

PSELDNets: Pre-trained Neural Networks on a Large-scale Synthetic Dataset for Sound Event Localization and Detection

Jinbo Hu, *Student Member, IEEE*, Yin Cao, *Member, IEEE*, Ming Wu, *Member, IEEE*,
Fang Kang, *Member, IEEE*, Feiran Yang, *Member, IEEE*, Wenwu Wang, *Senior Member, IEEE*
Mark D. Plumbley, *Fellow, IEEE*, Jun Yang, *Senior Member, IEEE*

Abstract—Sound event localization and detection (SELD) has seen substantial advancements through learning-based methods. These systems, typically trained from scratch on specific datasets, have shown considerable generalization capabilities. Recently, deep neural networks trained on large-scale datasets have achieved remarkable success in the sound event classification (SEC) field, prompting an open question of whether these advances can be extended to the development of SELD foundation models. In this paper, leveraging the power of pre-trained SEC models, we propose pre-trained SELD networks (PSELDNets) on a large-scale synthetic dataset. The synthetic dataset, generated by convolving sound events with simulated spatial room impulse responses (SRIRs), contains 1,167 hours of audio clips with an ontology of 170 sound classes. These PSELDNets are applied to various SELD scenarios. When we adapt PSELDNets to specific scenarios, particularly in cases of low-resource data, we introduce a data-efficient fine-tuning method, AdapterBit. PSELDNets are evaluated on *synthetic-test-set* using collected SRIRs from the TAU Spatial Room Impulse Response Database (TAU-SRIR DB) and achieve satisfactory performance. We also carried out experiments to validate the transferability of PSELDNets to three publicly available datasets and our own real-world recordings. The results demonstrate that PSELDNets surpass state-of-the-art systems across all publicly available datasets. Given the need for direction-of-arrival estimation, SELD generally relies on sufficient multi-channel audio clips. However, incorporating the AdapterBit, PSELDNets show more efficient adaptability to various scenarios using minimal multi-channel or even just monophonic audio clips, outperforming traditional fine-tuning approaches.

Index Terms—Sound event localization and detection (SELD), pre-trained SELD networks, data-efficient fine-tuning.

I. INTRODUCTION

Sound event localization and detection (SELD) combines sound event detection (SED) with direction-of-arrival (DOA)

estimation, with the goal of recognizing the categories, onsets, offsets, and DOAs of various sound sources. SELD frameworks represent audio sources in both spatial and temporal domains, making them suitable for applications such as robot listening, audio surveillance, and smart home environments.

A. Existing learning-based SELD methods

In recent years, there have been notable advancements in learning-based SELD methods. Adavanne et al. [1] introduced SELDnet, an end-to-end network designed for simultaneous sound event detection and DOA estimation. Nevertheless, SELDnet faces challenges in identifying overlapping sound events of the same class from different locations. To address this homogenous overlap issue, the Event-Independent Network V2 (EINV2) is proposed [2]–[4]. EINV2 uses a track-wise output format and permutation invariant training to predict a single sound event and its corresponding DOA for each track. Unlike SELDnet and EINV2, the Activity-coupled Cartesian DOA (ACCD OA) combines SED and DOA tasks into a single output and embeds activity information into Cartesian DOA vectors [5]. The Multi-ACCD OA (mACCD OA) [6] extends ACCD OA by incorporating a track-wise output format and employing auxiliary duplicated permutation invariant training to address the homogenous overlap problem.

On the other hand, numerous learning-based SELD investigations [2]–[11] have predominantly utilized synthetic datasets from SELD challenge events [12]–[17], showing promising performance in both simulated and real spatial environments. However, these systems have two limitations. Firstly, the target sound event classes that the systems predict must be pre-specified before training, posing a challenge since each application scenario may require different target classes. Secondly, learning-based SELD approaches can suffer from performance degradation when exposed to acoustic environments not encountered during training, i.e., a phenomenon known as environment shift [18].

One of the effective ways to address the problems of unknown sound event classes and unseen acoustic environments is by acquiring significant scenario-specific data for training. However, creating spatial sound event signals is a complex task involving extensive data collection and computational generation. This process requires convolving dry sound source signals with measured spatial room impulse

Jinbo Hu, Ming Wu, and Jun Yang are with the Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China (e-mail: hujinbo@mail.ioa.ac.cn; mingwu@mail.ioa.ac.cn; jyang@mail.ioa.ac.cn). Jinbo Hu and Jun Yang are also with the University of Chinese Academy of Sciences, Beijing 100049, China. (Corresponding author: Jun Yang, Yin Cao)

Yin Cao is with the Department of Intelligent Science, Xi'an Jiaotong Liverpool University, Suzhou 215123, China (e-mail: yin.k.cao@gmail.com).

Fang Kang is with the Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Oulu 90570, Finland (e-mail: fang.kang@oulu.fi).

Feiran Yang is with the State Key Laboratory of Acoustics, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China (feiran@mail.ioa.ac.cn).

Wenwu Wang and Mark D. Plumbley are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: w.wang@surrey.ac.uk, m.plumbley@surrey.ac.uk).

This work was supported by the National Key Research and Development Project (NO. 2022YFB2602003), Grant XJTLU RDF-22-01-084, and Engineering and Physical Sciences Research Council (EPSRC) Grant EP/T019751/1.

responses (SRIRs). Moreover, manually collecting and annotating real-world spatial sound event recordings is very costly, and publicly accessible real-scene SELD data is limited [19], [20]. To mitigate these challenges, the zero-and-few-shot SELD system [21] and environment-adaptive Meta-SELD [18], [22] utilize pre-trained models to operate effectively with limited data. Despite these developments, there is still a notable lack of foundation models for SELD. In contrast, several foundation models have recently been developed [23]–[26] for sound event classification (SEC), which focus on identifying the categories of sound events without concerning their timestamps (i.e., onset and offset). These SEC models are highly relevant to SELD tasks, but their potential benefits for the SELD systems are not adequately investigated.

B. Foundation models in SEC

Deep neural networks have made substantial strides in SEC research [23]–[26]. A key milestone was the introduction of AudioSet [27], which is a comprehensive dataset featuring over 2 million human-annotated 10-second audio clips and an ontology of 527 sound classes, and utilized for general-purpose sound event recognition. Convolutional neural networks (CNNs), exemplified by Pre-trained Audio Neural Networks (PANNs) [23], extract local features from audio spectrograms and enhance performance by optimizing the network's depth and breadth.

Recently, Transformer architectures [28], which have proven effective in sequence modeling, have been adapted to computer vision by partitioning images into smaller patches [29], [30]. Inspired by these approaches, several studies, such as the Audio Spectrogram Transformer (AST) [24], the Patchout faSt Spectrogram Transformer (PaSST) [25], and the Hierarchical Token-Semantic Audio Transformer (HTS-AT) [26], apply purely attention-based models to audio spectrograms to capture long-range global context. AST [24] leverages the self-attention mechanism, overlapping patches from audio spectrograms, and pre-trained parameters from computer vision to build the first convolution-free model for SEC. Drawing inspiration from SpecAugment [31] and the mask technique used in Bidirectional Encoder Representations from Transformers (BERT) [32], PaSST [25] offers an efficient implementation of AST by omitting segments of the Transformer's input sequence during training. This method encourages the Transformer to classify the events using an incomplete sequence. In comparison, HTS-AT [26] uses Swin Transformer blocks with shifted window attention [30], enhancing efficiency by limiting self-attention calculations to local non-overlapping windows while allowing cross-window connections. These models achieve state-of-the-art (SOTA) SEC results on AudioSet.

Furthermore, these models, which were pre-trained on large-scale datasets, offer the potential for transferability to other audio-related tasks to further improve performance [23]–[26], such as acoustic scene classification, music genre classification, and speech emotion classification. Nevertheless, the efficient transfer of these pre-trained models to various audio tasks remains challenging. One common method for adapting

the pre-trained models to these tasks involves fine-tuning all the parameters of the pre-trained models using the datasets designed for these tasks. However, this technique requires significant computational resources and memory capacity. On the other hand, it can result in a loss of model generalization, possibly due to catastrophic interference among tasks [33].

