector and Feature-Wise Linear Modulation (FiLM) layers to localize only the selected class [27]. According to the authors' experiments on the DCASE 2021 dataset, which contains interfering sound events from non-target classes [14], their model achieved a greater improvement over baseline localization than their unconditional model, indicating that selectively focusing localization efforts on only the target class can potentially help increase robustness against competing interfering sound events from non-target classes [27].

6.2. Reverberation Effects

The presence of reverberation within a space degrades the ability to detect and localize a source. Due to reverberation, the time-domain characteristics of the signal are smeared, making it difficult to determine the onsets and offsets of a source. Furthermore, reverberation introduces secondary indirect spatial cues that can confound the determination of a source's location [14,15]. Edge-SELD systems are especially affected because they often rely on small microphone arrays, limited spatial cues, and low-latency processing requirements [1,19,20].

Prior DCASE-style SELD datasets have increasingly incorporated more realistic reverberant spatial conditions. For example, the DCASE 2019 SELD dataset utilized measured spatial room impulse responses from actual physical locations along with ambient noise; however, most event types were static, and the overall conditions were somewhat constrained [13]. The DCASE 2020 SELD dataset [15] increased diversity in terms of the number of rooms and various reverberant conditions, as well as movement of the sound sources. Continuous DOA trajectories and ambient noise levels were also varied [15]. The DCASE 2021 dataset continued the trend toward greater realism through the inclusion of directional interferers, higher polyphony, and overlapping same-class events under reverberant and noisy conditions [14].

Politis et al. [15]'s DCASE 2020 dataset is particularly relevant for analyzing reverberation in conjunction with moving sources. The dataset consisted of synthetic reverberant spatial sound scenes generated from measured spatial room impulse responses collected in multiple rooms. Static and moving sound events were each created. Moving events followed previously measured DOA trajectories at speeds approximating 10°/s, 20°/s, or 40°/s [15]. Thus, the dataset is more representative of reality than previous fixed-grid static-source datasets.

Reverberation impacts spatial features since reflected sound diminishes the strength of the direct path. Intensity-vector-based features in FOA-based SELD may be negatively impacted if the sound field becomes more diffuse. Similarly, microphone-array-based features like GCC-PHAT may not provide reliable results in complex scenes involving multiple simultaneous events [21]. Additionally, analysis of the DCASE 2021 dataset [14] showed that reverberation consistently increases error rates, reduces F-scores and localization recall, and increases localization errors by approximately 2–4° [14]. Directional interferers were shown to be more detrimental than reverberation alone. Hence, developing robust solutions will need to account for both room acoustics and other localized competing sound sources [14].

The use of beamforming or SRP-PHAT-based localization techniques can be beneficial for emphasizing directional information by utilizing signals across multiple microphones. Belloch et al. examined real-time sound-source localization on embedded GPU platforms using SRP-PHAT with multi-microphone arrays [50]. Similar research has been conducted using embedded GPUs in combination with spherical microphone arrays, demonstrating that denser arrays can provide real-time 3D localization with increased computational demands [33]. Although these prior works addressed sound-source localization techniques rather than complete SELD systems, they did require adequate numbers of microphone channels and appropriate geometric arrangements.

Dereverberation preprocessing may improve SELD performance in highly reverberant environments, but it introduces additional processing overhead. Therefore, its robustness benefits should be evaluated together with the deployment costs described in Section 5.

Training on a variety of reverberant conditions provides improved generalizability. Room impulse responses may be measured or simulated to present models with disparate room geometries, reflectivity patterns, direct-to-reverberant ratio values, and source trajectories. The use of measured room impulse responses and ambient noise from real rooms to create more realistic SELD conditions demonstrates this approach through the creation of the DCASE 2020 and 2021 datasets [14,15]. However, even these datasets may fail to encompass all possible deployment environments, resulting in potential domain-shift issues when the trained model is deployed on edge devices.

6.3. Overlapping Events and Polyphony

Localization and detection of multiple sound events that overlap in time and/or frequency are significant problems in SELD [23,24]. This is especially challenging when more than one instance of the same type of sound exists within the same scene, i.e., when multiple instances of the same event class occur at the same time, since the system needs to detect and localize more than one source, all labeled with the same class label [24].

Activity-Coupled Cartesian Direction of Arrival (ACCDOA) representation provides a solution to Sound Event Localization and Detection (SELD) problems by capturing both sound event activity and direction in a single Cartesian DOA vector. The length of the vector captures the degree of the event's activity, while the vector's direction reflects the estimated DOA of the event. Hence, SELD can be represented as a single regression task instead of having two independent tasks for SED and DOA prediction [25]. A drawback of the original representation is that it is class-wise and assigns only one location to each event class. This makes it difficult to capture simultaneous events of the same class [24].

To overcome this limitation, the Multi-ACCDOA method extends the single-ACCDOA representation by representing multiple ACCDOA vectors per sound class using a class-and-track-wise output format. As a result, the method can accurately locate overlapping sound events that are classified as being from the same class. In addition, Shimada et al. proposed auxiliary duplicating permutation-invariant training (ADPIT), which assists in solving the track permutation issue without losing benefits related to ACCDOA-style training [24]. In DCASE 2021 Task 3 dataset experiments, Multi-ACCDOA with class-wise ADPIT outperformed other methods for overlapping same-class events while retaining similar accuracy under non-overlapping conditions [24].

Another angular-distance-based YOLO-style approach called AD-YOLO was introduced by Kim et al. [26] for polyphonic SELD. Unlike previous approaches that assign class responsibility to a predetermined DOA prediction for each class, AD-YOLO uses the angular distance to assign responsibility for each predicted DOA. By doing so, AD-YOLO can handle both unknown polyphony and overlapping events belonging to the same class [26]. Performance evaluations conducted on DCASE 2020–2022 Task 3 datasets indicated that AD-YOLO performs better than other approaches for class-homogeneous polyphony cases [26].

Overlapping events create additional complexities in terms of edge deployment due to their requirement for sufficient model capacity to represent concurrent sources. While methods like Multi-ACCDOA and AD-YOLO help improve modeling capabilities in polyphonic situations, they may introduce increased output dimensions, inference complexity, or additional processing requirements [24,26]. Thus, when achieving polyphonic capabilities in edge-SELD systems, these systems must trade off polyphonic capability against parameter count, latency, memory usage, and power consumption [19,20,29].

Additionally, both temporal resolution and input context impact performance in polyphonic environments. Processing short hop sizes yields greater temporal detail in quickly changing events. However, this results in processing more frames per second. Longer input contexts provide more history and therefore potentially more accurate context. However, both yield increased buffering and inference costs. Studies by Yeow et al. [20] demonstrated that decreasing input window size negatively impacts the performance of SELD tasks. Additionally, depending on model complexity, some models are capable of handling shorter inputs with less loss of accuracy but do so at higher computational loads [20]. For example, Shah et al.'s Raspberry Pi SELD system utilized a 32 kHz sampling rate, a 1024-sample analysis window, and a 320-sample hop size, demonstrating that feature selection should always consider real-time constraints [28].

6.4. Domain Shift and Generalization

Domain shift, which refers to the disparity between how models are trained and how they are applied, represents an important problem for real-world SELD deployment. Training models on controlled datasets, such as DCASE SELD benchmarks, which may lack realistic acoustic characteristics, microphone configurations, source locations, interfering sound sources, etc., may lead to poor generalization when applied to a variety of deployment scenarios [14,15,20].

The DCASE SELD benchmarks have become increasingly representative of real-world deployments, reducing the disparity between testing and deployment. In the DCASE 2019 dataset, the sound events were primarily stationary and were created using measured room impulse responses to create a sense of space [13]. In contrast, the DCASE 2020 dataset had greater variability in terms of rooms, moving objects/sound sources, continuously changing DOAs over time, and ambient noise collected at the point of recording [15]. The DCASE 2021 dataset introduced new challenges such as directional interferers, increased polyphony, and more instances of same-class overlapping sounds [14]. Recently released STARSS-based datasets have included actual spatial recordings and have been used to train and test real-scene SELD applications and source-distance estimation [22,35,39].

Microphone-array mismatch is an important source of domain shift. As discussed in Section 5.3, SELD systems are commonly developed using standardized FOA or MIC configurations, whereas edge devices may use stereo, smartphone, or wearable microphone layouts [1,2,13,15,19]. Consequently, models developed for one array configuration may generalize poorly to devices with different spatial characteristics [21]. Yeow et al. highlighted the impact of MIC format in their work focused on edge-device application of STARSS23 data and noted its applicability to non-tetrahedral microphone arrangements and future smart-device applications [20]. Kim et al. also noted that stereo SELD is a much more difficult task due to the loss of elevation cues associated with the stereo format; however, it is also a more feasible format for most mobile robots and consumer devices [19].

Similarly to spectral features being affected by changes in acoustic environment, variations in reverberant characteristics, direct/reverberant energy ratio, room geometry, background noise spectrum, source distance, and event distribution can cause significant variation between training data and deployment environments [14,15]. Although the DCASE 2020 dataset was created to increase acoustic diversity by including more rooms, more varied reverberation conditions, and moving sources, there is still evidence that models may not generalize uniformly to unseen rooms, as evidenced by varying performance among unseen evaluation spaces [15].

Data augmentation is frequently utilized as a means to enhance generalizability. Common reported techniques employed in SELD systems include SpecAugment/SpecMix-style masking, mixup-style augmentation, ensemble or multichannel data augmentation, synthetic spatial scene generation, and combinations of real recordings with simulated spatial data [15,25,28,35]. Shah et al. implemented SpecMix augmentation in a Raspberry Pi SELD system to improve training performance [28]. Similarly, several of the top-performing DCASE 2024 systems implemented real STARSS23 recordings along with simulated spatialized data and additional augmentation techniques to mitigate limitations in available real training data [35].

Transfer learning and domain adaptation may provide some relief against domain shift by mapping generalized acoustic representations onto the target SELD domain. While pre-training on large-scale audio datasets may provide generalized acoustic features, fine-tuning on target-domain data may enable the model to learn about the particular characteristics of the deployment microphone or acoustic environment [47]. However, very large pre-trained models may present significant challenges in terms of direct deployment onto

edge devices. Thus, compression, distillation, or lightweight adaptation methods may need to be employed to facilitate deployment onto smaller form-factor devices [20,37,42,43].

On-device adaptation is an area that has not yet been fully explored for edge SELD. A deployed system could theoretically adapt to local room acoustics, microphone characteristics, and persistent sound events through unsupervised or semi-supervised learning. However, on-device adaptation adds computational overhead and memory use. It also introduces risks such as catastrophic forgetting or degraded behavior caused by noisy self-training labels. As such, domain adaptation for edge SELD remains a significant challenge.

Far-field deployment presents an additional robustness challenge for edge and mobile SELD. Sounds occurring at greater distances may have lower signal levels and may be more strongly affected by reverberation, occlusion, background noise, and microphone-array limitations [14,15]. These conditions can also increase domain mismatch when training datasets do not adequately represent the source distances and acoustic conditions encountered during deployment. Therefore, far-field SELD requires improved robustness to distance, noise, reverberation, and microphone constraints while remaining within practical computational constraints for edge and mobile devices [20,48].

6.5. Comparative Summary of Robustness Evidence

The overall results from this review indicate that no single acoustic factor is solely responsible for robustness degradation. Both noise and reverberation can reduce the reliability of spectral and spatial cues, while both polyphony and directional interference cause increased ambiguity as to which source(s) are present. An additional challenge is presented when there is a "domain shift" between the training environment and the operational environment, where room acoustics, microphone configurations, and the distribution of events are different from those used during training. A number of approaches have been shown to potentially help improve robustness, such as data augmentation, improved output representations, class-conditioned localization, and domain adaptation. While some of these techniques have been demonstrated to improve robustness, the costs associated with deploying them have not been reported consistently. Augmentation at training time does not necessarily add to the processing cost at test/inference time. However, some other techniques, such as dereverberation, more complex output representations, larger pre-trained models, and on-device adaptation, can increase the computational resources required for preprocessing, computation, memory use, and/or latency. Therefore, evaluation of robustness improvements for edge SELD should take place in conjunction with the deployment cost of the approach instead of solely using SELD accuracy measures. The main robustness conditions, mitigation approaches, and associated deployment costs are summarized in Table 5.

