Proceedings of the 7th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2027.39477

Accurate, Label-Efficient, and Edge-Deployable Sound Event Detection: Methods for Complex Acoustic Scenes

Guowan Weng

School of Communication and Electronic Engineering, East China Normal University, Shanghai, China
10252150554@stu.ecnu.edu.cn

Abstract. Sound event detection must answer two questions at once: what is sounding, and for how long. This requirement separates sound event detection from clip-level tagging and explains why a detector can perform well on classification while still failing in time-sensitive applications. This review examines the tension between localization quality, learning from incomplete labels, and execution on constrained hardware. Convolutional recurrent networks established useful temporal baselines, whereas attention models and differentiable frame-to-event mappings have improved long-range context and boundary handling. Weak-label disentanglement, pseudo-label training, and separation-guided detection reduce annotation demands, but each introduces biases that remain difficult to measure. Lightweight conformers and spiking networks make edge operation increasingly realistic, although reported efficiency is rarely tied to a complete hardware workload. Across these lines, the field lacks a common basis for deciding whether an F1 gain justifies extra annotation, computation, or complexity. This review argues for evaluation that reports localization, robustness, and resource cost together, and identifies open-set operation, continual adaptation, and reproducible benchmarking as priorities.

Keywords: Sound event detection, weak supervision, semi-supervised learning, lightweight models, edge deployment

1. Introduction

Sound is often the earliest signal available to a monitoring system: glass breaks before an alarm is raised, a person calls for help before a camera is turned toward them, and a machine begins to fail long before its visual appearance changes. Converting such observations into reliable alerts requires more than recognizing a sound class. The detector must also decide where the event begins and ends, since a late onset or an excessively long interval can make the output useless to the downstream system [1, 2]. This combination of recognition and timing is what distinguishes sound event detection (SED) from acoustic scene classification and audio tagging.

The literature has approached this problem along several paths. Convolutional recurrent networks (CRNNs) were among the first systems to combine local time–frequency analysis with explicit temporal context, and their basic design remains a useful control condition for new methods [3]. Lin et al. separated event and background information before classification [4], while Kong et al. used a

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2027.39477

convolutional neural network (CNN)–Transformer model and automatically optimized the decision threshold [5]. Where labels are scarce but unlabeled audio is plentiful, researchers have turned to self-training and local temporal embeddings [6, 7]. Transfer from general-purpose audio representations has also attracted interest, although such features may not transfer equally well to every acoustic domain [8].

A parallel concern is whether these methods can run where the recording is made. The lightweight convolutional iConformer reduces the cost of combining local and longer-range context [9], and event-driven spiking networks offer a route to low-power inference on sensor nodes [10]. Boundary quality has improved through differentiable mappings from frame predictions to event intervals [11], while language-queried source separation provides a way to remove interference before detection [12]. Tutorial and survey papers have documented the datasets, label conventions, and metrics that frame these comparisons [1, 13].

Weak supervision can localize an event incorrectly even when the clip-level decision is right, especially when background sounds co-occur with the target [4, 5]. Dense overlaps and uneven class frequencies further encourage representations that serve common events better than rare ones [6, 7]. At the same time, larger models tend to improve benchmark scores while making always-on inference less attractive [9, 10]. Comparisons become more uncertain when studies choose different label granularity, thresholds, and evaluation tolerance [11, 13].

The review is therefore organized around three questions that cut across model families. How much temporal information can be inferred under a given supervision regime? Which representations and training procedures help when events overlap or labels are incomplete? And what evidence is needed before a model can be described as efficient rather than merely smaller than its predecessor? Section 2 compares the relevant formulations, architectures, learning strategies, and deployment constraints. Section 3 turns the unresolved issues into a research agenda, and Section 4 draws the main conclusions.

2. Background and related work

2.1. SED: problem formulation and evaluation

An SED system assigns class labels to time regions in audio. In the strong-label case, each class has a binary activity trace from which onsets and offsets can be recovered; frame-level classifiers can then be trained directly. A weak label supplies only the presence of a class somewhere in a clip. Real annotation sets frequently sit between these positions: some segments may be labelled precisely, while others are incomplete, uncertain, or absent from the annotation altogether. Because these distinctions change the information available to both the learner and the evaluator, comparisons that ignore them are difficult to interpret [1, 13].

Several metrics are used because no single number describes both recognition and timing. Segment-based F1 measures agreement over fixed intervals. Event-based F1 matches predicted and reference events under onset, offset, and class tolerances. Polyphonic sound detection score (PSDS) summarizes performance over multiple operating points and penalizes cross-triggering [1, 13]. A strikingly high score can therefore reflect generous tolerance, a convenient threshold, or a class distribution dominated by long events. Reporting only the best operating point hides this dependence on post-processing; the choice of threshold is part of the detector, not an external detail. Table 1 summarizes how supervision settings differ in temporal information, learning signal, and evaluation emphasis.