C. Parameter-efficient fine-tuning

To mitigate the challenges associated with efficient transfer, the parameter-efficient fine-tuning (PEFT) methodology, which only fine-tunes a small number of (extra) parameters to achieve strong performance, has been extensively investigated across the domains of natural language processing [34]–[37] and computer vision [33], [38]–[40].

Prominent PEFT methods include Low-Rank Adaptation (LoRA) [34], Adapter tuning [33], [37], [38], prompt tuning [40], and others. The fundamental principle of these PEFT methodologies involves freezing the primary or all pre-trained parameters and introducing additional trainable parameters for fine-tuning. Expanding on these PEFT techniques, various researchers working on audio processing have integrated some model-specific Adapters into their frameworks [41]–[44]. The Adapter, a straightforward plug-and-play module, is designed for attention-based networks and entails incorporating a few lightweight bottleneck networks into the Transformer layers. These methodologies retain the generality of the pre-trained model, conserve computational resources, reduce data requirements, and attain competitive or even superior performance.

D. Our contributions

In this study, we endeavor to develop SELD foundation models that can be applied to various real-world scenarios. We introduce pre-trained SELD networks (PSELDNets) on a large-scale synthetic dataset. This dataset, comprising approximately 1,167 hours of audio clips and featuring an ontology of 170 sound classes, is scenario-agnostic and generated by convolving sound event clips from FSD50K [45] with simulated SRIRs. The PSELDNets, inheriting the architectures of our previously proposed CNN-Conformer [4], [9] and pre-trained models that achieve SOTA results in SEC, such as PANNs [23], PaSST [25] and HTS-AT [26], extract spatial and global features from multi-channel spectrograms. Unlike common model architectures tailored for the SELD problem, e.g., CRNN [8] and CNN-attention [4], [10], to our knowledge, PSELDNets represent the first SELD model that exploits the Transformer model. We evaluate the performance of PSELDNets on *synthetic-test-set* that uses measured SRIRs from TAU Spatial Room Impulse Response Database (TAU-SRIR DB) [46] and obtain satisfactory performance.

We transfer PSELDNets to multiple publicly available datasets, including the Detection and Classification of Acoustic Scenes and Events (DCASE) 2021 Challenge Task 3 [14], the L3DAS22 Challenge Task 2 [16], the Sony-TAU Realistic Spatial Soundscapes 2023 (STARSS23) [20] dataset, and our own audio recordings. Experimental results demonstrate the transferability of PSELDNets, showing that the models consistently exceed SOTA benchmarks [4], [10], [47]–[49] across all these publicly available datasets.

Inspired by PEFT techniques [33], [36], [38], we introduce a data-efficient fine-tuning method for SELD, AdapterBit, which fine-tunes only the additionally inserted multi-layer perceptron (MLP) Adapter and bias terms, enabling the efficient utilization of low-resource data, including minimal multi-channel or even monophonic clips. By employing AdapterBit for transfer learning in specific SELD scenarios using low-resource data, PSELDNets exhibit superior performance, compared to the conventional full fine-tuning methods. Notably, when utilizing monophonic clips, pseudo-multi-channel clips are generated by convolving the audio sources with the theoretical responses of the microphone array to ensure compatibility with the input to PSELDNets.

The contribution of this work includes:

- 1) Synthesizing a large-scale SELD dataset designed to include numerous sound event instances and various acoustic environments.
- 2) Introducing PSELDNets trained on the large-scale synthetic SELD dataset to develop foundation models.
- 3) Transferring PSELDNets to several SELD scenarios and achieving SOTA performance.
- 4) Proposing a data-efficient fine-tuning technique to adapt PSELDNets to specific scenarios using limited data.
- 5) Releasing the source code, the pre-trained parameters of PSELDNets, and the large-scale synthetic SELD dataset¹.

II. DATA SYNTHESIS

The synthesis of SELD clips is achieved through the convolution of clean sound event clips from FSD50K [45] with simulated SRIRs. The important components for accurately simulating spatial sound event recordings are the acquisition of high-quality sound event clips and SRIRs.

A. Sound event clips

Various efforts have been dedicated to developing audio datasets for SEC [27], [45], [50]–[55]. In this study, we select sound event clips based on the AudioSet Ontology [27] and emphasize strong labeling, single-source clips and high label quality.

1) *AudioSet Ontology*: We focus on creating SELD foundation models. To meet this objective, the selected classes need to cover a comprehensive range of everyday sounds and be scalable regarding data and vocabulary. Accordingly, we use the AudioSet Ontology² for organizing our data. The AudioSet Ontology includes 632 sound event classes, arranged hierarchically with up to 6 levels, and encompasses a variety of everyday sounds. The class annotations in the datasets, e.g., AudioSet [27], FSDnoisy18K [50], FSDKaggle2019 [51], and FSD50K [45], make use of the vocabularies provided by this ontology.

2) *Strong data labeling*: The SELD task, akin to the SED task, necessitates predicting the exact start and end times of sound events, which is essential for accurately predicting the trajectory of moving sound sources. Nonetheless, audio

datasets that provide such detailed timestamp annotations [54], [56], i.e., strong labels, are quite rare. Most datasets provide annotations at the clip level without precise timestamps, i.e., weak labels [27], [45], [50]–[53], [55].

3) *Single-source clips*: Individual sound sources are spatially isolated to be distinguished in SELD. Conversely, typical audio datasets usually include audio clips annotated with multiple class labels, indicating that each clip may contain several overlapping sound events or a single event with hierarchical propagation labels [27], [45]. When synthesizing SELD clips, each selected sound event clip must contain only a single sound source at a time to guarantee an accurate representation of spatialized sound events.

4) *Label quality*: The quality of datasets is important for model performance. Early studies on audio datasets often relied on small and exhaustively annotated datasets [52], [54], [55], and as large-scale datasets like AudioSet [27] have emerged, label inaccuracies have become more common [45], due to the impracticality of exhaustive manual annotation. Some datasets [50], [51] focus on learning under noisy label conditions, which is out of the scope of this work. Label accuracy remains critical for ensuring model reliability, as noisy labels can introduce interference and lead to performance degradation [57]. This work prioritizes label quality while maintaining a substantial dataset size.

Based on the above considerations, we select single-source clips from FSD50K [45] for synthesis. FSD50K encompasses a collection of 51,197 audio clips totaling 108 hours, manually labeled with 200 classes derived from the AudioSet Ontology. Despite its weak labeling, FSD50K exhibits a high label density [58]. Label density refers to the portion of the recording duration during which the annotated sound event is actually present. The clips with high label density allow us to treat sound events throughout the entire clip as active. Notably, strong annotations in AudioSet [56] are not used for data synthesis, due to significant imbalance and incompleteness in certain classes.

B. Spatial room impulse responses

SELD generally necessitates multi-channel audio inputs for effective source localization. First-order ambisonics (FOA), well known as an array-agnostic format, is widely employed in various SELD datasets [12]–[17], [19], [20], [59]. Numerous advanced methods utilize FOA signals, instead of original microphone-array-format signals, to achieve SOTA results [10], [11], [48], [60]. Additionally, several studies [61], [62] explore ambisonics encoding of arbitrary microphone arrays, yielding competitive performance. Therefore, we employ FOA-format SRIRs to synthesize SELD clips.