Table 5. Summary of robustness conditions, reported SELD performance trends, mitigation approaches, and deployment-related costs in the reviewed studies.

Robustness Condition	Representative Studies	Representative Evidence	Reported Performance Trend	Deployment Cost/Limitation
Background noise	Ganguly et al. [2]; Shah et al. [28]; Yin and Verhelst [30]	Ganguly et al. [2] reported a mean RMSE of approximately 2.29° at −5 dB SNR for speech-source localization. Shah et al. [28] reported improved SELD performance using SpecMix. Noise can still degrade detection and localization when spatial cues become unreliable.	Training-time data augmentation; phase-based spatial features such as GCC-PHAT and SRP-PHAT [28,30]	Training-time augmentation does not necessarily add inference cost. However, spatial-feature extraction still contributes to the total processing cost of edge systems.

Table 5. *Cont.*

Robustness Condition	Representative Studies	Representative Evidence	Reported Performance Trend	Deployment Cost/Limitation
Reverberation	DCASE 2020/2021 [14,15]	Reverberation negatively affected error rate, F-score, and localization recall, while increasing localization error by approximately 2–4° [14].	More robust spatial-feature extraction and/or dereverberation	Dereverberation adds another preprocessing stage and may increase computation and latency, which is important for edge deployment.
Polyphony and overlapping events	Multi-ACCDOA [24]; AD-YOLO [26]	Multi-ACCDOA showed better performance than comparison methods for overlapping same-class events while maintaining similar performance in non-overlapping conditions. AD-YOLO also performed better for class-homogeneous polyphony across DCASE 2020–2022.	Multiple-location output representations such as Multi-ACCDOA and AD-YOLO	More complicated output representations may require additional processing. Sufficient edge latency, memory, and power-consumption data were generally not reported.
Directional interference	DCASE 2021 [14]; Slizovskaia et al. [27]	Directional interference was more detrimental to SELD performance than reverberation alone [14]. Class-conditioned DOA estimation showed greater improvement over the baseline than the unconditional version under directional interference [27].	Class-conditioned DOA estimation using a target-class vector and FiLM layers [27]	This approach adds additional conditioning to the localization process, but additional edge latency, memory use, and power consumption were not reported in the reviewed material.
Domain shift	[1,2,14,15,19,20, 28,35,47]	Performance can still degrade in unseen rooms or when microphone configurations and other deployment conditions differ from training conditions.	Data augmentation, real + simulated spatial data, pre-training, fine-tuning, and domain adaptation	Training-time augmentation may add little or no inference overhead. Large pre-trained models can increase model size, computation, and memory requirements, while on-device adaptation can add processing and memory costs during deployment.

7. Applications

Real-time SELD capabilities in edge and mobile devices enable an extensive variety of applications, including assistive technologies, smart environments, automotive safety, and immersive experiences. In this section, we discuss major application domains along with their respective requirements.

7.1. Environmental Sound Awareness

Environmental sound awareness systems utilize acoustic detection and localization techniques to inform users of their surroundings. These types of systems provide a significant advantage when visual attention is limited (e.g., reading or driving), users wear noise-canceling headphones, and/or sound events provide critical information related to user safety.

Yeow et al. designed a lightweight ASC-conditioned SELD framework to support improved situational awareness in wearable audio systems by developing a system that first detects the acoustic environment, or "scene," and then uses the detected scene to enable the SELD portion to detect and locate audible environmental sound events, including bicycle bells, car horns, sirens, babies crying, dogs barking, and knocking on doors [1]. This work was conducted using simulated headphone data and demonstrates the potential

of wearable-oriented SELD; however, additional steps are required before deploying this technology on commercially available wearable hardware [1].

Vehicle emergency detection has significant implications for public safety. Munirathinam et al. developed a low-cost system for recognizing siren sounds and determining the location of an approaching emergency vehicle using an STM32F411RE microcontroller [52]. To accomplish these tasks, they employed multiple digital microphones, GCC-PHAT-based sound localization methods, Mel-frequency cepstral coefficients (MFCCs) for feature extraction, and a machine-learning classification model to identify siren sounds and determine the direction of the sound source. Although this work provides evidence of how acoustic awareness can be integrated into low-power hardware for roadway safety applications, its focus is limited to detecting siren sounds and providing directional information about them [52].

I-SPOT research addresses real-time acoustic perception for automotive applications and includes detecting, locating, and tracking various types of audible environmental signals, such as car horn noises and emergency vehicle sirens [8], while addressing several automotive-specific issues, including high levels of ambient noise, overlapping sounds from multiple directions, changing acoustic environments due to traffic patterns or vehicle movement, constrained sensor placement, and low-power requirements for embedded processing and memory resources. The primary emphasis of I-SPOT is to design efficient low-footprint algorithms for automotive acoustic awareness and develop corresponding hardware designs that optimize performance with respect to size, weight, and cost [8].

In addition to automotive applications, distributed acoustic sensing may be used in smart city/Internet-of-Things applications to monitor environmental factors. Wang et al. have proposed an edge-computing-based cooperative sound event detection method utilizing multi-sensor platforms to reduce latency associated with cloud computing [5]. In their method, each acoustic device communicates directly with an edge server to classify and localize unusual sounds. While this is not full SELD, Wang et al.'s method is most closely aligned with cooperative edge-based sound event detection and localization and supports distributed environmental sound awareness [5].

Public-safety monitoring is another potential application of edge and mobile SELD. Systems capable of detecting and localizing emergency sirens, distress calls, impacts, structural-collapse sounds, gunfire, or vehicle crashes could support early warning and emergency response applications [5,8,52]. In search-and-rescue scenarios, mobile or wearable systems could also potentially assist in locating audible signals from trapped or injured individuals, such as shouting or knocking.

7.2. Assistive Hearing Devices

Devices such as hearing aids, hearables, and assistive listening systems are able to take advantage of both acoustic localization and SELD-like environmental awareness. Spatial location information can help users who experience some degree of hearing loss improve their ability to enhance speech, perform beamforming, select sources, and become aware of important background sounds.

Ganguly et al. provided a study of non-uniform and nonlinear microphone arrays designed for enhancing robust speech-source localization in smartphone-assisted hearing-aid applications [2]. Ganguly's use of a three-element L-shaped array, which utilizes smartphone microphone geometries, supports 360-degree DOA scanning and reduces spatial aliasing, ambiguity, and confusion associated with other approaches. The results of Ganguly's simulations showed a mean RMSE (root mean square error) of approximately 2.29 degrees at an SNR of −5 dB. It appears, however, that Ganguly's focus is primarily centered upon performing single speech-source localization as opposed to full SELD. In

addition to providing evidence of the feasibility of this localization application for real-time operation, Ganguly reports that his algorithm performs within a total processing time of approximately 9.4 milliseconds per 20-millisecond frame when run on a 2 GHz Intel processor (Intel Corporation, Santa Clara, CA, USA) [2].

Küçük has performed research on sound-source localization for the purpose of improving hearing-aid studies through mobile platforms [3]. Küçük provides evidence that smartphones and tablets can provide useful mobile platform solutions for hearing-aid research, including real-time deep-learning-based DOA estimation with two microphones, DOA estimation with an eight-microphone circular array, and single-microphone distance estimation [3]. This paper may be considered mobile sound-source localization and distance estimation but does not represent full SELD.

Pertilä et al. studied mobile microphone-array speech detection and localization performance across various everyday environments [53]. The Pertilä research is directly applicable to mobile and assistive audio systems since it examines the robustness of speech detection and localization utilizing a mobile-phone-format microphone array in diverse everyday environments. While Pertilä's work provides insight into how mobile microphone arrays can detect and localize speech, it does not address full SELD.

Spatial acoustic signal processing can enable hearing-assistance capabilities by supporting beamforming, source selection, and suppression of unwanted interference. Full SELD deployment in hearing aids and hearables continues to be challenging due to limited resources: extreme latency requirements for active noise control and hear-through processing, the extremely limited amount of processing hardware available in hearable form factors, and severely restricted battery life [7]. Additionally, previous low-latency work on speech enhancement in hearables shows that sub-millisecond algorithmic latency and low end-to-end latency are achievable with customized designs; however, memory and compute budgets continue to remain heavily restricted [49].

7.3. Smart Home and IoT

Smart homes and IoT systems can benefit from SELD and similar acoustic sensing methods, which can enhance voice interfaces, improve safety notifications, support context-aware automation, monitor acoustic environments, and detect important sounds. Because distributed edge processing can reduce data-processing latency, limit raw-audio transmission, and improve user privacy by processing data at or near the source of the audio, it is highly desirable for smart-home and IoT applications.

Chen et al. illustrated that sound event detection could be accomplished with one microcontroller unit (MCU) using a low-power, portable, and lightweight design [4]. The researchers used LTSD feature extraction and an SVM classifier to perform SED on a CC1350 microcontroller without an external analog front end. Chen et al. achieved onboard recognition using only 33.4 kB flash memory, 9.94 kB SRAM, and 23.1 mW of power consumption [4]. As such, this research is applicable to battery-powered acoustic sensors in smart-home and IoT applications. However, the research only accomplishes SED instead of full SELD.

DiPassio et al. researched embedded machine learning for sound event detection/classification and acoustic sensing in the built environment, framing acoustic sensing as a component of "smart acoustic spaces" [54]. As such, DiPassio et al.'s research supports the utilization of low-power embedded models to monitor indoor environments, buildings, and other acoustic spaces and to utilize local event recognition to enable automation and situational awareness [54].

Lee et al. developed a system capable of predicting a resident's activity of daily living (ADL) in real time via edge-based time-series ambient sound recognition [6]. Such edge-

based acoustic recognition is relevant to both elderly care and smart-home automation because it has the potential to locally infer contextual activity information, thereby reducing reliance upon cloud processing [6].

Yarkan proposed a lightweight acoustic direction-finding framework for edge-AI, IoT, and smart sensing devices [51]. Yarkan's framework was implemented using two MAX4466 microphone-preamplifier modules (Analog Devices, Inc., Wilmington, MA, USA) on an STM32F429 microcontroller with Arm CMSIS-DSP optimized bare-metal processing, demonstrating real-time applicability on low-cost embedded hardware [51]. Although Yarkan's framework does not accomplish full SELD, the results are applicable to spatial acoustic awareness in IoT devices.

Wang et al. researched learning-based cooperative sound event detection using edge computing [5]. Their framework utilizes multiple acoustic devices and an edge server to support abnormal sound event detection and localization in city-surveillance scenarios, reducing processing delay compared with remote cloud processing. While this framework is better described as cooperative edge-based SED/localization rather than full SELD, it is highly relevant to distributed environmental acoustic monitoring in smart-city, campus-scale, and urban-surveillance applications [5].

Roy et al. researched TinyML-driven sensor nodes for energy-efficient acoustic event detection within pervasive acoustic wireless sensor networks [55]. Roy et al.'s approach utilizes resource-limited TinyML models deployed on acoustic sensor nodes, transmits only relevant detected events, reduces communication overhead, and improves energy efficiency. Roy et al.'s research provides evidence of how distributed acoustic event detection may scale to large wireless sensor networks with low energy budgets [55].

7.4. Extended Reality and Spatial Audio Interfaces

Extended Reality (XR) has been defined as a broad category encompassing virtual reality (VR), augmented reality (AR), and mixed-reality (MR) technologies. More and more Extended Reality applications are utilizing spatial audio technology to generate immersive and reactive experiences. As XR continues to grow as a field, edge-based SELD will continue to provide relevant solutions for XR use cases. Specifically, XR can use SELD because detected and localized real-world sounds can be transformed into spatial audio cues, visual alerts, or haptic feedback.

In addition to providing relevant solutions for XR, Gupta et al. reviewed augmented and mixed-reality audio for hearables. They focused specifically on sensing, real-sound control, and virtual-sound rendering, and provided a comprehensive framework to describe how hearables collect the acoustic properties of real sounds, such as frequency content, spatial location, timbre, etc., and combine them with virtual-sound information to produce the auditory aspects of AR/MR experiences [7]. Thus, Gupta's work provides a foundation for understanding how SELD could be used to enhance the sensory components of XR experiences.