DOI: 10.54254/2755-2721/2027.39477

Table 1. Supervision settings and their evaluation implications

Setting	Temporal information	Typical learning signal	Evaluation emphasis	Main limitation
Strong supervision	Class activity at each frame or annotated onset/offset	Frame-wise binary cross-entropy with strong labels	Event-based F1, segment-based F1, PSDS	Costly annotations; boundary ambiguity
Weak supervision	Clip-level presence only	Multiple-instance or attention-pooled clip classification	Classification and indirectly inferred localization	Weak temporal evidence; threshold and pooling sensitivity
Semi-supervised or partial	Few strong labels; unlabeled audio or incomplete labels	Pseudo-labeling, consistency regularization, partial-loss masking	Detection quality under missing supervision	Confirmation bias; uneven class gains

2.2. Model architectures for sound event detection

2.2.1. Convolutional recurrent networks under strong supervision

CRNNs exploit a natural division of labour: convolutional layers encode local spectral structure, recurrent layers accumulate evidence over time, and a frame classifier turns the resulting state into activity scores [3]. When strong labels are available, recurrence can bridge brief gaps and suppress isolated false activations. Computation proceeds sequentially, long contexts are difficult to summarize compactly, and a single recurrent state may blend simultaneous events rather than preserve their separate identities. CRNN results therefore remain an important reference point, but they should not be treated as evidence that temporal context alone solves polyphonic detection [3, 13].

2.2.2. Attention- and transformer-based detectors

Table 2. Representative sed methods and their trade-offs

Method family	Representative study	Key mechanism	Main strength	Remaining limitation
CNN–RNN	CRNN [3]	Convolution plus recurrent temporal context	Reliable frame-level baseline	Sequential cost; limited handling of mixtures
Disentangled weak SED	Lin et al. [4]	Feature disentanglement and specialized decisions	Reduces background interference	Correlated background sounds remain difficult
CNN–Transformer	Kong et al. [5]	Global attention; automatic thresholding	Better temporal inference from weak labels	Data and computation requirements
Self-training	Cross-referencing network [6]	Pseudo-label generation using multiple views	Uses unlabeled mixtures	Confirmation bias and class imbalance
Lightweight conformer	Convolutional iConformer [9]	Efficient convolution and context modeling	Lower computation for deployment	Accuracy–energy trade-off across hardware
Event-driven network	Spiking detector [10]	Sparse spike-based computation	Low-power continuous sensing	Training complexity and limited ecosystem support

DOI: 10.54254/2755-2721/2027.39477

Table 2. (continued)

Method family	Representative study	Key mechanism	Main strength	Remaining limitation
Separation-guided SED	Language-queried separation [12]	Target separation before detection	Handles noise and interfering events	Separation error; extra computation

Note:Synthesized by the author from the cited studies; entries summarize conceptual trade-offs rather than directly comparable benchmark results.

Table 2 brings these families together, listing for each the representative study, the key mechanism, the main strength and the limitation that remains, so that the trade-offs outlined above can be compared at a glance. Attention lets each time position draw information from a wider acoustic context. CNN–Transformer detectors retain convolutional encoding but replace part of the recurrent pathway with self-attention; the model of Kong et al. also searches for a better operating threshold, linking representation learning to post-processing [5]. Attention may capture long-range relationships without producing sharper onsets, and high-resolution attention is expensive on embedded hardware. Lightweight conformer variants attempt to preserve local convolutional structure while limiting the cost of global aggregation [9]. A different route makes the frame-to-event relationship explicit through differentiable onset and offset objectives, which can improve boundary learning while keeping frame-level evidence interpretable [11].

2.3. Learning from weak and partially labeled data

2.3.1. Weakly labeled learning and disentangled features

A clip-level label says that an event occurred, not which part of the recording supports that conclusion. Multiple-instance and attention-pooling objectives have to discover this correspondence, and they often settle on a salient local peak even when the event occupies a much wider interval. The problem becomes worse when rain, machinery, or room noise appears consistently with a target: the classifier can learn the background as a shortcut. Lin et al. address the issue by disentangling event and background variation and by shaping the decision surface around the target category [4]. Yet disentanglement cannot be complete when the two sources share nearly identical spectral or temporal cues.