Ambisonics represents a data format that decomposes a sound field on an orthogonal basis of spherical harmonic functions. This format is typically derived by converting the spherical microphone array signals [63]. The FOA signal comprises four channels (W, Y, Z, X) with W denoting an omnidirectional microphone and (Y, Z, X) referring to three

¹<https://github.com/Jinbo-Hu/PSELDNets>

²<https://research.google.com/audioset/ontology/index.html>

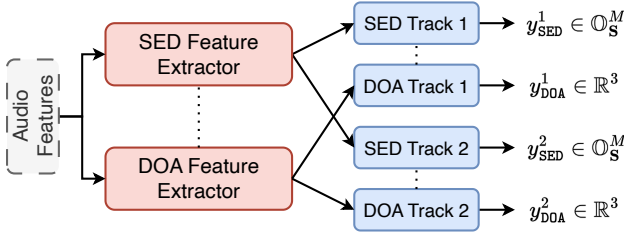


Fig. 1: The architecture of EINV2. The solid line boxes represent trainable neural networks. The dotted lines are learnable parameters connecting two branches.

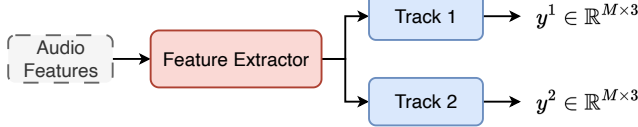


Fig. 2: The mACCDOA representation of the SELD model. There is no track dimension in the ACCDOA representation.

bidirectional microphones aligned along the Cartesian axes. The theoretical spatial responses of FOA are [12], [63], [64]:

$$\begin{aligned} H_1(\phi, \theta) &= 1, \\ H_2(\phi, \theta) &= \sin(\phi) \cos(\theta), \\ H_3(\phi, \theta) &= \sin(\theta), \\ H_4(\phi, \theta) &= \cos(\phi) \cos(\theta), \end{aligned} \quad (1)$$

where θ and ϕ denote elevation and azimuth.

The computational generation method for FOA-format SRIRs involves a two-step process: microphone-array RIRs simulation and ambisonics format conversion. Microphone-array RIRs are generated using the image source method [65]. The procedure for converting microphone-array signals to FOA signals can be found in [18], [63], [64], [66], [67].

III. THE SELD SYSTEMS

A. Related SELD methods

We introduce two existing learning-based SELD methodologies: Event-Independent Network V2 (EINV2) [2]–[4] and Activity-coupled Cartesian DOA (ACCDOA) [5], [6]. Both approaches perform frame-wise prediction of the temporal activity and spatial trajectory of sound events. For simplicity, subsequent illustrations omit the temporal dimension and instead focus on the SELD output format for individual frames.

1) *EINV2*: EINV2 [3] comprises two branches: SED and DOA estimation, which are linked through a soft parameter-sharing strategy, e.g., multiple sets of learnable parameters. Each branch has multiple event-independent tracks, forming track pairs. Each pair can only predict a sound event with its corresponding DOA. Permutation-invariant training is used to handle the track misalignment between the ground truth and the prediction. The architecture is shown in Fig. 1. For the i -th track, y_{SED}^i indicates one-hot encoding of M sound event classes in the set \mathcal{S} , and y_{DOA}^i represents Cartesian DOA output. The number of tracks depends on the maximum number of overlapping events.

2) *ACCDOA*: The ACCDOA approach represents the presence of a sound event through the amplitude of its corresponding Cartesian DOA vector. Unlike EINV2, the ACCDOA representation merges two branches into one, thereby obviating the necessity of balancing the loss between SED and DOA branches and avoiding an increase in model parameters.

However, the ACCDOA representation cannot detect multiple instances of the same event type from various locations. To address this issue, the mACCDOA representation has been proposed [6]. mACCDOA integrates both class-wise and track-wise output formats, as shown in Fig. 2. Additionally, auxiliary duplicating permutation invariant training (ADPIT) is introduced to tackle problems of track misalignment and sparse target outputs.

B. Network architectures

The SEC field has seen substantial advancements due to deep neural networks [23]–[26] and large-scale datasets [27], [45]. Utilizing pre-trained models that exhibit superior performance in SEC, e.g., PANNs [23], AST [24], PaSST [25] and HTS-AT [26], may improve the SELD task. Consequently, the structures of PSELDNets align with these pre-trained SEC models for effective transfer learning. PSELDNets take as input the concatenation of log-mel spectrograms and intensity vectors extracted from FOA signals, and predict active sound events with corresponding DOA vectors for each timestamp, adhering to SELD output formats described in Sec. III-A.

1) *PANNs*: PANNs [23] are convolution-based models trained from scratch on AudioSet. Current SELD techniques predominantly utilize CNN-attention hybrid models, which have demonstrated superior performance, e.g. ResNet-Conformer [10], [11] and our previously proposed CNN-Conformer [4], [9].

Following the above model architectures, we extend CNN-Conformer to CNN14-Conformer, where a Conformer block [68] follows the main body of the CNN14 [23] module. CNN14 contains a stack of 6 VGG-like [69] CNN blocks, and the Conformer comprises two feed-forward layers with residual connections sandwiching the multi-head self-attention and convolution modules. The CNN block extracts local fine-grained features, while the Conformer block captures both local and global context dependencies in an audio sequence.

2) *PaSST*: PaSST [25] is an advanced and efficient variant of AST [24]. The overall architecture of PaSST is shown in Fig. 3, and the specific structure of the attention-based blocks is detailed in Fig. 4(a). The 2D audio spectrogram is split into a sequence of overlapping patches, which are subsequently flattened and linearly projected to a sequence of 1D patch embeddings. These embeddings are processed using a standard Transformer Encoder [28]. Additionally, in analogy to the mask technique used in BERT [32] and SpecAugment [31], PaSST adopts unstructured Patchout and structured Patchout, where unstructured Patchout randomly omits patches from any position, and structured Patchout picks some frequency bins or time frames and remove the corresponding row or column of extracted patches. These two approaches improve generalization and reduce computation and memory complexity.

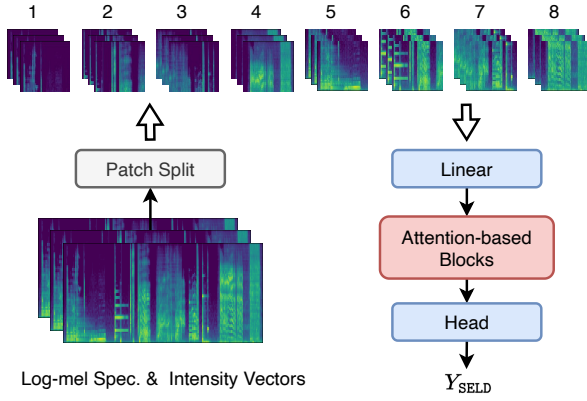


Fig. 3: The architecture of the purely attention-based network.

We employ PaSST with the structured Patchout approach [25] in the frequency axis to ensure the valid output of each time frame. PaSST utilizes the classification and distillation token mechanism for label prediction, inherently preventing it from predicting event start and end times in audio clips. The output embeddings from the Transformer layer contain temporal information for each frame. To tackle the prediction of sound events with precise timestamps, we project the final layer embeddings into SELD output formats via a linear head.

3) *HTS-AT*: HTS-AT [26] combines the Swin Transformer [30] and a token-semantic module [70]. The Swin Transformer focuses on self-attention within each local window, which comprises fixed-size and non-overlapping patches. A key design element of the Swin Transformer is the shifted window attention across successive self-attention layers, introducing connections between neighboring non-overlapping windows from the preceding layer. Moreover, Swin Transformer builds hierarchical feature maps by gradually merging neighboring patches in deeper Transformer layers to reduce the sequence size. The token-semantic module [70] employs a simple convolution layer as the head layer and converts output feature maps from the final Swin Transformer Block into activation maps for the prediction of each timestamp. The details of HTS-AT are depicted in Fig. 3 and Fig. 4(b), with the detailed architecture of the Swin Transformer Block being identical to the standard Transformer Encoder in Fig. 4(a).

C. Data augmentation

Data augmentation is a valuable technique for improving the generalization capabilities of a system. Given the successful application of our previously proposed data augmentation chains [4], [9] in L3DAS22 Task 2 [16] and DCASE 2022 Task 3 [19], we adopt this technique to increase the data diversity.