Kealey et al. evaluated a wearable device consisting of an 8-microphone array and a camera mounted on glasses. When paired with an embedded system and headphones, the device enhanced video−audio streams in real time by detecting and identifying a specific speaker via face detection, tracking, and verification, followed by beamforming and neural post-filtering to improve the quality of the identified speaker's voice in noisy environments [34]. Although Kealey's work did not evaluate SELD directly, it demonstrated a viable approach to wearable audio−visual pipelines applicable to assistive-listening technologies and spatial-audio interfaces in AR-type systems.

Jo et al. proposed EgoSep, an egocentric on-screen sound-source separation system for real-time edge computing [56]. Jo et al.'s EgoSep system integrated both visual appear-

ance/motion data and audio data to enable on-screen sound sources to be separated from off-screen ambient noise. In evaluating EgoSep's feasibility for real-time implementation on an NVIDIA Jetson Nano processor, Jo et al. found that EgoSep could operate at an average rate of 60 Hz with an overall *RTF* of 0.19 [56]. Although EgoSep is primarily an example of audio−visual sound-separation technology rather than true SELD technology, Jo et al.'s findings show how egocentric visual perspective can be combined with auditory scene analysis, which is relevant to XR-style systems.

Gurvich et al. [36] developed an active-speaker-detection system that utilized a combination of microphone-array audio data, 360-degree visual inputs, and spatial query mechanisms. Gurvich et al.'s system is designed to run efficiently on low-power edge-computing platforms and was originally designed for teleconferencing devices; therefore, the system detects active speakers in meetings, thereby supporting "virtual cinematography" [36]. Although Gurvich et al.'s system was primarily designed to determine who among multiple speakers in a meeting is actively speaking during a conversation, their research illustrates the potential for combining audio−visual spatial querying with real-time multimodal interfaces.

The amount of latency allowed in XR and hearable audio depends upon the processing block and the intended perceptual goals. Gupta et al. noted that real-sound-control tasks like active noise cancellation (ANC) and hear-through require extremely tight latency constraints, whereas virtual-sound-rendering tasks allow longer total system latencies depending upon the tasks being performed [7]. Kealey et al. estimated the total latency associated with their wearable audio/video enhancement prototype to be approximately 120 ms [34], which they report is sufficient for their prototype's speech-enhancement function; however, this total latency value may be too long for several tightly coupled XR interactive tasks. Consequently, XR-oriented edge-based SELD will need to be implemented according to the appropriate use case, while considering factors such as detection accuracy, localization accuracy, latency, power consumption, and multimodal presentation.

8. Datasets and Benchmarks

8.1. DCASE SELD Datasets

The Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge has developed a common benchmark series for SELD through the Task 3 evaluations. The DCASE datasets have been continually updated since their inception to improve acoustic realism, polyphony, spatial complexity, and task difficulty.

In Politis et al. [13]'s review and analysis of sound event localization and detection in DCASE 2019 [13], the authors reported on a benchmark for joint sound event detection and localization based on spatialized sound events. The challenge employed multiple formats, including both First-Order Ambisonics (FOA) and microphone-array (MIC) formats. The official DCASE 2019 challenge evaluated sound event detection using error rate and F-score, while localization was evaluated separately using DOA error and frame recall [13].

Compared with DCASE 2019, the DCASE 2020 dataset [15], described by Politis et al., increased acoustic realism by utilizing additional types of rooms, spatially measured room impulse responses, continuous DOA trajectories, stationary and moving sound events, ambient noise fields recorded on site, and both FOA and MIC four-channel formats. Therefore, this dataset provided a more difficult benchmark for dynamic SELD compared to the mostly static-source benchmark of the earlier dataset.

The DCASE 2021 dataset [14] further increased the realism and difficulty of the task by incorporating directional interferers, increasing polyphony, and incorporating more same-class overlapping sound events. Therefore, the task became more representative of actual real-world acoustic environments in which sounds outside the desired classes

can affect the detection and localization of desired events. An analysis of the DCASE 2021 dataset demonstrated that directional interferers were particularly detrimental to the performance of SELD tasks [14].

DCASE 2022 continued this trend toward increased realism of spatial-sound scenes. Later work such as AD-YOLO examined output representations across the DCASE 2020, 2021, and 2022 Task 3 datasets [26]. Additionally, Kim et al. demonstrated that AD-YOLO produced better results than several other previous SELD output formats in class-homogeneous polyphony cases, thereby providing relevance to the evolving benchmark focus on overlapping and same-class events [26]. Furthermore, class-conditioned sound event localization was investigated on DCASE 2020 [15] and DCASE 2021 [14]; however, especially significant improvements were shown on the DCASE 2021 dataset when considering directional interference [27].

Finally, DCASE 2024 expanded the SELD task by incorporating source-distance estimation; thus, systems estimate not only the sound class and direction of arrival of each active sound source, but also the distance from the recording device to each source [35]. This extended task is commonly denoted as SELD with Source-Distance Estimation (SELDDE). Yeow et al. evaluated Squeeze-and-Excite ResNet-Conformer systems on the DCASE 2024 audio-only SELDDE task using the STARSS23 dataset and DCASE 2024 metrics that include source-distance estimation [35]. Similarly, Dong et al. presented a study of joint modeling for SELD with source-distance estimation using STARSS23-based real spatial recordings [22].

Although DCASE benchmarks provide standardized measures of SELD accuracy, they are not specifically designed to evaluate complete edge-deployment performance. Deployment-oriented evaluation is discussed in Section 8.4 [20]. Table 6 summarizes the differences among the DCASE benchmarks and evaluation protocols discussed above [13–15,26,35].

Table 6. Comparison of evaluation protocols across relevant DCASE SELD Task 3 editions.

DCASE	Task Formulation	Main Evaluation Metrics	Matching/Spatial Tolerance	Submission/ Output Format	Source Distance
DCASE 2019	Joint SED + DOA estimation on spatialized scenes	ER, F-score; DOA error, frame recall	SED and localization evaluated separately; frame-wise Hungarian assignment for DOA estimates; no fixed angular TP threshold	Frame, class, azimuth, elevation; 20 ms	No
DCASE 2020	SELD with static/moving and overlapping sources	ER20°, F20°, LE_CD, LR_CD	Joint class/location-dependent evaluation; 20° tolerance for ER/F; LE/LR unthresholded	Frame, class, azimuth, elevation; 100 ms	No
DCASE 2021	SELD with directional interferers and increased overlap	ER20°, F20°, LE_CD, LR_CD	Same joint framework as 2020; 20° tolerance for ER/F	Frame, class, azimuth, elevation; 100 ms	No
DCASE 2022	SELD in real spatial sound scenes	ER20°, F20°, LE_CD, LR_CD	Per-class Hungarian matching; 20° threshold for ER/F; F-score macro-averaged	Frame, class, azimuth, elevation; 100 ms	No

Table 6. *Cont.*

DCASE	Task Formulation	Main Evaluation Metrics	Matching/Spatial Tolerance	Submission/ Output Format	Source Distance
DCASE 2024	Audio/audio −visual SELD with source-distance estimation (SELDDE)	F20°/1, DOAE, RDE	Per-class Hungarian matching; DOA $\leq 20°$ and relative-distance error ≤ 1 for F-score TP; frame-based evaluation	Frame, class, azimuth, elevation, distance; 100 ms	Yes

Note: SED, Sound Event Detection; SELD, Sound Event Localization and Detection; DOA, Direction of Arrival; ER, Error Rate; F-score, F1-score; LE_CD, class-dependent Localization Error; LR_CD, class-dependent Localization Recall; DOAE, Direction-of-Arrival Error; RDE, Relative Distance Error; TP, true positive; ms, milliseconds. ER20° and F20° denote location-dependent ER and F-score evaluated using a 20° spatial threshold; F20°/1 additionally incorporates the source-distance criterion used in DCASE 2024.

Results from different DCASE editions are not necessarily directly comparable, since each edition's dataset, task formulation, matching methodology, metric definition, averaging strategy, and reporting format have changed across editions. Specifically, DCASE 2019 separately evaluated detection and localization; DCASE 2020 introduced joint location-dependent metrics, as well as a spatial threshold of 20 degrees; DCASE 2022 altered its method for calculating the average F-score; and DCASE 2024 introduced source-distance estimation along with frame-based evaluation.

8.2. TAU-NIGENS Spatial Sound Events

The TAU-NIGENS Spatial Sound Events datasets were produced by the audio research team at Tampere University. They provide spatial audio information for developing and evaluating SELD systems. Additionally, the datasets have been utilized within previous versions of the DCASE Task 3 benchmarks. These datasets are important because they provide controlled but increasingly realistic spatialized sound environments containing labeled sound events, time-activity labels, and direction-of-arrival (DOA) annotations.

The TAU-NIGENS Spatial Sound Events 2020 dataset was selected for DCASE 2020 Task 3 and provided a new type of reverberant spatialized sound environment that included moving sources [15]. The dataset was created using samples of sound events from the NIGENS database. Additionally, the dataset contained measured spatial room impulse responses from various rooms and ambient background noise collected at the point of recording. The dataset contains both stationary and moving sound events and provides continuous DOA trajectories for each sample. Two types of spatial formats were applied to all samples, including first-order ambisonics (FOA) and microphone-array (MIC) format [15].

Each sample within the TAU-NIGENS Spatial Sound Events 2020 dataset is one minute long. The dataset is typically accessed via the DCASE development set. A common method for accessing these samples involves utilizing subsets of the recordings; for example, one study utilized subsets of 400 recordings for training, 100 recordings for validation, and 100 recordings for testing [26]. Marhoff et al. utilized the TAU-NIGENS Spatial Sound Events 2020 dataset to develop and test a machine learning-based approach to estimating the DOA of a source sound based upon beamforming maps and CNN-based source-map recognition [41]. However, this approach only estimates the DOA of a source sound without providing full SELD classification [41].

The TAU-NIGENS Spatial Sound Events 2021 dataset extended the original 2020 dataset by adding directional interferers, increased polyphony, and more same-class overlapping sound events [14]. The addition of these elements made the dataset more difficult than the original dataset and more reflective of what occurs in real-world SELD applica-

tions, i.e., there can be non-target sounds in close proximity to the target sound event(s). Slizovskaia et al. [27] utilized the TAU-NIGENS 2020 and 2021 datasets to develop a class-conditioned sound event DOA estimator. Their results demonstrated greater improvements with the 2021 dataset when compared to the original condition under directional interference [27].

The TAU-NIGENS datasets allow researchers to utilize different representations of spatial audio. DCASE-style SELD systems often use log-mel spectrograms and intensity-vector features with FOA input, while MIC input is typically paired with log-mel spectrograms and GCC-PHAT features [15]. As such, the datasets are valuable tools for comparison among different spatial formats and/or representation methods. However, results from experiments may not translate well to devices equipped with different microphone configurations, e.g., smartphones, hearables, etc.

TAU-NIGENS/DCASE data have also been used in edge-oriented evaluations; for example, Shah et al. evaluated a single-stage CRNN on Raspberry Pi using DCASE 2020 data [15,28]. However, deployment-specific measurements must be evaluated separately, as discussed in Section 8.4.

8.3. STARSS and L3DAS

In addition to the above-mentioned developments that enhance the realism and complexity of acoustic environments to reflect what happens in the field, we are now able to develop datasets that much more closely approximate real-world acoustic environments. The use of STARSS-type datasets represents an effort to make SELD evaluations even more similar to what occurs in real acoustic spaces. Additionally, L3DAS datasets allow researchers to obtain large-scale 3D audio data for both speech enhancement and related SELD research.

As noted earlier, the STARSS23 dataset (Sony-TAU Realistic Spatial Soundscapes 2023) [35] includes real spatial sound recordings that were recently used in several DCASE SELD benchmarks. For example, in DCASE 2024 Task 3, the SELD task included source-distance estimation in addition to estimating the location of sound sources in terms of direction of arrival and activity. Therefore, SELD systems needed to estimate not only the active sound event class and direction of arrival, but also the distance between the recording device and the active sound source [35]. Yeow et al. tested Squeeze-and-Excite ResNet-Conformer models on STARSS23 for this audio-only SELD with source-distance estimation task [35].