2.3.2. Semi-supervised and self-training strategies

Unlabeled recordings are abundant compared with strong event annotations, and self-training is a straightforward way to exploit them. A teacher model proposes pseudo-labels, a student model learns from both confident predictions and labelled examples, and the procedure repeats. Cross-referencing networks reduce dependence on a single teacher and target overlapping events in audio mixtures [6]. Local temporal embeddings take a complementary view: the representation around each frame supplies context that can stabilize pseudo-labels rather than treating frames as isolated examples [7]. Separation-guided detection adds yet another option by isolating a queried source before recognition, although residual separation errors and extra computation can propagate into the detector [12]. The unresolved danger is confirmation bias. A rare class that is initially mislabelled is unlikely to recover unless confidence filtering is paired with class-balanced sampling or an independent teacher.

DOI: 10.54254/2755-2721/2027.39477

2.4. Efficient and edge-deployable sound event detection

Edge deployment imposes constraints absent from standard benchmark training. An always-on detector must share memory, power, and thermal budget with other processes, and a model that is fast in a batch experiment may be unsuitable for continuous streaming. The lightweight convolutional iConformer reduces computation while retaining attention-style context [9], and spiking networks use sparse, event-triggered computation for sensor-node operation [10]. Source separation and high-resolution attention may be useful in difficult acoustic scenes, yet they consume the resources that light models are meant to save [9, 12]. A credible comparison should include parameter count, FLOPs or MACs, peak memory, measured latency, energy per inference, duty cycle, and detection quality on the same workload; otherwise "lightweight" describes a model file, not a deployable system [9, 10].

2.5. Summary of research gaps

Three gaps recur across the studies. The first is temporal: weak labels often produce boundaries that are systematically too broad or too narrow, and existing metrics do not always expose the downstream consequence. The second is representational: overlap and class imbalance encourage shared features that obscure rare events, while pseudo-label methods can amplify the same bias. The third is operational: efficiency claims are made on different devices and workloads, making it difficult to tell whether an architectural simplification actually improves deployment. More reliable temporal evidence could improve pseudo-labels; balanced representation learning could make those labels more useful; and common efficiency reporting would show whether the resulting model remains practical.

3. Open challenges and future directions

The next stage of SED research is less about finding a universally superior architecture than about making trade-offs visible and testable. Four directions follow from the limitations above.

3.1. Pretrained audio foundation models and multimodal cues

Pretrained audio encoders offer a plausible way to compensate for scarce labels, but transfer depends on the relationship between the pretraining corpus and the target scene [8]. A generic embedding may encode what a sound is without representing where it begins or ends. Parameter-efficient adaptation, selective freezing, and uncertainty-aware transfer therefore deserve more attention than simple feature replacement. Language interfaces introduce another possibility: a textual query can specify a target source and has been used to guide separation before detection [12]. The next step is to test whether such queries improve temporal localization under realistic latency and memory constraints, rather than treating semantic flexibility as sufficient evidence of multimodal capability.

3.2. Hardware–software co-design for always-on detection

Pruning, quantization, temporal downsampling, and event-driven execution can all reduce computation, but each changes the detector's sensitivity to short events or quiet classes [9, 10]. Evaluations should therefore report a Pareto surface over accuracy, latency, and energy instead of a single compressed operating point. Reproducibility also requires the measurement conditions to be

DOI: 10.54254/2755-2721/2027.39477

stated: sampling rate, input duration, batch size, processor, power mode, and instrumentation. Without those details, a latency result cannot be transferred from one study to another.

3.3. Open-set, continual, and personalized detection

Closed-set evaluation assumes that deployment will encounter only the classes seen during training. Real homes, streets, and workplaces violate this assumption. A detector may map a new sound onto a familiar class, and repeated updating can overwrite earlier knowledge or expose private audio to unnecessary retention [1, 2]. Open-set rejection, continual learning, and personalization are therefore connected rather than separate features. Progress will depend on evaluation in genuinely new environments and on adaptation rules that can be audited by users.

3.4. Standardized benchmarks and reproducible evaluation

Cross-paper comparison remains fragile when dataset versions, label rules, event tolerances, and thresholds differ. A benchmark report should identify the data split, annotation assumptions, post-processing, operating points, uncertainty, and computational budget [1, 13]. Explicit frame-to-event objectives are useful in this context because they expose how frame evidence becomes an event decision, but they should be compared with conventional metrics at matched cost [11]. Shared protocols and released configurations would make negative results as informative as improvements.

4. Conclusion

SED has advanced through several distinct mechanisms. Strong supervision made CRNNs reliable frame-level baselines; attention and Transformer models expanded temporal context; disentangled representations reduced reliance on background cues; self-training exploited unlabeled mixtures; and lightweight or event-driven architectures moved detection toward constrained devices. None of these advances removes the central difficulty. A model can improve classification while leaving boundaries unstable, and it can reduce parameters while increasing energy use on a particular processor. A model that ranks first on one benchmark may therefore be a poor choice in another room, at another sampling rate, or on another processor.