Each data augmentation chain comprises various augmentation operations, which are randomly selected and linked together. Following the methodology described in [4], [9], we randomly sample $k = 3$ augmentation chains and select Mixup [71], Cutout [72], SpecAugment [31], and frequency shifting [8] as data augmentation operations. Additionally, we employ the rotation of FOA signals [10], [73] as an independent

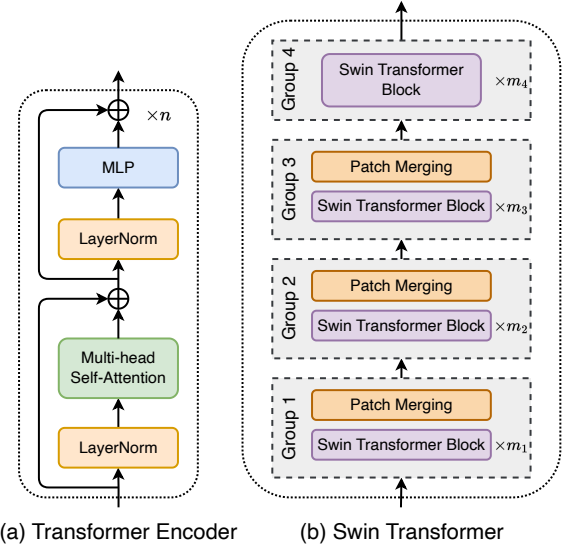


Fig. 4: The detailed architecture of the attention-based blocks in Fig. 3. The Transformer Encoder is employed in AST and PaSST, while the Swin Transformer is employed in HTS-AT.

spatial augmentation operation, which is not a part of data augmentation chains.

IV. DATA-EFFICIENT FINE-TUNING

Fine-tuning involves deploying a pre-trained model for a new task, where all parameters are initialized from the pre-trained model, with the possible exception of the head layer. The conventional full fine-tuning approach may result in the loss of model generalization, potentially due to catastrophic interference among tasks [33]. Inspired by PEFT techniques [33], [36]–[38], we introduce a data-efficient fine-tuning strategy, AdapterBit.

SELD generally necessitates multi-channel audio inputs for source localization. By utilizing AdapterBit, PSELDNets can be more efficiently adapted to various SELD scenarios using limited data, with a particular emphasis on the monophonic sound event clips. Specifically, when employing monophonic signals for fine-tuning, we generate pseudo-FOA signals based on theoretical responses of the employed microphone array to align with the input to PSELDNets.

A. AdapterBit

We design the AdapterBit structure as depicted in Fig. 5. AdapterBit integrates MLP Adapters following the architecture outlined in [33], [38], [41], while additionally introducing bias-term tuning. In bias-term tuning, only the bias terms from the pre-trained model, such as those in linear and convolution layers, are fine-tuned. The Adapter is constructed with a Gaussian Error Linear Unit (GELU) non-linearity positioned between two linear layers, which is subsequently integrated into the standard Transformer Encoder layer through a scaling factor s . The scaling factor serves to balance the general features produced by the frozen branch and the scenario-specific features produced by the trainable branch. For a given

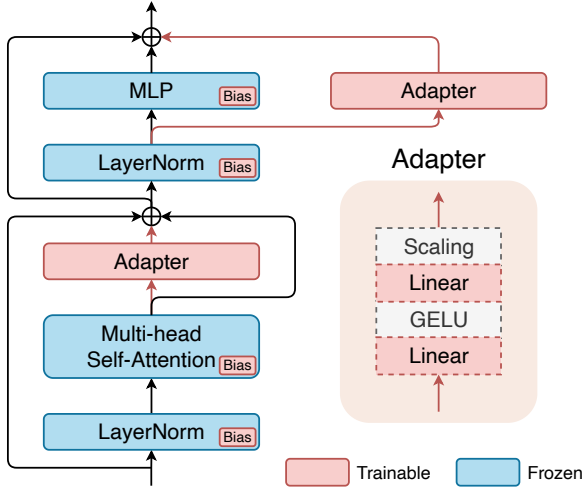


Fig. 5: The brief illustration of AdapterBit.

input feature x_l , the Adapter produces adapted features \tilde{x}_l as follows:

$$\tilde{x}_l = s \cdot (W_2 \cdot \text{GELU}(W_1 \cdot x_l + b_1) + b_2), \quad (2)$$

where W_1 , W_2 , b_1 and b_2 denote the weights and biases of the linear layers. The parameters W_1 and b_1 are randomly initialized, while W_2 and b_2 are set to zero. The rationale behind zero initialization is to ensure that the adapted model keeps the same state as the pre-trained model at the beginning of fine-tuning by utilizing the parallel design and the residual connection of the Adapter.

During the fine-tuning stage, we focus on optimizing just the newly added parameters and the bias term from PSELD-Nets, while keeping the other parameters frozen. Specifically, PSELDNets initializes its weights from the pre-trained checkpoint and maintains all parameters, except the bias terms, in a frozen state. The Adapters and the bias term of the adapted model are updated using the SELD data from the specific scenario, such as different microphone array types, acoustic environments and sound event categories. During the inference phase, we reload all pre-trained parameters, including those that were previously kept frozen, along with the newly inserted and fine-tuned parameters.

B. Pseudo-FOA signals

SELD generally necessitates multi-channel audio input to enable effective DOA estimation, e.g., four-channel FOA signals input to PSELDNets. To fulfil this input requirement, we can generate pseudo-FOA signals from monophonic signals by utilizing the theoretical responses of FOA described in Eq. (1). Additionally, the monophonic signals must contain non-overlapping sound events to satisfy the single-source clip requirement described in Sec. II-A3.

The pseudo-FOA signals are obtained as follows:

$$\begin{bmatrix} W(t, f) \\ Y(t, f) \\ Z(t, f) \\ X(t, f) \end{bmatrix} = \begin{bmatrix} 1 \\ \sin(\phi) \cos(\theta) \\ \sin(\theta) \\ \cos(\phi) \cos(\theta) \end{bmatrix} S(t, f), \quad (3)$$

where $S(t, f)$ denotes the monophonic signal spectrogram. ϕ and θ can be randomly sampled from the desired distribution of the azimuth and elevation.

The pseudo-FOA signals can be considered a form of regularization for input monophonic signals, as they preserve information related to sound events while mitigating the loss of inter-channel connections wherever possible.

V. EXPERIMENTAL SETUPS

A. Datasets

Audio clips from FSD50K [45] are selected according to the criteria in Sec. II-A. We select single-source sound event clips and filter out the classes with less than 30 clips and those that pose recognition challenges. As a result, a total of 170 classes are selected. The selected audio clips comprise 31,444 samples for training, amounting to 43.4 hours, and an additional 3,701 samples for testing, totaling 5.3 hours.

SRIRs are mainly generated via simulation [74]. We simulate diverse shoebox-shaped rooms employing frequency-dependent absorption coefficients. This approach avoids the requirement for sampling from a distribution of reverberation times and estimating absorption coefficient values, thereby preventing unrealistic scenarios such as long reverberation times in small rooms. Absorption materials from typical acoustic material databases^{3,4} are randomly allocated to the wall, ceiling and floor surfaces of each simulated room. We use these simulated SRIRs to synthesize spatialized static sound sources for training. To generate overlapping data, we convolve each monophonic single-source clip with SRIRs from different spatial locations within the same room configuration and then mix the resulting signals. Furthermore, additional spatialized sound events, including moving sources, are synthesized using collected SRIRs from TAU-SRIR DB [46] to evaluate the simulated SRIRs and various network architectures. We adopt the publicly available code for data synthesis⁵.

In total, we synthesize 67,000 1-minute clips amounting to approximately 1,117 hours for training, where each clip is simulated using a unique room configuration, termed as *synthetic-training-set*. Additionally, we synthesize 3,060 1-minute clips amounting to roughly 51 hours for testing, denoted as *synthetic-test-set*. The distribution of maximum polyphony of 1, 2, and 3 in the synthetic dataset follows a ratio of approximately 10:5:2.