In their studies, Dong et al. assessed the performance of joint modeling approaches for sound event detection (SED), localization (DOA), and source-distance estimation using the DCASE 2024 Task 3 configuration [22]. Dong et al.'s SED-DOA + SED-SDE approach trained DOA and SDE independently prior to generating final predictions and produced a development-set SELD score of 0.237 without model ensemble methods; the paper also reported that the proposed method was ranked #1 among all participants in the DCASE 2024 Task 3 evaluation [22]. These results emphasize the need to move SELD beyond directional-only localization into distance-aware 3D spatial understanding.

L3DAS datasets, although somewhat older than some other databases being discussed here, provide yet another important source of 3D audio data. L3DAS21 and L3DAS22 are focused on both 3D speech enhancement and 3D SELD. Both datasets contain Ambisonic audio generated by convolving either speech or sound event samples with impulse responses recorded using two First-Order Ambisonic microphones [47]; these datasets contain both a speech-enhancement subset and a SELD subset. The SELD subsets of these datasets contain both SED and DOA annotations, while later work has used the speech-enhancement subsets for unsupervised or self-supervised training of spatial-audio pre-training models [47].

The w2v-SELD study employed both L3DAS and TAU Spatial Sound Events datasets for self-supervised spatial-audio pre-training [47]. In their approach, the authors adapted wav2vec 2.0 to multichannel spatial audio and pre-trained on unlabeled 3D audio prior to fine-tuning on labeled SELD datasets. As stated previously, this type of self-supervised learning is one way in which future SELD dataset designers can potentially create fewer dependencies on time-consuming manual SELD labeling efforts [47].

Although these datasets increasingly reflect realistic acoustic conditions, deployment readiness requires additional hardware-oriented evaluation beyond dataset-based SELD accuracy, as discussed in Section 8.4 [20].

8.4. Evaluation Protocols for Edge Deployment

Current SELD benchmarks mainly focus on accuracy metrics, such as error rate, F-score, localization error, localization recall, and SELD score [13]. Nonetheless, those metrics do not fully describe a system's level of deployment readiness on edge and mobile devices [20]. Yeow et al. identified a gap within current metrics in relation to real-time SELD deployment and demonstrated that real-time SELD deployments require assessment of the entire processing chain, which includes audio buffering, feature extraction, neural network inference, and post-processing, all running on target hardware [20]. Belloch et al. likewise stated that practical acoustic localization for IoT devices needs to take into account real-time performance, energy efficiency, use of multiple processors on hardware, and other limitations specific to each platform [48].

The dimensions of evaluation protocols specifically designed for edge deployment should include the following:

1. Computational Metrics: In order to assess the suitability of edge devices for certain models, it is essential to report model size, number of parameters, memory footprint, computational cost, time required for feature extraction, and time required for inference, in addition to the usual accuracy metrics. The importance of assessing these computational aspects stems from the fact that edge devices are often limited by memory bandwidth, the computational costs associated with feature extraction, and the availability of accelerators [19,28,48]. For example, Kim et al. provided empirical evidence related to the embedded real-time capabilities of MMT-Net via demonstrations on two NVIDIA Jetson platforms, namely, the Xavier NX and Orin AGX [19]. Additionally, Shah et al. assessed a single-stage CRNN for SELD on Raspberry Pi and reported *RTF* values, demonstrating the merits of conducting experiments directly on the targeted hardware [28].
2. Latency Measurements: To assess end-to-end latency for SELD systems, we recommend measuring latency throughout the entire process, from when audio is captured through SELD output, including buffering, feature extraction, inference, and post-processing [20]. A widely used measure of processing speed is the real-time factor (*RTF*). An *RTF* below 1.0 indicates that the measured processing time was shorter than the duration of the evaluated audio under the reported test conditions [28]. As discussed in Section 5.1, however, this alone does not establish low-latency causal streaming or end-to-end deployment readiness. As such, latency should be measured on the target hardware using realistic streaming conditions because fast offline inference does not necessarily mean that the system can operate in real time during deployment [20].
3. Power Consumption: Assessing power consumption/energy usage is vital for devices powered solely by batteries, such as hearables, smartphones, and IoT nodes, since power consumption affects both longevity and functionality [3]. Küçük has performed studies to measure power consumption for a mobile sound-source localization proto-

type [3]. Likewise, Silva et al. have described low-power FPGA-based sound-source localization architectures suitable for IoT-type deployment [32]. While neither study represents a complete SELD system, they demonstrate why evaluation protocols for edge deployment need to incorporate measures of energy usage in conjunction with accuracy assessments.

4. Robustness Assessment: An assessment of the robustness of edge-SELD systems is necessary across a wide variety of acoustic environments, including varying signal-to-noise ratio (SNR) levels and reverberation conditions; various backgrounds and types of background noise; differing angular orientations of interfering sounds; varying distances between sources and microphones; and varied microphone configurations. DCASE 2020 and 2021 datasets contain numerous examples of increasingly realistic representations of acoustic environments that can be utilized for evaluation purposes. Examples include moving sources, reverberant environments, ambient noise, directional interference, and increased polyphony [14,15]. Nonetheless, robustness assessments also need to occur on the specific edge device/microphone configuration being deployed due to potential differences between deployment conditions and those represented in benchmark datasets [20].
5. Constraints of Microphone Arrays: Edge deployment assessment protocols need to assess systems utilizing microphone configurations that represent actual edge devices, such as stereo microphones, smartphone arrays, and headphone-mounted microphones [2,19]. The reason is that performance in FOA four-channel or MIC tetrahedral formats may not translate directly to performance in compact consumer devices.

9. Comparative Summary of Reviewed Edge/Mobile Acoustic Systems

To evaluate the reviewed articles comparatively, this section summarizes the systems discussed in Sections 3–8. Table 7 provides a compact synthesis of complete SELD, SELDDE, and SELD-like detection-localization systems by grouping them according to system category, representative studies, deployment evidence, and main limitation or takeaway. Related edge/mobile acoustic systems that are not complete SELD systems are summarized separately in Table 8. A more detailed study-level comparison, including input format, microphone setup, features, model architecture, datasets, evaluation metrics, hardware setting, model size, computation/latency values, and memory/power reporting, is provided in Supplementary Tables S1 and S2.

Table 7. Compact synthesis of deployment readiness for complete SELD, SELDDE, and SELD-like systems.

System Category	Representative Studies	Deployment Evidence	Main Limitation/Takeaway
Automotive SELD-like acoustic perception/MCUs	I-SPOT [8]	For deployment, the authors report hardware-flow results on a Raspberry Pi 4B. The reported results include 8.59 ms/frame, a 7.26× speedup, an approximately 10× latency improvement, and an approximately 50% reduction in coefficients. The authors also report an optimized edge-device version that is approximately 86% smaller and approximately 47% faster than the previous version.	A main limitation is that although the study provides strong evidence for edge deployment, it is not a full DCASE-style SELD benchmark.

Table 7. *Cont.*

System Category	Representative Studies	Deployment Evidence	Main Limitation/Takeaway
Lightweight embedded SELD	Kim et al., MMT-Net [19]	The model has approximately 2M parameters and uses nearly seven times fewer parameters than Conformer-based models. The authors also report that the model runs at more than 20 frames per second on both NVIDIA Jetson Xavier NX and NVIDIA Jetson Orin AGX devices.	The study provides clear evidence that the model can be deployed in an embedded setting. However, because it uses only two microphones, it may have less spatial information than systems that use larger microphone arrays.
Foundational and model-efficient SELD architectures	Adavanne et al., SELDNet [23]; Brignone et al., QSELD-TCN [29]; Guirguis et al., SELD-TCN [31]	SELDNet is mainly used as a desktop/benchmark baseline. QSELD-TCN reports reduced trainable parameters, storage requirements, runtime memory use, and inference time. SELD-TCN reports parameters, training time, inference time, and MACs.	These studies are useful for showing the shift from foundational SELD models toward more efficient architectures. However, most of the reported results are still model-level or desktop-based rather than full end-to-end edge-deployment evaluations.
Edge-SELD pipeline studies	Yeow et al. [20]; Shah et al. [28]	Shah et al. reported Real-Time Factor (*RTF*) values ranging from 0.50 to 0.70 on a Raspberry Pi. Yeow et al. identified feature extraction and neural-network inference as significant contributors to latency in an edge-SELD pipeline.	These studies show that edge SELD is a system-level problem rather than solely a neural-network inference problem.
Wearable-specific SELD	Yeow et al. [1]	The authors report results from compact versions of both the ASC and SELD modules. They also report computational complexity in MACs and latency breakdowns, with 13.1 ms of ASC latency and 38.0 ms of SELD latency.	Although promising for wearable situational awareness, this work has been evaluated through simulation-based or wearable-oriented experiments rather than by deployment as a hearable system.
Spatial-feature SELD	Nguyen et al., SALSA-Lite [21]	The authors report computational time for NIPD as 0.30 s, which is 30 times faster than full SALSA and nine times faster than MELSPECGCC.	While the authors present very strong results regarding the efficiency of their approach, they do not report all key performance parameters for an end-to-end system deployed on edge hardware, such as latency, memory usage, and power consumption.
Polyphony and overlapping-event SELD output representations	Shimada et al., Multi-ACCDOA [24]; Kim et al., AD-YOLO [26]; Shimada et al., ACCDOA/RD3net [25]	The majority of these works were evaluated using common DCASE-style SELD metrics such as ER20°, F20°, LECD, LRCD, and SELD score.	These approaches improve performance in complex acoustic environments. However, although improved performance is observed when compared to other approaches, limited data are reported regarding the deployment costs associated with these approaches, such as edge latency, memory usage, and/or power consumption.
Distance-aware SELDDE systems	Yeow et al., SE ResNet-Conformer SELDDE [35]; Dong et al. [22]; Mu et al., SELD-Mamba [40]	These systems were typically evaluated using SELDDE-style metrics such as F20°/1, DOAECD, RDECD, ESELDDE, and/or SELD score.	While these works are important for achieving 3D spatial awareness in audio processing, there is still limited deployment evidence showing how well these approaches can run on edge devices. This is mainly because information related to the computational resources used during deployment is often not reported.

Table 7. *Cont.*

System Category	Representative Studies	Deployment Evidence	Main Limitation/Takeaway
Self-supervised and alternative SELD-like modeling	Dos Santos et al., w2v-SELD [47]; He et al., SoundDet [46]; Rosero et al., G-SELD [38]; Shul et al. [39]	These studies demonstrate some degree of accuracy improvement or real-time potential in benchmark/system evaluations.	These approaches are useful for future SELD design, but deployment-related information such as edge latency, memory use, power consumption, and *RTF* is often missing.

Table 8. Compact synthesis of related edge/mobile acoustic systems that are not complete SELD systems.

System Category	Representative Studies	Deployment Evidence	Main Limitation/Takeaway
Edge and embedded sound-source localization systems	Yin and Verhelst [30]; Belloch et al. [33,48,50]; Yarkan [51]; Silva et al. [32]	There is some evidence related to the deployment of these systems. The reported evidence includes extreme-edge localization models, embedded Internet of Things (IoT) localization frameworks, real-time localization on mobile SoCs, STM32-based direction finding, and FPGA-based localization with reported power consumption data.	Although the systems in this entry are highly relevant to the localization component of SELD, they do not demonstrate full sound event detection and classification together with localization.
Mobile and hearing-aid-oriented localization systems	Ganguly et al. [2]; Pertilä et al. [53]; Küçük [3]	Examples of deployment evidence include evaluations of mobile-phone microphone arrays, tests of speech detection and localization on mobile devices, Android smartphone/tablet prototypes for room-level sound localization, and reports on direction-of-arrival and distance estimation using mobile-device microphone configurations.	While these studies provide useful insight into what can be done with a limited number of microphones on mobile devices, they mainly focus on speech or localization. They do not demonstrate the ability to detect and classify a wide range of general sound events.
Embedded systems for low-power sound event detection and acoustic recognition	DiPassio et al. [54]; Cerutti et al. [37]; Chen et al. [4]; Roy et al. [55]; Lee et al. [6]; Cerutti et al. [42]	Examples include low-energy acoustic event detection, SED using an MCU, distributed acoustic sensor nodes, edge-based acoustic recognition, and binary-neural-network-based sound event detection.	This work is important for providing the detection component of a SELD system. However, it does not provide a solution for estimating source direction. Thus, it is not a complete SELD system.
Cooperative/distributed edge acoustic sensing	Wang et al. [5]; Roy et al. [55]; Belloch et al. [48]	Wang et al. demonstrate an edge-server solution that provides lower latency than traditional cloud processing. Roy et al. investigate low-cost and low-resource TinyML acoustic sensor nodes. Belloch et al. investigate real-time performance and energy efficiency for embedded SoCs.	While these studies contribute to distributed acoustic awareness, they cannot be considered standard full SELD benchmark systems.
Audio-visual and wearable assistive-listening systems	Kealey et al. [34]; Gurvich et al. [36]	The wearable glasses-mounted system developed by Kealey et al. has an eight-microphone array, a camera, and headphones. The authors estimate the total latency of their system to be about 120 milliseconds. In addition, Gurvich et al. describe a speaker-detection system designed for use with low-power edge-computing devices.	While both systems could be useful as part of XR-style or assistive acoustic pipelines for SELD applications, the primary functions of these systems are speech enhancement or active-speaker detection rather than general SELD.