What matters, then, is not whether one architecture dominates the others, but whether its claims are supported under the conditions that matter for use. Weak-label localization, class-imbalanced learning, and hardware-aware evaluation should be studied together, because progress in one area can change the cost of another. That framing turns architecture choice into an empirical question rather than a matter of fashion. Pretrained models and separation-guided detection are promising extensions, but they should be evaluated against the same localization, robustness, and efficiency criteria as smaller systems. The same rule applies to newer foundation models and separation pipelines: any improvement must be separated from the effects of additional training data, model capacity, or preprocessing.

The review is limited by its conceptual organization: the selected studies use different datasets, annotation conventions, thresholds, and hardware, so their numerical results are not pooled. Its conclusions concern design relationships rather than a definitive ranking. Applications in bioacoustics or music may also require different event definitions and tolerances. These are constraints of the review rather than a claim that the underlying technical comparisons are invalid. Even within this scope, one implication is clear: the field should treat measured resource use,

DOI: 10.54254/2755-2721/2027.39477

reproducible evaluation, and an explicit account of failure modes as core research outcomes, not as details left to implementation.

References

[1] Mesaros, A., Heittola, T., Virtanen, T., & Plumbley, M. D. (2021). Sound event detection: A tutorial. *IEEE Signal Processing Magazine, 38*(5), 67–83. https://doi.org/10.1109/MSP.2021.3090678

[2] Neri, M., Battisti, F., Neri, A., & Carli, M. (2022). Sound event detection for human safety and security in noisy environments. *IEEE Access, 10*, 134230–134240. https://doi.org/10.1109/ACCESS.2022.3231681

[3] Cakir, E., Parascandolo, G., Heittola, T., Huttunen, H., & Virtanen, T. (2017). Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25*(6), 1291–1303. https://doi.org/10.1109/TASLP.2017.2690575

[4] Lin, L., Wang, X., Liu, H., & Qian, Y. (2020). Specialized decision surface and disentangled feature for weakly-supervised polyphonic sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28*, 1466–1478. https://doi.org/10.1109/TASLP.2020.2989575

[5] Kong, Q., Xu, Y., Wang, W., & Plumbley, M. D. (2020). Sound event detection of weakly labelled data with CNN-transformer and automatic threshold optimization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28*, 2450–2460. https://doi.org/10.1109/TASLP.2020.3014737

[6] Park, S., Han, D. K., & Elhilali, M. (2023). Cross-referencing self-training network for sound event detection in audio mixtures. *IEEE Transactions on Multimedia, 25*, 4573–4585. https://doi.org/10.1109/TMM.2022.3178591

[7] Gao, L., Mao, Q., & Dong, M. (2024). On local temporal embedding for semi-supervised sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32*, 1687–1698. https://doi.org/10.1109/TASLP.2024.3369529

[8] Abeßer, J., Grollmisch, S., & Müller, M. (2023). How robust are audio embeddings for polyphonic sound event tagging? *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31*, 2658–2667. https://doi.org/10.1109/TASLP.2023.3293032

[9] Chan, T. K., & Chin, C. S. (2023). Lightweight convolutional-iConformer for sound event detection. *IEEE Transactions on Artificial Intelligence, 4*(4), 910–921. https://doi.org/10.1109/TAI.2022.3173582

[10] Hernández, R. M., & Baeza, V. M. (2026). Event-driven acoustic event detection using spiking neural networks for edge sensor nodes. *IEEE Sensors Journal, 26*(11), 17876–17883. https://doi.org/10.1109/JSEN.2026.3683904

[11] Yoshinaga, T., Tanaka, K., Bando, Y., Imoto, K., & Morishima, S. (2025). Onset-and-offset-aware sound event detection via differentiable frame-to-event mapping. *IEEE Signal Processing Letters, 32*, 186–190. https://doi.org/10.1109/LSP.2024.3509336

[12] Chen, Y., Xiao, Y., Yin, H., Guan, Y., & Liu, X. (2025). Noise-robust sound event detection and counting via language-queried sound separation. *IEEE Signal Processing Letters, 32*, 3974–3978. https://doi.org/10.1109/LSP.2025.3614476

[13] Chan, T. K., & Chin, C. S. (2020). A comprehensive review of polyphonic sound event detection. *IEEE Access, 8*, 103339–103373. https://doi.org/10.1109/ACCESS.2020.2999388