B. Hyper-parameters

The sampling rate is 24 kHz. We extract 64-dimension log mel spectrograms from FOA signals using a Hanning window of 1024 points and a hop size of 240. Each audio clip is segmented to a fixed duration of ten seconds for training and inference. All hyper-parameters of PSELDNets are consistent with those in the pre-trained SEC models [23], [25], [26]. The number of event-independent tracks is 3 in EIN2 and mACCDOA. A batch size of 32 and the AdamW

³<https://www.acoustic-supplies.com/absorption-coefficient-chart/>

⁴<https://pyroomacoustics.readthedocs.io/en/pyri-release/pyroomacoustics.materials.database.html>

⁵<https://github.com/Jinbo-Hu/SELD-Data-Generator>

TABLE I: Description of the utilized SELD models involving various stages.

Method	Initial parameters	Datasets
Pre-train	Pre-trained SEC checkpoints	<i>synthetic-training-set</i> (1,117 h) One dataset from $\{L3DAS22 \text{ Task 2 (7.5 h), DCASE 2021 Task 3 (13.3 h), STARSS23 (7.5 h), scenario-specific synthetic datasets (1 to 45 h)}\}$ for different experiments
From-scratch	Pre-trained SEC checkpoints	
Fine-tune	Pre-trained SELD checkpoints	
AdapterBit		

[75] optimizer are employed for training. The learning rate is set to 10^{-4} for the first 20 epochs and subsequently decreased to 10^{-5} for the following 5 epochs.

C. Evaluation metrics

We use a joint metric of localization and detection [76], [77] in this work: two location-dependent detection metrics, F-score (F_{20°) and error rate (ER_{20°), and two class-dependent localization metrics, localization recall (LR_{CD}) and localization error (LE_{CD}). In contrast to the standard computation of F-score and error rate in SED [78], F_{20° and ER_{20° consider a prediction a true positive only if the sound event is accurately detected and its estimated direction lies within a spatial threshold of 20° away from the ground truth, while the predictions falling outside this threshold are treated as false positives. LE_{CD} and LR_{CD} compute the mean angular error and true positive rate in the case when sound event classes are predicted correctly. Note that LR_{CD} can also be interpreted as the unthresholded recall.

We use an aggregated SELD metric for the method comparison and hyper-parameter selection:

$$\mathcal{E}_{\text{SELD}} = \frac{1}{4} \left[ER_{20^\circ} + (1 - F_{20^\circ}) + \frac{LE_{CD}}{180^\circ} + (1 - LR_{CD}) \right]. \quad (4)$$

For comparison and consistency across different task setups, a macro-average of F_{20° , LR_{CD} , LE_{CD} , and $\mathcal{E}_{\text{SELD}}$ across classes is utilized in STARSS23 and *synthetic-test-set*, and a micro-average of those metrics across instances is employed in other datasets. An effective system should demonstrate a low ER_{20° , a high F_{20° , a low LE_{CD} , a high LR_{CD} , and a low $\mathcal{E}_{\text{SELD}}$.

VI. EXPERIMENTS

Firstly, the performance of PSELDNets is evaluated on *synthetic-test-set*, investigating various networks and SELD output formats. Secondly, PSELDNets are transferred to multiple publicly available datasets. Subsequently, the efficiency of the data-efficient fine-tuning approach is validated on low-resource data. Finally, the effectiveness of PSELDNets and the data-efficient fine-tuning approach are tested using our own audio recordings, termed *Indoor Recordings*.

The SELD models involving various stages are presented in Table I. If not specified, the Pre-train method in this work denotes the SELD models trained on *synthetic-training-set*, initialized with pre-trained SEC checkpoints, such as those from PANNs, PaSST or HTS-AT. The From-scratch, Fine-tune, and AdapterBit methods represent different strategies for applying these SELD models to specific SELD scenarios. The From-scratch method utilizes pre-trained SEC checkpoints, specifically HTS-AT in this work, without using any pre-trained SELD checkpoints. In contrast, Fine-tune and AdapterBit methods directly utilize the pre-trained SELD checkpoints,

TABLE II: Results of various networks with the mACCD OA representations.

Network	# Params	$ER_{20^\circ} \downarrow$	$F_{20^\circ} \uparrow$	$LE_{CD} \downarrow$	$LR_{CD} \uparrow$	$\mathcal{E}_{\text{SELD}} \downarrow$
CNN14-Conformer	179.4 M	0.805	25.4%	17.3°	32.1%	0.582
PaSST	52.3 M	0.773	29.2%	17.6°	33.2%	0.562
HTS-AT	34.6 M	0.784	27.6%	17.6°	33.9%	0.567

TABLE III: Results of HTS-AT with various SELD methods.

Method	$ER_{20^\circ} \downarrow$	$F_{20^\circ} \uparrow$	$LE_{CD} \downarrow$	$LR_{CD} \uparrow$	$\mathcal{E}_{\text{SELD}} \downarrow$
ACCD OA	0.777	27.9%	17.1°	33.0%	0.566
mACCD OA	0.784	27.6%	17.6°	33.9%	0.567
EINV2	0.801	24.7%	15.4°	25.3%	0.597

specifically those integrating HTS-AT with the mACCD OA output format in this work.

A. Results on the synthetic dataset

1) *Network architecture*: We evaluate the performance of CNN14-Conformer, PaSST and HTS-AT, and choose the mACCD OA [6] representations as the main SELD output formats. Each network employs its respective pre-trained SEC checkpoints [23], [25], [26], excluding the additional Conformer [68] module, which is randomly initialized due to the absence of an appropriate pre-trained checkpoint.

Table II presents the comparison of various networks. We observe that PSELDNets achieve LR_{CD} of over 32% and LE_{CD} of approximately 17° on a macro-average across all classes. Notably, when comparing the result on *synthetic-test-set* with those on scenario-specific datasets presented in Tables IV, V and VI, the performance gap in localization can be attributed to the discrepancy of SRIRs and source motion characteristics between *synthetic-training-set* and *synthetic-test-set*, as illustrated in Sec. V-A, whereas the performance gap in detection is primarily due to the diversity of sound events, spanning a total of 170 classes, as well as label noise, including imprecise timestamps and incorrect sound event annotations in FSD50K.

Although CNN14-Conformer has a substantial number of parameters, it performs the worst among the three networks, possibly due to the challenges in optimizing such a large model. Compared to PaSST, HTS-AT achieves similar performance but with fewer parameters. Consequently, we select HTS-AT as the baseline model for further investigation.

2) *SELD methods*: We evaluate the performance of three SELD methods employing HTS-AT: ACCD OA representations [5], mACCD OA representations [6], and EINV2 [3]. Within the EINV2 method, we utilize two relatively independent SED and DOA branches, both with identical architectures and pre-trained checkpoints. These branches are connected through

TABLE IV: Results on L3DAS22 Task 2.

Method	Aug.	ER _{20°} ↓	F _{20°} ↑	LE _{CD} ↓	LR _{CD} ↑	ℰ _{SELD} ↓
(#1) Hu et al. [4]	✓	0.437	65.1%	11.9°	73.2%	0.280
From-scratch	✗	0.617	49.6%	17.9°	67.6%	0.386
	✓	0.445	63.8%	13.4°	72.4%	0.289
Fine-tune	✗	0.370	70.6%	11.6°	79.0%	0.235
	✓	0.361	70.3%	12.2°	77.0%	0.239
Fine-tune *	✓	0.330	73.6%	11.3°	80.4%	0.213

* denotes post-processing methods.

several sets of trainable parameters [3] following each group of Swin Transformers [30] as illustrated in Fig. 4(b).