Table 8. *Cont.*

System Category	Representative Studies	Deployment Evidence	Main Limitation/Takeaway
Speech enhancement, source separation, and hearable low-latency systems	Srinivas et al. [43]; Dementyev et al. [49]; Jo et al., EgoSep [56]	Sub-millisecond algorithmic latency and an overall end-to-end latency of approximately 3.35 milliseconds were reported by Dementyev et al. as part of their speech-enhancement hearable-oriented system. In addition, the EgoSep study reported successful operation at 60 Hz along with a Real-Time Factor (*RTF*) of 0.17 when run on an NVIDIA Jetson Nano.	These studies are useful for understanding low-latency audio processing and multimodal edge auditory scene analysis. However, they do not perform complete SELD, so they mainly serve as supporting references rather than direct SELD systems.
Systems that learn spatial features or DOA estimation	Wang et al., IPDnet2 [45]; Marhoff et al. [41]; Slizovskaia et al. [27]	The main way these studies were tested was through localization evaluation or DCASE-style benchmark evaluation. A few provide better performance against directional interference, and most do not specify the target edge hardware.	While the above studies contribute to spatial estimation related to SELD, they are not complete SELD systems. Some do not detect sound events, some assume a target class, and none evaluate a complete SELD pipeline.
Systems that perform detection and localization for a specific type of event or set of events	Munirathinam et al. [52]; Wang et al. [5]	Munirathinam et al. used an STM32F411RE microcontroller with multiple digital microphones and employed GCC-PHAT and MFCCs. Wang et al. deployed several acoustic devices connected to an edge server.	Compared with detection-only or localization-only systems, the above systems are closer to SELD because they combine detection and localization. However, their task scope is narrower than general SELD, and neither system represents a standard full SELD benchmark system.
Acoustic perception based on ultra-low-power edge hardware and event-driven sensors	Ceolini et al. [44]; Chen et al. [4]; Cerutti et al. [42]	This body of work explores ultra-low-power edge-hardware-based acoustic perception using event-driven sensors, MCU-based processing, and/or binary neural networks.	The results reported by Ceolini et al. include keyword spotting greater than 93%, speaker verification greater than 88%, and the lowest system latency of 5 ms. The results reported by Chen et al. include deployment on a CC1350 MCU with 33.4 kB flash memory, 9.94 kB SRAM memory, and 23.1 mW power consumption.

To avoid treating all reported "real-time" results as equivalent evidence of deployment readiness, the studies were interpreted according to the strength of their deployment evidence. Direct deployment evidence was considered strongest when a complete SELD pipeline was evaluated on target edge/mobile hardware with measured timing or resource metrics. Model-level efficiency measurements or evaluations lacking complete pipeline timing were considered partial deployment evidence. Studies addressing related tasks such as SED, SSL, speech enhancement, or audio−visual processing were treated as supporting evidence and were not considered direct demonstrations of complete SELD deployment.

Because no formal study-level risk-of-bias assessment tool was applied, no study-level risk-of-bias ratings are reported.

Overall, the comparison shows that real-time edge/mobile SELD is becoming feasible, although the strength of deployment evidence remains uneven across studies. As discussed in Section 8.4, standardized reporting of deployment-oriented metrics on target hardware remains necessary for meaningful comparison.

This review has several methodological limitations. The literature search was restricted to English-language publications retrieved through SciSpace, SciSpace Full Text, Google Scholar, and arXiv up to 30 June 2026, and relevant studies indexed elsewhere may therefore have been missed. Study selection and data extraction were performed by the first author, with automated tools used to assist prioritization and information extraction, rather than through independent duplicate screening and extraction. In addition, the PRISMA study-selection counts were reconstructed retrospectively during revision from archived search outputs and manuscript-version history rather than from a contemporaneously maintained screening log. Finally, no formal risk-of-bias or certainty-of-evidence framework was applied because of the substantial heterogeneity of the included engineering and deployment studies. These limitations may have increased the risk of missed or misclassified records and introduced additional uncertainty into the retrospectively reconstructed screening counts.

10. Future Directions

Significant advances have occurred in recent years in the development of SELD systems for edge and mobile devices. However, several limitations remain, particularly in standardized deployment evaluation, hardware efficiency, robustness to domain shift, and real-world adaptation, as summarized in Sections 6 and 9. The following sections highlight several promising research directions that could help address these challenges, while broader directions are discussed separately as the authors' perspectives.

10.1. Standardized End-to-End Edge Benchmarks

To our knowledge, there are currently no widely adopted standardized benchmarks for evaluating edge-based SELD deployment. The lack of standardized benchmarks prevents advancement within this area because it is difficult to compare systems and assess true deployment readiness. Most published studies have focused on reporting SELD accuracy, model size, FLOPs, or real-time factor; however, these measurements are typically taken using various combinations of hardware, software, feature extraction, and latency assumptions. Collectively, these publications demonstrate that real-time edge acoustic perception is a full-system problem that encompasses features, hardware, latency, energy use, and robustness to interference/noise.

1. Hardware-Specific Benchmarks: As discussed in Sections 5.1 and 8.4, future edge-SELD benchmarks should evaluate the complete processing pipeline rather than neural-network inference alone. Such evaluation should include feature extraction and other processing stages whose costs may become significant on constrained hardware [20,21].
2. Deployment Metrics: Hardware-specific benchmarks will enable a more equitable method of comparison across multiple edge platforms, such as Raspberry Pi, NVIDIA Jetson, mobile SoCs, embedded GPUs, and microcontroller-class devices. Each benchmark must provide a defined set of standard measurement protocols for latency, memory use, power consumption, model size, real-time factor, startup time, and sustained operation under thermal constraints. The deployment metric(s) must be provided in conjunction with the SELD accuracy metric(s), such as error rate, F-score, localization error, localization recall, and/or SELD score.
3. Diverse Evaluation Scenarios: Additionally, benchmarks must utilize multiple acoustic and deployment scenarios. The evaluation must assess varying signal-to-noise ratios, reverberation times, background noise types, microphone-array geometries, channel counts, and device conditions. This is particularly critical for mobile and wearable systems that vary significantly from controlled laboratory environments

with regard to microphone placement, device orientation, motion, and non-standard array configurations.

4. Open-Source Reference Implementations: Open-source reference implementations will help increase reproducibility and reduce the barrier to entry. The open-source reference implementation should include training code, model architectures, feature-extraction pipelines, deployment scripts, quantization or compression settings, and hardware measurement tools. Without such open-source reference implementations, reported efficiency results are likely to continue being difficult to reproduce across multiple devices.
5. Challenge Tracks for Edge Deployment: A challenge track for edge deployment in DCASE or a comparable challenge would expedite progress by providing an assessment mechanism that evaluates SELD systems based on both accuracy and deployment metrics. Such a challenge track would encourage practical research into low-latency, energy-efficient, and hardware-aware SELD systems and assist in bridging the gap between laboratory benchmarks and real-world applications.
6. Microphone Calibration and Device Profiles: Another issue that future edge-based SELD benchmarks should consider is calibration and device profiling. The main reason SELD systems have difficulty transferring from one device to another is that every device has its own unique microphone placement, geometry, number of channels, spacing, frequency response, gain, phase response, and noise characteristics. Therefore, if a model is trained using FOA, tetrahedral MIC arrays, or a specific research microphone, the model's performance will differ significantly when deployed on smartphones, hearables, smart speakers, and other small IoT devices equipped with non-traditional microphone configurations [1,2,13,15,19].

One potential avenue is to establish simple calibration procedures that allow the creation of a device profile prior to deployment. As an example, a very short calibration sound, such as a chirp, a clap, a known source location, or a built-in test signal, could be used to generate estimates for differences in microphone gain, differences in microphone phase, channel delays, differences in frequency responses, and approximate geometries. Once these device profiles have been established, they can be applied to either the SELD model or the feature-extraction pipeline to compensate for the specific microphone characteristics of that particular device.

Device-profile standardization is important since edge and mobile SELD systems are heavily impacted by microphone-array limitations and the effects of domain shift. Reporting not only SELD accuracy in future benchmarking efforts, but also whether the SELD systems were tested under calibrated vs. uncalibrated conditions, whether the systems utilize device-specific profiles, and how well the SELD models transfer to multiple different microphone configurations, would greatly facilitate comparisons of SELD systems across multiple smartphone brands, hearable manufacturers, Raspberry Pi devices, and other embedded platforms [1,2,19,20].

Simulation-based SELD benchmarks have the potential to assist researchers in evaluating edge-based deployment more systematically than traditional benchmarks. By rendering the same acoustic scene with different microphone configurations, device locations, and hardware assumptions, researchers can compare how well a particular model transfers between smartphones, hearables, Raspberry Pi devices, embedded processors, etc. For this direction to be beneficial, future simulation-based benchmarks will need to provide researchers with metrics beyond SELD accuracy. Researchers will want to know about latency, feature-extraction cost, memory usage, real-time factor, and power consumption on target hardware [20].

10.2. Hardware-Aware Model and Pipeline Optimization

Neural Architecture Search (NAS) can offer a promising direction toward creating SELD systems that are optimized for real-time operation on edge and mobile devices. In contrast to most existing SELD models, which were created based upon architectures designed for image recognition, speech processing, or sequence modeling, such as CRNNs, TCNs, Conformers, Transformers, and state-space models [24,29,47], the majority of these architectures were chosen manually and empirically rather than systematically optimized for specific hardware constraint sets.

Therefore, in terms of edge SELD, NAS may be used to search for architectures that jointly optimize detection accuracy, localization accuracy, latency, memory use, and power consumption. Since a SELD system must perform sound event detection, direction estimation, and simultaneous handling of multiple sources, all within real-time constraints, this is inherently a multi-objective problem. A hardware-aware NAS framework would therefore seek models located on the Pareto frontier between accuracy and efficiency. Thus, different constrained devices, such as smartphones, Raspberry Pi devices, NVIDIA Jetson platforms, microcontrollers, and wearables, may require different optimal configurations.

Future NAS approaches for SELD, however, should not be limited solely to searching for the best possible neural network architecture. Real-time SELD performance is dependent upon the entire processing pipeline and includes factors such as feature extraction, input window length, number of microphone channels, spatial feature representation, output format, post-processing, quantization strategy, etc. Therefore, future NAS approaches should consider end-to-end pipeline search. That is, they should search for the best possible model architecture combined with the best possible feature representation and deployment strategy [20,48].

Candidate systems could be evaluated using both SELD metrics, such as F-score and localization error, and deployment metrics, such as latency, memory footprint, FLOPs, *RTF*, and energy consumption. Differentiable NAS or weight-sharing NAS may potentially reduce architecture-search cost, while transfer learning from related audio tasks could also improve search efficiency.

Another promising practical area for onboard SELD is energy-aware adaptation and duty cycling. Edge and mobile devices typically cannot process the entire SELD pipeline at peak performance levels continuously due to the total power consumed by feature extraction, model inference, buffering, and post-processing [20,48]. One way to achieve this would be to implement a low-power always-on detector to identify potential audio events and, once identified, initiate full SELD processing. For example, smart-home devices can be placed into a low-power monitoring mode during periods of quiet operation; however, when an alarm, door knock, glass break, or another type of sound event occurs, they can automatically switch to full SELD operation. Wearable/mobile devices can dynamically control their sample rates, feature-extraction parameters, model sizes, and/or inference frequencies based on the specific characteristics of the acoustic environment, available battery life, and/or user prioritization.

Related studies have demonstrated low-power acoustic sensing using MCU-based SED [4], binary neural networks for power-constrained SED [42], and FPGA-based sound localization [32]. However, these approaches do not provide complete edge-SELD solutions. Future work should therefore evaluate whether energy-aware operation and duty cycling can reduce power consumption while maintaining acceptable SELD accuracy, wake-up latency, and real-time performance [20,48].

10.3. Robustness, Domain Generalization, and On-Device Adaptation

At present, most SELD systems are designed to be trained offline using predefined or fixed training sets and then run in a static manner once deployed. However, operating conditions in edge and mobile applications may differ from typical test conditions because of variations in room acoustics, background noise levels, microphone location, microphone gain/phase imbalance, device orientation, and user-specific sound-priority requirements. These differences can contribute to degradation of system performance after deployment. On-device adaptation therefore represents a promising approach for reducing this domain difference without requiring complete retraining in the cloud [14,15,20].

1. Lightweight Adaptation: Several forms of lightweight adaptation may be useful for edge SELD. Continual learning could allow a system to adapt gradually to changes in acoustic conditions, while few-shot learning could allow users to provide a small number of examples of important sounds in their own environments. Unsupervised or self-supervised adaptation could also use unlabeled audio collected during operation to estimate local background-noise characteristics, microphone differences, or reverberant effects. These approaches are particularly relevant to mobile and wearable devices, where microphone configuration, device orientation, and acoustic conditions may change during use. However, future systems will need adaptation methods that require only limited computation and do not depend on complete SELD annotations.
2. Human-in-the-Loop Feedback: Another interesting direction for practical adaptation involves using "human-in-the-loop" (HITL) adaptation. Most edge and mobile applications will likely have users who cannot provide the required SELD labels, such as event class, azimuth, elevation, and, when distance estimation is included, source distance [13,22]; however, they can still provide some form of weak feedback. Examples include confirming whether an alert was correct, marking a sound as important/unimportant, identifying a repeated source of a particular sound, or simply correcting the system when it has confused events. This type of feedback would allow for personalized learning with respect to environmental context, i.e., adapting to what sounds are relevant within a given environment, without requiring complete manual annotation. As a result, over time, a home device, wearable device, or mobile device could potentially learn which sounds are most significant in a particular environment, such as doorbells, alarms, vehicles, appliances, or warning signals. However, the primary challenges relate to designing effective HITL adaptation strategies that leverage weak user feedback while limiting computational complexity and potential power consumption, minimizing privacy risks, and avoiding degradation of the initial SELD model [20,48].
3. Environment-Specific SELD Model Development: Environment-specific SELD models are another viable option for practical model development. Rather than developing a single generic SELD model that is expected to perform equally well across multiple environments, future systems could develop, adapt, or select models based on their environment. For example, a smart-home model may prioritize alarms, appliances, and human activity, while an outdoor mobile model may prioritize vehicles, sirens, and warning sounds. This direction is relevant because SELD performance can be affected by differences in acoustic environment, microphone configuration, reverberation, noise, and event distribution [14,15]. However, these specialized models must be trained, selected, updated, and stored on resource-constrained devices without creating excessive memory, latency, or maintenance costs [20,48].
4. Simulation-Based Training and Sim-to-Real Transfer: Training and testing SELD models using simulated spatial audio could represent an additional direction for SELD research. Simulation-based approaches can provide controlled source locations, mo-

tion, timing, reverberation conditions, and other acoustic properties while automatically providing ground-truth spatial information. This can facilitate the generation of labeled training and evaluation data and the testing of challenging conditions such as overlapping sound sources, moving sources, reverberation, and directional interference [14,15].

The primary challenge associated with simulation-based training environments is sim-to-real transfer. A SELD model trained on a simulated dataset may fail to generalize well when deployed in a real-world acoustic environment because differences exist in the physical properties of real microphones, room acoustics, background noise levels, and source characteristics compared to those assumed in the simulated training environment. Thus, this represents yet another example of domain shift and is directly related to the difference between the performance of SELD models reported in benchmark studies and their performance in real-world deployment settings. Future investigations into methods for combining simulated SELD data with real recording datasets, self-supervised pre-training strategies, domain-adaptation techniques, and on-device fine-tuning will likely lead to improved generalization capabilities for SELD models [20,47].

Simulation can also provide controlled variation of microphone geometry, source distance, room characteristics, background noise, and device motion before evaluation on actual hardware.

In summary, a major limitation of practical on-device adaptation and domain-generalization methods is computational overhead. Training is inherently significantly more computationally costly than testing/inference. Many existing edge-based SELD platforms currently operate close to their limits regarding latency, memory consumption, and power use. Thus, practical implementations may require parameter-efficient updates, including adapters, low-rank adaptation, sparse updates, recalibration of normalization parameters, updating thresholds, or simply re-tuning lightweight classifier heads. In the future, researchers are expected to investigate not only whether on-device adaptation improves SELD performance metrics but also whether it can do so in accordance with reasonable energy use, acceptable latency constraints, sufficient memory availability, and reasonable privacy bounds.

10.4. Self-Supervised and Data-Efficient Spatial Learning

As opposed to traditional supervised learning, for which annotated training data are required, self-supervised learning has become increasingly popular for reducing a SELD system's reliance on costly and labor-intensive labeled training data. Most SELD systems rely heavily on several types of annotation, including labeled activity regarding what type of sound is occurring, the identity of each sound source, i.e., class, the direction from which each sound event originates (DOA), and occasionally the distance from each sound event. Collecting these types of annotations, particularly in realistic environments where there may be multiple simultaneous source signals, reverberation, etc., presents significant challenges. Large amounts of unlabeled multichannel audio can be used to train a model to learn useful acoustic and spatial representations using self-supervised pre-training followed by fine-tuning on a labeled SELD dataset [47].

For the purposes of SELD, self-supervised learning needs to address both the spatial and semantic aspects of audio. Potential pretext tasks for this area might include predicting masked multichannel spectrograms, restoring removed inter-channel phase differences, estimating missing microphone channels, estimating how consistent a particular sound event is across channels, or performing contrastive learning based on different views of the same acoustic environment. All of these would promote the learning of representations that contain temporal relationships and direction-related information.

Spatial Self-Supervised Learning: Potential approaches include contrastive learning using different views of the same acoustic scene and masked prediction of missing time-frequency regions, spatial-feature channels, or inter-channel phase patterns. However, spatial information must be preserved during training, since operations such as channel shuffling, time shifting, or phase modification can damage direction-of-arrival cues. These approaches could potentially improve robustness to noise, reverberation, missing channels, or degraded microphone input.

One major barrier to deploying the resulting models at the edge is the high computational cost typically associated with pre-trained self-supervised models. To address this issue, a reasonable solution is to perform all pre-training on large collections of unlabeled data offline, then transfer the learned representations into compact SELD models via fine-tuning, knowledge distillation, adapter layers, or frozen lightweight feature extraction [37,42,43,47]. Future work should explore whether self-supervised spatial representations can contribute to improved SELD system performance, including accuracy and robustness, while remaining within the bounds established for edge deployment, including latency, memory usage, and power consumption.

10.5. Authors' Perspectives

Beyond the gaps directly identified in the reviewed edge and mobile SELD studies, several broader directions may also be considered. A number of edge and mobile devices have sensors besides microphones, including cameras, inertial measurement units (IMUs), depth sensors, and head-tracking modules. Because these additional sensors can provide complementary information about visible sound sources, device motion, user orientation, and environmental context, multi-modal integration has potential as a broader direction for real-time SELD and other spatial audio tasks [34,36,56].

1. Audio-Visual SELD: An example of how integrating visual and auditory information can enhance spatial audio awareness is by providing additional localization information through facial features, objects detected in the scene, lip motion, and scene understanding. Visual information can help disambiguate overlapping or same-class sources and/or degraded localization due to reverberant and noisy environments. Other researchers have shown success in similar tasks using audio−visual systems, including active speaker detection, audio−visual speech enhancement, and egocentric on-screen sound-source separation [34,36,56]. Therefore, future SELD models may benefit from using visual information as a prior for spatialization or as a contextual cue instead of relying only on microphone signals.
2. Spatial Audio Rendering: Also important for AR/VR/hearables, SELD can be integrated with spatial audio rendering and head tracking. While sound localization is primarily utilized for recognition purposes, it can also be used for presentation. Sounds detected from the real world can be amplified, enhanced, muted, relocated, captioned, etc., as spatial cues within an immersive interface [7]. To achieve this type of functionality, there needs to be close cooperation between sensing, localization, sound control, rendering, user modeling, and device-state estimation with minimal delay [7,34].
3. Beamforming in Mobile Devices: One possible application of SELD technology includes combining it with mobile beamforming. There are numerous applications where SELD can provide information about what sound sources are present and where they are located. At the same time, it may be beneficial to enhance these sound sources by using beamforming to emphasize them and suppress unwanted background noise. Examples of such applications include assistive-listening systems, such as smartphones and hearables, noisy public places, such as shopping malls, emer-

gency situations, such as earthquakes, and virtual reality (VR) interfaces where certain sounds need to stand out over others. Prior research on mobile microphone arrays, smartphone-assisted hearing-aid localization, and wearable audio−video enhancement suggests that this direction is technically feasible and relevant [2,3,34]. However, as previously stated, SELD-guided beamforming on resource-constrained devices presents significant challenges, including small microphone spacings, changing device orientations, limited power resources, and the need to maintain real-time processing while performing both localization and enhancement.

4. Distributed and Cooperative Device Fusion: One other area to consider for future edge SELD is cooperative edge-device localization. As opposed to having a single device use a small number of microphones, multiple devices can work together to create a larger virtual sensing system. For example, a user's smartphone, smartwatch, smart speaker, smart earbud/hearable, or even a nearby IoT sensor could send some level of acoustic data, possibly low-resource feature data, or a local estimate of events. Combining all these signals or predictions could potentially increase sound localization accuracy, reduce the effects of occlusion, and provide improved robustness against reverberation and noise. There are many reasons why this approach is beneficial, since most edge devices today have very little microphone spacing and limited channel counts, making it difficult to achieve localization accuracy [1,2,19]. Edge-based sound event detection and localization were demonstrated previously in city-surveillance-type applications utilizing cooperative acoustic sensing [5]; however, broader literature regarding distributed edge-based acoustic localization also focuses on platform-level issues, such as communication costs, energy usage, timing requirements, and hardware constraints [48]. Distributed SELD will add additional difficulties, such as device synchronization, clock drift, network delay, privacy, and estimating device position. Therefore, future work should explore ways to perform cooperative SELD to improve localization while limiting communication cost, delays, energy consumption, and user-privacy risks.

However, the integration of multiple sensors introduces its own set of issues. The addition of multiple sensors increases computational overhead, power consumption, synchronization complexity, and privacy concerns. Additionally, audio and visual sensors need to be calibrated into a single coordinate system in order to fuse them effectively. Further, if the timing between the modalities differs significantly, then the overall effectiveness of the system will likely decrease. Therefore, further investigation of multi-modal fusion methods for edge deployment is needed, including early fusion, intermediate feature fusion, late decision fusion, and attention-based fusion, each of which offers trade-offs among accuracy, latency, robustness, and hardware cost [20,48].

Privacy, Security, and Trustworthy Deployment: Edge-based SELD systems may continuously or near-continuously process ambient audio that can include speech and information about activities in homes, workplaces, vehicles, and other environments. Although local edge processing can reduce the transmission of raw audio to remote servers, it does not eliminate privacy or security risks. Future systems should therefore consider data-minimization approaches, such as local feature extraction, selective event reporting, anonymized representations, and privacy-preserving model updates.

Trustworthy SELD deployment should also consider false alarms, misleading or adversarial audio inputs, and uncertainty in event predictions. These issues are particularly important for safety-related applications, where an incorrect or low-confidence detection may result in inappropriate downstream actions. Future evaluation should therefore consider not only SELD accuracy, but also resilience to unreliable inputs, confidence or un-

certainty estimates, privacy protection, and continued operation under realistic deployment conditions, particularly under noise, reverberation, and domain shift [14,15,20].