The results of various SELD methods are presented in Table III. Among the three methods, the EINV2 method exhibits the worst performance, especially in detection. The EINV2 method has considerably worse LR_{CD} but significantly better LE_{CD} compared to the ACCDOA-based methods. One potential explanation for this discrepancy is that EINV2 utilizes multiple tracks, with each track predicting only one of 170 event classes, resulting in sparse outputs in the SED branch. In contrast, the mACCDOA representation is trained using auxiliary duplicated permutation invariant training (ADPIT), enabling each track to learn with the same target as the ACCDOA format. This training mechanism leads to dense outputs for mACCDOA, thereby achieving performance similar to that of ACCDOA.

B. Transfer to various scenarios

In this section, we investigate one application of PSELD-Nets. We employ HTS-AT with mACCDOA for transfer learning and apply it to several SELD scenarios using the full fine-tuning method. Notably, some systems only report their results using either ensemble or single models with post-processing. For a fair comparison, we adopt a post-processing method containing moving average (MA) and dynamic threshold (DT). During inference, test samples are segmented into 10-second clips with a 0.5-second hop length. The results for each 0.5-second segment are averaged across all corresponding time-overlapped segments, referred to as the MA method. Unlike the common approach that uses a uniform threshold for predicting sound event classes, DT employs class-specific thresholds.

For each scenario, we evaluate the following strategies: 1) Fine-tuning a model initialized from the SEC checkpoints pre-trained on AudioSet, denoted as the *From-scratch* method; 2) Fine-tuning a model initialized from the SELD checkpoints pre-trained on *synthetic-training-set*, referred to as the *Fine-tune* method. The difference between the *From-scratch* and *Fine-tune* methods is illustrated in Table I. Notably, the performance difference between these two methods highlights the improvements achieved through the pre-trained SELD models, as shown in the following results.

1) *L3DAS22 Task 2*: L3DAS22 Task 2 [16] focuses on investigating 3D sound event localization and detection using two groups of FOA microphones in a large office room. The

TABLE V: Results on DCASE 2021 Task 3.

Method	Aug.	ER _{20°} ↓	F _{20°} ↑	LE _{CD} ↓	LR _{CD} ↑	ℰ _{SELD} ↓
(#1) Shimada et al. [47], [48] **	✓	0.320	79.1%	8.5°	82.8%	0.187
(#2) Nguyen et al. [47], [79] **	✓	0.320	78.3%	10.0°	78.3%	0.202
From-scratch	✗	0.484	60.5%	15.8°	71.1%	0.314
	✓	0.435	61.7%	17.1°	80.3%	0.278
Fine-tune	✗	0.394	69.3%	12.9°	76.7%	0.251
	✓	0.329	74.7%	11.6°	79.6%	0.213
Fine-tune *	✓	0.285	79.0%	10.0°	82.0%	0.183

* denotes post-processing methods. ** denotes ensemble models.

dataset, synthesized using measured SRIRs, contains 900 30-second audio recordings with 14 classes of sound events. Table IV shows the performance comparison between PSELDNets and our previously proposed system [4], CNN-Conformer, on the evaluation set. The latter system obtained the top rank in L3DAS22 Task 2. We modify the output format of CNN-Conformer to predict the corresponding DOAs rather than 3D Cartesian coordinates of corresponding sound events. For a fair comparison, we utilize only the centre FOA microphones and disregard the secondary FOA microphones. Experimental results indicate the From-scratch method performs comparably to CNN-Conformer [4], while the Fine-tune method surpasses CNN-Conformer by a large margin. Nonetheless, the Fine-tune method exhibits no performance improvement with data augmentation, possibly due to the high degree of similarity between the simulated environments in *synthetic-training-set* and recording environments in the target dataset. Additionally, the post-processing method further improves performance. The Fine-tune method with post-processing achieves superior performance across all metrics.

2) *DCASE 2021 Task 3*: The dataset in DCASE 2021 Task 3 [14], synthesized using measured SRIRs, comprises 800 1-minute audio recordings. Different from the dataset in L3DAS22 Task 2, the DCASE 2021 Task 3 dataset encompasses moving sources and directional interference outside of the target classes. Table V presents the performance difference between PSELDNets and the top two systems [47], [48], [79] on the evaluation set. Notably, these top two systems report exclusively the results of the ensemble models with post-processing. Experimental results reveal that both fine-tuned PSELDNets and the data augmentation method significantly improve performance. Moreover, when comparing the aggregated SELD metric ℰ_{SELD}, our single model fine-tuned PSELDNets with post-processing perform even better than the SOTA system proposed by Shimada et al. [48], which was achieved by ensemble models with post-processing.

3) *STARSS23*: The STARSS23 [20] dataset, an extended version of the STARSS22 [19] dataset, was collected in real-world environments, annotated manually, and served as the dataset for DCASE 2023 Task 3 and DCASE 2024 Task 3. Various clips in STARSS23 were recorded with combinations of moving and static participants acting in simple scenes and interacting among themselves and with the sound props. STARSS23 comprises roughly 7.5 hours of recordings in its development set. Due to the limited size of STARSS23, DCASE provides an additional official synthetic dataset [81]

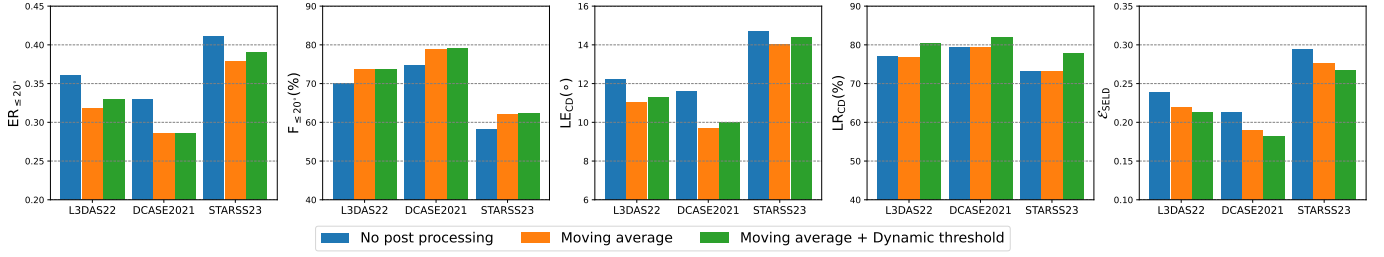


Fig. 6: The effect of the post-processing method on PSELDNets.

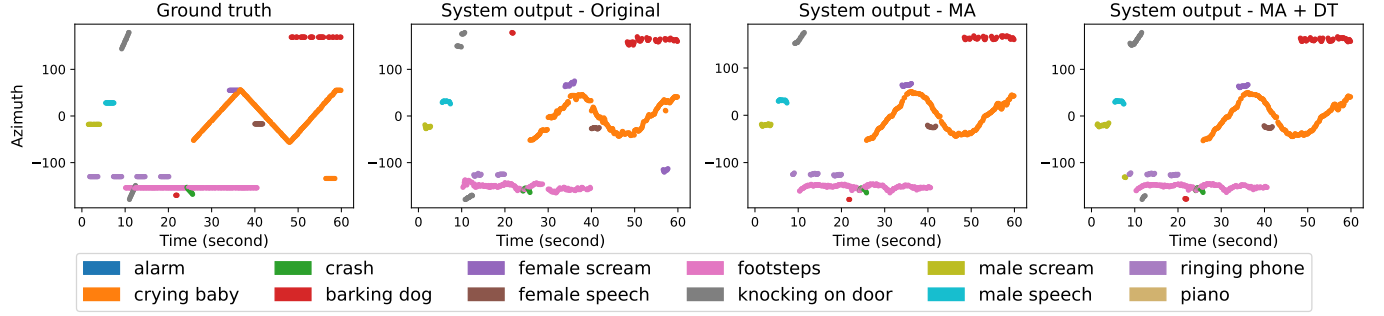


Fig. 7: Visualization of the ground truth and the system output w/ or w/o post-processing for a clip from the DCASE 2021 Task 3 evaluation set. SED predictions with the corresponding azimuth estimations are presented.