11. Conclusions

This review evaluated the existing state of the art in real-time Sound Event Localization and Detection (SELD) for edge and mobile devices, with an emphasis on deployment readiness. The reviewed literature indicates that SELD is undergoing a transformation. The field is shifting from primarily offline or GPU-centric experimentation toward lightweight architectures, efficient spatial features, and demonstrations on platforms including Raspberry Pi, NVIDIA Jetson, mobile processors, and other constrained hardware. This indicates that real-time edge SELD is technically possible under specified deployment constraints.

11.1. Present Capability

Modern lightweight SELD systems have achieved considerable progress. Compact architectures such as MMT-Net, quaternion-domain SELD models, SELD-TCNs, and single-stage CRNN-based systems have demonstrated that it is possible to decrease model size, memory usage, and inference cost without significantly reducing detection and localization quality [19,28,29,31]. Additionally, efficient feature representations such as SALSA-Lite have further decreased the costs associated with extracting spatial features in real time, making real-time processing on edge platforms more practical [21]. Collectively, these contributions represent the first step toward demonstrating that real-time SELD is no longer solely theoretical.

11.2. Persistent Limitations

Significant limitations remain before practical edge/mobile SELD becomes widespread. These include constraints imposed by compact microphone arrays and resource-limited hardware, as discussed in Section 5, together with robustness and domain-generalization challenges under realistic acoustic conditions, as discussed in Section 6 [1,2,14,15,19,20,24,26,33].

11.3. Deployment Gap

Deployment readiness cannot be determined from SELD accuracy alone. As discussed in Section 8.4 and summarized in Section 9 and Supplementary Table S1, deployment-related measurements are reported inconsistently across the reviewed literature. Future evaluations should therefore assess SELD performance together with standardized deployment metrics on the intended target hardware [20,48].

11.4. Readiness of Applications

The readiness of edge SED/SSL/SELD-like systems varies by application domain. Smart-home, IoT monitoring, environmental awareness, and acoustic safety systems have shown reasonable promise using currently available lightweight SED, SSL, and SELD-like systems [1,4,5,51]. On the other hand, more stringent applications such as assistive hearing, automotive safety, and XR interfaces require lower latency, robustness under uncontrolled acoustic conditions, lower power consumption, support for sparse microphone arrays, and reliable user feedback. Although current systems show promise for these more demanding applications, additional testing on actual hardware and in uncontrolled deployment environments would be beneficial for large-scale deployment.

11.5. Directions for Research

Several research areas could help close the above-described gaps. Hardware-aware model design and hardware−algorithm co-design could improve the alignment of SELD systems with their target devices [8,48]. Self-supervised spatial-audio representation

learning could reduce reliance on costly annotated SELD datasets and enhance generalizability [47]. Improved output representations and localization approaches, including Multi-ACCDOA, AD-YOLO, and class-conditioned sound event localization, could potentially alleviate some problems related to overlapping events, polyphony, and directional interference [24,26,27]. Further improvements may also arise from on-device adaptation, energy-aware duty cycling, model compression, spatial-feature efficiency, multimodal sensing, distributed device fusion, microphone calibration and device profiling, privacy-aware processing, and standardized edge-deployment benchmarks [2,5,20,48].

11.6. Summary Conclusion

Real-time SELD for edge and mobile devices should be regarded as an emerging reality rather than a solved problem. The evidence summarized in Section 9 and Supplementary Table S1 demonstrates that real-time operation is achievable under selected hardware and acoustic conditions, but deployment readiness remains inconsistent across systems. Future progress will depend on jointly improving efficiency, robustness, hardware compatibility, and standardized end-to-end evaluation.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/asi9100198/s1, Table S1: Full deployment-oriented comparison of reviewed SELD, SELDDE, and SELD-like acoustic systems; Table S2: Comparative summary of related edge/mobile acoustic systems that are not complete SELD systems. Supplementary File S1: PRISMA 2020 Checklist.

Author Contributions: Conceptualization, P.T. and P.V.; methodology, P.T.; investigation, P.T.; resources, P.T.; writing—original draft preparation, P.T.; writing—review and editing, P.T. and P.V.; visualization, P.T.; supervision, P.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used in this systematic review were derived from published sources. The complete search strategies are reported in Table 1, and the extracted study-level characteristics and deployment information are provided in Supplementary Tables S1 and S2. No analysis code was used for the synthesis.

Acknowledgments: During the preparation and revision of this manuscript, SciSpace, accessed in June–July 2026, was used to assist with literature discovery, search prioritization, and the identification of relevant technical information. ChatGPT (GPT-5.6 Sol) by OpenAI, accessed in July–September 2026, was used to assist with exploratory research, organization of technical material, and preliminary checking of whether selected statements were adequately supported by the cited references. The first author independently reviewed the relevant source publications and verified the bibliographic information, citations, technical claims, and interpretations; both authors reviewed and edited the final manuscript. The authors take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACCDOA	Activity-Coupled Cartesian Direction of Arrival
ADL	Activities of Daily Living
ADPIT	Auxiliary Duplicating Permutation-Invariant Training

AD-YOLO	Angular-Distance YOLO
AI	Artificial Intelligence
ANC	Active Noise Cancellation
AR	Augmented Reality
ASC	Acoustic Scene Classification
AV	Audio−Visual
BNN	Binary Neural Network
CNN	Convolutional Neural Network
CQT	Constant-Q Transform
CRNN	Convolutional Recurrent Neural Network
DCASE	Detection and Classification of Acoustic Scenes and Events
DOA	Direction of Arrival
FOA	First-Order Ambisonics
FLOPs	Floating-Point Operations
FPGA	Field-Programmable Gate Array
GCC-PHAT	Generalized Cross-Correlation with Phase Transform
GPU	Graphics Processing Unit
HITL	Human-in-the-Loop
IMU	Inertial Measurement Unit
IoT	Internet of Things
IPD	Inter-Channel Phase Difference
L3DAS	Learning 3D Audio Sources
LTSD	Long-Term Spectral Divergence
MCU	Microcontroller Unit
MFCC	Mel-Frequency Cepstral Coefficient
MIC	Microphone Array
MMT-Net	Modern Multi-Scale Temporal Network
MR	Mixed Reality
NAS	Neural Architecture Search
NIGENS	Non-Individualized Generated Environmental Sounds
NIPD	Normalized Inter-Channel Phase Difference
QSELD	Quaternion Sound Event Localization and Detection
RIR	Room Impulse Response
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RTF	Real-Time Factor
SALSA	Spatial Cue-Augmented Log-Spectrogram
SED	Sound Event Detection
SELD	Sound Event Localization and Detection
SELDDE	Sound Event Localization and Detection with Source-Distance Estimation
SDE	Source-Distance Estimation
SNR	Signal-to-Noise Ratio
SoC	System on Chip
SSL	Sound-Source Localization
STARSS	Sony-TAU Realistic Spatial Soundscapes
SVM	Support Vector Machine
TAU	Tampere University
TCN	Temporal Convolutional Network
VR	Virtual Reality
XR	Extended Reality
YOLO	You Only Look Once

References

1. Yeow, J.-W.; Tan, E.-L.; Peksi, S.; Ong, Z.-T.; Gan, W.-S. Enhancing Situational Awareness in Wearable Audio Devices Using a Lightweight SELD System. In *Proceedings of the 2025 AES International Conference on Headphone Technology, Espoo, Finland, 27–29 August 2025*; Audio Engineering Society: New York, NY, USA, 2025; pp. 77–84.
2. Ganguly, A.; Panahi, I. Non-Uniform Microphone Arrays for Robust Speech Source Localization for Smartphone-Assisted Hearing Aid Devices. *J. Signal Process. Syst.* **2018**, *90*, 1415–1435. Erratum in *J. Signal Process. Syst.* **2018**, *90*, 1437. https://doi.org/10.1007/s11265-017-1297-8. [CrossRef] [PubMed]
3. Küçük, A. Sound Source Localization for Improving Hearing Aid Studies Using Mobile Platforms. Ph.D. Thesis, The University of Texas at Dallas, Richardson, TX, USA, 2021. Available online: https://utd-ir.tdl.org/items/8b475ada-9db2-4a92-ab47-1e51ca2202fe (accessed on 13 July 2026).
4. Chen, H.; Zhu, M.; Zhou, J.; Qiao, J.; Liu, B. Sound Event Detection on a Single MCU: Lightweight, Low-Power and Portable. In *Proceedings of the 2022 International Conference on Sensing, Measurement & Data Analytics in the Era of Artificial Intelligence (ICSMD), Harbin, China, 30 November 2022–2 December 2022*; IEEE: New York, NY, USA, 2022; pp. 300–304. [CrossRef]
5. Wang, J.; Liu, K.; Tzanetakis, G.; Pan, J. Learning-Based Cooperative Sound Event Detection with Edge Computing. In *Proceedings of the 2018 IEEE 37th International Performance Computing and Communications Conference (IPCCC), Orlando, FL, USA, 17–19 November 2018*; IEEE: New York, NY, USA, 2018; pp. 1–8. [CrossRef]
6. Lee, C.; Yuh, A.H.; Kang, S.J. Real-Time Prediction of Resident ADL Using Edge-Based Time-Series Ambient Sound Recognition. *Sensors* **2024**, *24*, 6435. [CrossRef] [PubMed]
7. Gupta, R.; He, J.; Ranjan, R.; Gan, W.-S.; Klein, F.; Schneiderwind, C.; Neidhardt, A.; Brandenburg, K.; Välimäki, V. Augmented/Mixed Reality Audio for Hearables: Sensing, Control, and Rendering. *IEEE Signal Process. Mag.* **2022**, *39*, 63–89. [CrossRef]
8. Yin, J.; Damiano, S.; Verhelst, M.; van Waterschoot, T.; Guntoro, A. Real-Time Acoustic Perception for Automotive Applications. In *Proceedings of the 2023 Design, Automation & Test in Europe Conference & Exhibition (DATE), Antwerp, Belgium, 17–19 April 2023*; IEEE: New York, NY, USA, 2023; pp. 1–6. [CrossRef]
9. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
10. Nakadai, K.; Kumon, M.; Okuno, H.G.; Hoshiba, K.; Wakabayashi, M.; Washizaki, K.; Ishiki, T.; Gabriel, D.; Bando, Y.; Morito, T.; et al. Development of Microphone-Array-Embedded UAV for Search and Rescue Task. In *Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017*; IEEE: New York, NY, USA, 2017; pp. 5985–5990. [CrossRef]
11. Han, C.; Kang, Y.S.; Sim, Y.; Park, H.W.; Jun, M.B.-G. LISTEN: Lightweight Industrial Sound-Representable Transformer for Edge Notification. *arXiv* **2025**, arXiv:2507.07879. [CrossRef]
12. Wu, X.; Ni, R.; Cui, X.; Guo, T.; Peng, H. Sound Source Localization Based on Three-Element Microphone Array and RepMobileViT Model. *Appl. Acoust.* **2025**, *231*, 110480. [CrossRef]
13. Politis, A.; Mesaros, A.; Adavanne, S.; Heittola, T.; Virtanen, T. Overview and Evaluation of Sound Event Localization and Detection in DCASE 2019. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 684–698. [CrossRef]
14. Politis, A.; Adavanne, S.; Krause, D.; Deleforge, A.; Srivastava, P.; Virtanen, T. A Dataset of Dynamic Reverberant Sound Scenes with Directional Interferers for Sound Event Localization and Detection. In *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021), Barcelona, Spain, 15–19 November 2021*; IEEE: New York, NY, USA, 2021; pp. 125–129. [CrossRef]
15. Politis, A.; Adavanne, S.; Virtanen, T. 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, 2–4 November 2020*; DCASE: Tokyo, Japan, 2020; pp. 165–169.
16. Yeow, J.-W.; Tan, E.-L.; Peksi, S.; Gan, W.-S. Environmental Acoustic Intelligence through Sound Event Localization and Detection: A Review. *npj Acoust.* **2025**, *1*, 31. [CrossRef]
17. Mohmmad, S.; Sanampudi, S.K. A Parametric Survey on Polyphonic Sound Event Detection and Localization. *Multimed. Tools Appl.* **2025**, *84*, 22083–22120. [CrossRef]
18. Jekateryńczuk, G.; Piotrowski, Z. A Survey of Sound Source Localization and Detection Methods and Their Applications. *Sensors* **2024**, *24*, 68. [CrossRef] [PubMed]
19. Kim, D.; Park, J.; Kim, D. MMT-Net: A Lightweight Hybrid of Multi-Scale Modern TCN and Transformer for Real-Time SELD. In *Proceedings of the 2025 25th International Conference on Control, Automation and Systems (ICCAS), Incheon, Republic of Korea, 4–7 November 2025*; IEEE: New York, NY, USA, 2025; pp. 191–196. [CrossRef]
20. Yeow, J.-W.; Tan, E.-L.; Bai, J.; Peksi, S.; Gan, W.-S. Real-Time Sound Event Localization and Detection: Deployment Challenges on Edge Devices. *arXiv* **2024**, arXiv:2409.11700. [CrossRef]