TABLE VI: Results on the STARSS23 dataset.

Method	Ext. Data	Aug.	ER _{20°} ↓	F _{20°} ↑	LE _{CD} ↓	LR _{CD} ↑	E _{SEL0°} ↓
(#1) Wang et al. [11], [49] *	Wang-set	✓	0.400	64.0%	13.4°	74.0%	0.274
(#2) Xue et al. [49], [80] *	Xue-set	✓	0.430	54.8%	14.7°	68.0%	0.321
From-scratch	-	✗	0.749	24.1%	30.8°	51.0%	0.542
	-	✗	0.530	42.9%	18.2°	58.6%	0.404
Fine-tune	-	✗	0.640	41.4%	20.7°	62.7%	0.429
	-	✓	0.523	54.4%	15.4°	65.6%	0.352
	Base-set	✓	0.429	56.5%	14.5°	69.8%	0.312
	+ Synth-set	✓	0.411	58.2%	14.7°	73.2%	0.295
Fine-tune *	Base-set	✓	0.390	62.4%	14.4°	77.7%	0.267
	+ Synth-set	✓					

* denotes post-processing methods.

for baseline training of Task 3 of DCASE 2022-2023. The official synthetic dataset, denoted as *Base-set*, is simulated by convolving sound event clips with measured SRIRs from TAU-SRIR DB. Since the labels of the STARSS23 evaluation set are not publicly available, Table VI shows results of the top two systems [11], [49], [80] and PSELDNets on the STARSS23 validation set. These top two systems report the results for the single models with post-processing and also synthesize substantial scenario-specific datasets for training, referred to as *Wang-set* and *Xue-set*. For a fair comparison, we also synthesize scenario-specific datasets, termed as *Synth-set*. The SOTA system proposed by Wang et al. [11], incorporated a class-dependent sound separation approach [82] into the SELD system. The Fine-tune method with post-processing outperforms the system of Wang et al. [11] in terms of $E_{SEL0°}$, more specifically, in terms of LR_{CD} and $ER_{20°}$.

4) *Discussions*: We investigate the performance of PSELDNets applied to these scenario-specific downstream datasets to evaluate the effects of data augmentation and post-processing. Additionally, we discuss the limitations of PSELDNets.

Impact of data augmentation. Empirical evidence indicates that learning-based SELD methods highly rely on large amounts of data, and data augmentation can increase the diversity of samples [4], [10]. Despite utilizing pre-trained checkpoints of PSELDNets, data augmentation continues to improve performance significantly, as shown in Tables V and VI. When applied to downstream datasets, PSELDNets offer general prior knowledge, while data augmentation provides a technique to effectively exploit scenario-specific data.

Impact of post-processing. We surprisingly observe that the post-processing method can also provide significant performance improvement. Fig. 6 presents the effect of the post-processing method in the above three datasets. We see that MA improves performance notably in $ER_{20°}$, $F_{20°}$ and LE_{CD} , while DT offers substantial performance improvement in LR_{CD} . Fig. 7 illustrates a visualization of the ground truth and the system output for a clip from the evaluation set of DCASE 2021 Task 3. We present SED predictions along with the corresponding azimuth estimations. Overall, MA smooths the predicted trajectories of sound events, bringing them closer to the ground truth, compared to the model output without post-processing, such as the moving trajectories of *crying baby*. On the other hand, DT reduces the false negatives, such as the *knocking on door* event observed between 10 to 15 s.

Limitation of PSELDNets. We observe the Fine-tune method exhibits an increase of more than 1° in LE_{CD} compared to the SOTA system of Shimada et al. [48] on DCASE 2021 Task 3 and the SOTA system of Wang et al. [11] on STARSS23, which can be attributed to the output temporal resolution of HTS-AT. Specifically, the output temporal resolution is 0.1 seconds in those systems [4], [11], [48], but approximately 0.3 seconds in HTS-AT due to the effect of

TABLE VII: Results of data-efficient fine-tuning on L3DAS22 Task 2.

Dataset	# Channels	Method	ER _{20°} ↓	F _{20°} ↑	LE _{CD} ↓	LR _{CD} ↑	ℰ _{SELD} ↓
L3DAS22 - split_ov1	4	From-scratch	0.697	30.1%	23.2°	43.1%	0.523
	4	Fine-tune	0.409	64.3%	12.7°	65.6%	0.295
	4	AdapterBit	0.392	64.7%	13.3°	67.6%	0.286
	1	Fine-tune	0.617	32.7%	25.1°	60.4%	0.456
	1	LoRA	0.605	37.0%	24.7°	60.5%	0.442
	1	Adapter	0.620	34.4%	25.4°	62.0%	0.449
	1	AdapterBit	0.591	37.9%	24.5°	62.0%	0.432
Synthetic dataset	4	Fine-tune	0.603	38.0%	23.6°	56.4%	0.448
	4	AdapterBit	0.607	38.1%	23.5°	57.8%	0.445
	1	Fine-tune	0.634	35.1%	23.1°	49.0%	0.480
	1	AdapterBit	0.618	37.8%	22.1°	51.7%	0.461

split patches and the patch merging module. This discrepancy has minimal impact on the localization of static sources, for instance, in L3DAS22 Task 2, the Fine-tune method also shows performance improvement in LE_{CD}, compared to our previous system [4]. However, it introduces a systematic error when localizing moving sources, such as those in DCASE 2021 Task 3 and STARSS23. On the other hand, PSELDNets provide more substantial performance improvement in detection than in localization. Moreover, the systems mentioned in this work [4], [9]–[11], [48], [60], [79], [80] have shown more significant improvement in detection than in localization.

C. Results of data-efficient fine-tuning

In this section, we explore another application of PSELDNets: data-efficient fine-tuning. Specifically, we employ PSELDNets with AdapterBit in both low-resource-data and rich-source-data scenarios. Low-resource data refers to small synthetic datasets using only simulated SRIRs and no data augmentation technique, while rich-resource data includes substantial samples that have been augmented and derived from either real-world scenes or synthesis using collected SRIRs. All results are evaluated on the evaluation sets of DCASE 2021 Task 3 and L3DAS22 Task 2 and the validation set of STARSS23.

Similar to Sec. VI-B, we also compare the *From-scratch* method, the *Fine-tune* method, and additional the *AdapterBit* method. The Fine-tune method fine-tunes all parameters of PSELDNets, while the AdapterBit method only fine-tunes the inserted Adapter module and the bias item of PSELDNets.

1) *Effect of AdapterBit*: Our ablation studies employ *split_ov1* subset of the L3DAS22 Task 2 dataset for training. This subset contains 250 30-second recordings without overlapping sound events. For training on monophonic clips, we extract the first-channel signal from FOA signals and then generate pseudo-FOA signals using these extracted monophonic signals. We evaluate the performance of four methods on pseudo-FOA signals, including Fine-tune, AdapterBit, Adapter (AdapterBit without bias-tuning), and LoRA [34], as shown in the 1-channel part of the top block of Table VII. Experimental results indicate the effectiveness of the designed Adapter, which achieves an ℰ_{SELD} of 0.449, compared to 0.456 for the Fine-tune method. Incorporating additional bias items into

TABLE VIII: Computational efficiency of various fine-tuning methods.

Method	# Tunable Params	Inference FLOPs
Fine-tune	28.1 M	2.88 G
Adapter	4.9 M	3.34 G
AdapterBit	5.0 M	3.34 G
LoRA	5.0 M	2.88 G

Adapter tuning leads to further performance improvement. Additionally, AdapterBit exhibits superior performance relative to LoRA.