21. Nguyen, T.N.T.; Jones, D.L.; Watcharasupat, K.N.; Phan, H.; Gan, W.-S. SALSA-Lite: A Fast and Effective Feature for Polyphonic Sound Event Localization and Detection with Microphone Arrays. In *Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, 22–27 May 2022*; IEEE: New York, NY, USA, 2022; pp. 716–720. [CrossRef]
22. Dong, Y.; Wang, Q.; Hong, H.; Jiang, Y.; Cheng, S. An Experimental Study on Joint Modeling for Sound Event Localization and Detection with Source Distance Estimation. In *Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing, Hyderabad, India, 6–11 April 2025*; IEEE: New York, NY, USA, 2025; pp. 1–5. [CrossRef]
23. Adavanne, S.; Politis, A.; Nikunen, J.; Virtanen, T. Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks. *IEEE J. Sel. Top. Signal Process.* **2019**, *13*, 34–48. [CrossRef]
24. Shimada, K.; Koyama, Y.; Takahashi, S.; Takahashi, N.; Tsunoo, E.; Mitsufuji, Y. Multi-ACCDOA: Localizing and Detecting Overlapping Sounds from the Same Class with Auxiliary Duplicating Permutation Invariant Training. In *Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, 22–27 May 2022*; IEEE: New York, NY, USA, 2022; pp. 316–320. [CrossRef]
25. Shimada, K.; Takahashi, N.; Takahashi, S.; Mitsufuji, Y. Sound Event Localization and Detection Using Activity-Coupled Cartesian DOA Vector and RD3Net. DCASE 2020 Challenge Technical Report. *arXiv* **2020**. [CrossRef]
26. Kim, J.S.; Park, H.J.; Shin, W.; Han, S.W. AD-YOLO: You Look Only Once in Training Multiple Sound Event Localization and Detection. In *Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, 4–10 June 2023*; IEEE: New York, NY, USA, 2023; pp. 1–5. [CrossRef]
27. Slizovskaia, O.; Wichern, G.; Wang, Z.-Q.; Le Roux, J. Locate This, Not That: Class-Conditioned Sound Event DOA Estimation. In *Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, 22–27 May 2022*; IEEE: New York, NY, USA, 2022; pp. 711–715. [CrossRef]
28. Shah, S.O.; Darwaish, M.; Shah, M.; Khan, M.A.; Shujaat, M. Real-Time Sound Event Localization and Detection Using IoT Raspberry Pi Devices Based on Single-Stage CRNN. *Spectr. Eng. Sci.* **2025**, *3*, 1020–1033. [CrossRef]
29. Brignone, C.; Mancini, G.; Grassucci, E.; Uncini, A.; Comminiello, D. Efficient Sound Event Localization and Detection in the Quaternion Domain. *IEEE Trans. Circuits Syst. II Express Briefs* **2022**, *69*, 2453–2457. [CrossRef]
30. Yin, J.; Verhelst, M. CNN-Based Robust Sound Source Localization with SRP-PHAT for the Extreme Edge. *ACM Trans. Embed. Comput. Syst.* **2023**, *22*, 75. [CrossRef]
31. Guirguis, K.; Schorn, C.; Guntoro, A.; Abdulatif, S.; Yang, B. SELD-TCN: Sound Event Localization & Detection via Temporal Convolutional Networks. In *Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–21 January 2021*; IEEE: New York, NY, USA, 2021; pp. 16–20. [CrossRef]
32. da Silva, B.; Segers, L.; Braeken, A.; Steenhaut, K.; Touhafi, A. Design Exploration and Performance Strategies towards Power-Efficient FPGA-Based Architectures for Sound Source Localization. *J. Sens.* **2019**, *2019*, 5761235. [CrossRef]
33. Belloch, J.A.; Cobos, M.; González, A.; Quintana-Ortí, E.S. Real-Time Sound Source Localization on an Embedded GPU Using a Spherical Microphone Array. *Procedia Comput. Sci.* **2015**, *51*, 201. [CrossRef]
34. Kealey, J.; Gosselin, A.; Deshaies-Samson, É.; Cardinal, F.; Ducharme-Turcotte, F.; Bergeron, O.; Rioux-Joyal, A.; Bélec, J.; Grondin, F. Real-Time Audio Video Enhancement with a Microphone Array and Headphones. *arXiv* **2023**, arXiv:2303.00949. [CrossRef]
35. Yeow, J.-W.; Tan, E.-L.; Bai, J.; Peksi, S.; Gan, W.-S. Squeeze-and-Excite ResNet-Conformers for Sound Event Localization, Detection, and Distance Estimation for DCASE 2024 Challenge. DCASE 2024 Challenge Technical Report. *arXiv* **2024**. [CrossRef]
36. Gurvich, I.; Leichter, I.; Palle, D.R.; Asher, Y.; Vinnikov, A.; Abramovski, I.; Gopal, V.; Cutler, R.; Krupka, E. A Real-Time Active Speaker Detection System Integrating an Audio-Visual Signal with a Spatial Querying Mechanism. In *Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing, Seoul, Republic of Korea, 14–19 April 2024*; IEEE: New York, NY, USA, 2024; pp. 8781–8785. [CrossRef]
37. Cerutti, G.; Prasad, R.; Brutti, A.; Farella, E. Compact Recurrent Neural Networks for Acoustic Event Detection on Low-Energy Low-Complexity Platforms. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 654–664. [CrossRef]
38. Rosero Jacome, K.G.; Grijalva, F.L.; Masiero, B.S. Sound Events Localization and Detection Using Bio-Inspired Gammatone Filters and Temporal Convolutional Neural Networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 2314–2324. [CrossRef]
39. Shul, Y.; Choi, J.-W. Divided Spectro-Temporal Transformer for Sound Event Localization and Detection in Real Scenes. *J. Acoust. Soc. Am.* **2023**, *154*, A256–A257. [CrossRef]
40. Mu, D.; Zhang, Z.; Yue, H.; Wang, Z.; Tang, J.; Yin, J. SELD-Mamba: Selective State Space Model for Sound Event Localization and Detection with Source Distance Estimation. In *Proceedings of the 2025 Chinese Intelligent Automation Conference, Hefei, China, 4–6 July 2025*; Liu, H., Guo, D., Eds.; Lecture Notes in Electrical Engineering; Springer: Singapore, 2026; Volume 1502, pp. 176–183. [CrossRef]
41. Marhoff, P.; Kujawski, A.; Sarradj, E. Deep Learning for Sound Source Localisation Based on Data from the DCASE Challenge 2020. In *Proceedings of Fortschritte der Akustik—DAGA 2021, Vienna, Austria, 15–18 August 2021*; Deutsche Gesellschaft für Akustik e.V. (DEGA): Berlin, Germany, 2021; pp. 808–811.

42. Cerutti, G.; Andri, R.; Cavigelli, L.; Farella, E.; Magno, M.; Benini, L. Sound Event Detection with Binary Neural Networks on Tightly Power-Constrained IoT Devices. In *Proceedings of the 2020 ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED), Boston, MA, USA, 10–12 August 2020*; ACM: New York, NY, USA, 2020; pp. 19–24. [CrossRef]
43. Srinivas, V.; Itani, M.; Chen, T.; Eskimez, E.S.; Yoshioka, T.; Gollakota, S. Knowledge Boosting during Low-Latency Inference. In *Proceedings of Interspeech 2024, Kos Island, Greece, 1–5 September 2024*; ISCA: Baltimore, MD, USA, 2024; pp. 4338–4342. [CrossRef]
44. Ceolini, E.; Anumula, J.; Braun, S.; Liu, S.-C. Event-Driven Pipeline for Low-Latency Low-Compute Keyword Spotting and Speaker Verification System. In *Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12–17 May 2019*; IEEE: New York, NY, USA, 2019; pp. 7953–7957. [CrossRef]
45. Wang, Y.; Yang, B.; Li, X. IPDnet2: An Efficient and Improved Inter-Channel Phase Difference Estimation Network for Sound Source Localization. *arXiv* **2025**, arXiv:2509.21900. [CrossRef]
46. He, Y.; Trigoni, N.; Markham, A. SoundDet: Polyphonic Moving Sound Event Detection and Localization from Raw Waveform. In *Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021*; PMLR: Cambridge, MA, USA, 2021; Volume 139, pp. 4160–4170. Available online: https://proceedings.mlr.press/v139/he21b.html (accessed on 13 July 2026).
47. dos Santos, O.L.; Rosero, K.; Masiero, B.; Lotufo, R.d.A. w2v-SELD: A Sound Event Localization and Detection Framework for Self-Supervised Spatial Audio Pre-Training. *IEEE Access* **2024**, *12*, 181553–181569. [CrossRef]
48. Belloch, J.A.; Badia, J.M.; Igual, F.D.; Cobos, M. Practical Considerations for Acoustic Source Localization in the IoT Era: Platforms, Energy Efficiency, and Performance. *IEEE Internet Things J.* **2019**, *6*, 5068–5079. [CrossRef]
49. Dementyev, A.; Reddy, C.K.A.; Wisdom, S.; Chatlani, N.; Hershey, J.R.; Lyon, R.F. Towards Sub-Millisecond Latency Real-Time Speech Enhancement Models on Hearables. In *Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing, Hyderabad, India, 6–11 April 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–5. [CrossRef]
50. Belloch, J.A.; Badia, J.M.; Igual, F.D.; Cobos, M.; Quintana-Ortí, E.S. On the Use of a GPU-Accelerated Mobile Device Processor for Sound Source Localization. *Procedia Comput. Sci.* **2017**, *108*, 586–595. [CrossRef]
51. Yarkan, S. A Novel Acoustic Source Localization Technique for Edge AI Applications: A Lightweight Framework and Implementation for IoT and Smart Sensing Devices. *Electrica* **2025**, *25*, 0099. [CrossRef]
52. Munirathinam, R.; Vítek, S. Sound Source Localization and Classification for Emergency Vehicle Siren Detection Using Resource Constrained Systems. In *Proceedings of the 2024 34th International Conference Radioelektronika (RADIOELEKTRONIKA), Zilina, Slovakia, 17–18 April 2024*; IEEE: New York, NY, USA, 2024; pp. 179–183. [CrossRef]
53. Pertilä, P.; Cakir, E.; Hakala, A.; Fagerlund, E.; Virtanen, T.; Politis, A.; Eronen, A. Mobile Microphone Array Speech Detection and Localization in Diverse Everyday Environments. In *Proceedings of the 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 23–27 August 2021*; IEEE: New York, NY, USA, 2021; pp. 406–410. [CrossRef]
54. DiPassio, T.; Heilemann, M.C.; Thompson, B.R.; Rutowski, J.; Bocko, M.F. Toward Smart Acoustic Spaces: Embedded Machine Learning for Sound Event Detection and Classification in the Built Environment. *J. Acoust. Soc. Am.* **2024**, *155*, A282. [CrossRef]
55. Roy, B.B.; Das, S.; Mondal, U.K. TinyML-Driven Sensor Nodes for Energy-Efficient Acoustic Event Detection in Pervasive Acoustic WSNs. *J. Telecommun. Inf. Technol.* **2025**, *100*, 69–77. [CrossRef]
56. Jo, D.; Kim, J.-H.; Jeon, J.; Won, C.S. EgoSep: Egocentric On-Screen Sound Source Separation for Real-Time Edge Computing. *IEEE Access* **2025**, *13*, 6387–6396. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.