Moreover, we compare the performance of pseudo-FOA signals with the corresponding FOA signals, denoted as channels of 4 in the top block of Table VII. The primary difference between pseudo-FOA signals and FOA signals lies in inter-channel correlations. Our observations reveal that leveraging PSELDNets, all methods using only monophonic signals significantly outperform the From-scratch method using corresponding FOA signals, which only achieves ℰ_{SELD} of 0.523. When comparing the Fine-tune and AdapterBit methods using monophonic signals with the corresponding FOA signals, the primary performance difference lies in localization, due to similar performance in LR_{CD} but a significant performance difference in LE_{CD}, such as LE_{CD} of 12.7° and LR_{CD} of 65.6% in the 4-ch Fine-tune method and LE_{CD} of 25.1° and LR_{CD} of 60.4% in the 1-ch Fine-tune method. Original FOA signals of the target scenario contain more information about the acoustic environment and microphone array than pseudo-FOA signals.

Tables VII and VIII present a comparison of the performance and computational efficiency of various fine-tuning methods, respectively. AdapterBit exhibits superior performance compared to LoRA and Adapter despite using the same scale of tunable parameters. However, it requires more inference FLOPs than Fine-tune and LoRA. Notably, LoRA achieves inference FLOPs comparable to Fine-tune, as its additional parameters can be merged into the original model weights. Techniques such as knowledge distillation [83], model quantization [84], and model pruning [85], [86] offer the potential to further reduce the parameter scale and computational overhead, but their investigation is beyond the scope of this study.

TABLE IX: Results of data-efficient fine-tuning on DCASE 2021 Task 3.

# Channels	Method	ER _{20°} ↓	F _{20°} ↑	LE _{CD} ↓	LR _{CD} ↑	ε _{SELD} ↓
4	Fine-tune	0.621	40.4%	23.0°	56.9%	0.444
4	AdapterBit	0.610	43.4%	22.5°	61.5%	0.422
1	Fine-tune	0.594	42.3%	22.5°	58.7%	0.427
1	AdapterBit	0.595	44.9%	21.5°	59.2%	0.418

TABLE X: Results of data-efficient fine-tuning on the STARSS23 dataset.

# Channels	Method	ER _{20°} ↓	F _{20°} ↑	LE _{CD} ↓	LR _{CD} ↑	ε _{SELD} ↓
4	Fine-tune	0.732	23.6%	22.4°	41.4%	0.552
4	AdapterBit	0.736	23.5%	24.7°	44.8%	0.548
1	Fine-tune	0.719	19.2%	26.5°	37.3%	0.575
1	AdapterBit	0.745	22.9%	25.9°	43.3%	0.557

2) *Results on low-resource data:* Given the absence of polyphonic indications in each clip from STARSS23 and DCASE 2021 Task 3, we create scenario-specific datasets for training, as well as for L3DAS22 Task 2. Each dataset, consisting of 120 one-minute FOA audio clips, is generated using simulated SRIRs. The bottom block of Table VII, Table IX, and Table X illustrate the results of data-efficient fine-tuning. Tests with either multi-channel or monophonic signals demonstrate the efficacy of AdapterBit relative to the Fine-tune method. Remarkably, AdapterBit tuning on monophonic signals reaches performance levels comparable to those from synthesized FOA signals, thus potentially simplifying the adaptation process to a target scene and diminishing the requirement for synthesizing multi-channel clips. This suggests that the performance difference between original FOA signals and pseudo-FOA signals, as shown in the top block of Table VII, may indicate the discrepancy between the simulated environments and the acoustic conditions in the real world. Additionally, Tables VII, IX and X show the capacity of PSELDNets to generalize across diverse microphone array setups through the FOA format without incorporating scenario-specific information, despite differences in microphone array setups between *synthetic-training-set* and scenario-specific datasets.

3) *Results on rich-resource data:* Fig. 8 illustrates the performance of PSELDNets using AdapterBit, trained on varying proportions of the training set from downstream datasets, including clips from different numbers of rooms or splits. These datasets are either synthesized using measured SRIRs from real-world environments or directly collected in real-world environments. When employing STARSS23, a small-size synthesized dataset comprising spatial sound events of all corresponding classes is additionally utilized, because clips from any individual rooms in STARSS23 only encompass partial sound event classes. Experimental results indicate that both the Fine-tune method and AdapterBit exhibit superior performance compared to the From-scratch method, irrespective of the dataset size. Notably, AdapterBit also demonstrates more efficient data utilization than the Fine-tune method when trained on limited proportions of the training set, such as the clips from the first two splits of L3DAS22 Task 2, the

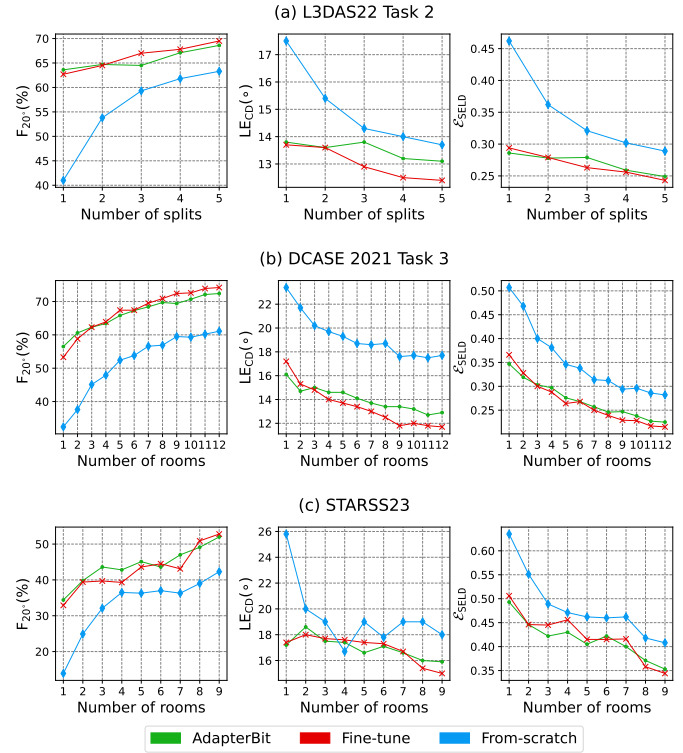


Fig. 8: The effect of AdapterBit on various proportions of the training set from downstream datasets.

first two rooms of DCASE 2021 Task 3 and the first five rooms of STARSS23. This aligns with the previous results on low-resource data. However, when trained on more data, AdapterBit's performance falls behind the Fine-tune method. This performance disparity can be attributed to the substantial distributional differences between the simulated environments and real-world acoustic environments, which AdapterBit may not effectively learn due to its limited trainable parameters.

D. Results on Indoor Recordings

In this section, we evaluate the transferability of PSELDNets and the impact of data-efficient fine-tuning on Indoor Recordings. As shown in Fig. 9, we exploit a 4-channel unbaffled spherical microphone array with a radius of 0.12 m, arranged in a tetrahedral configuration, to record sound sources emitted by loudspeakers in two distinct environments: an anechoic chamber and a meeting room. In the meeting room, we estimate a reverberation time of $T_{60} = 900$ ms and a signal-noise-ratio (SNR) of 6 dB. The microphone array was centrally positioned in a square, with loudspeakers placed at three vertical heights and eight predetermined horizontal locations corresponding to either the square's vertices or midpoints of the square's sides. The sides of the square have lengths of 4 m and 2.4 m. Therefore, there are a total of 48 sound source locations. We placed a loudspeaker among these locations and recorded one-minute audio clips with a maximum polyphony of 1 at each location, resulting in 48 non-overlapping audio clips. Additionally, we recorded 12 one-minute audio clips with a maximum polyphony of 2, where two loudspeakers

TABLE XI: Results on Indoor Recordings.

Method	# Channels	Datasets	Anechoic Chamber					Meeting Room				
			ER _{20°} ↓	F _{20°} ↑	LE _{CD} ↓	LR _{CD} ↑	ε _{SELD} ↓	ER _{20°} ↓	F _{20°} ↑	LE _{CD} ↓	LR _{CD} ↑	ε _{SELD} ↓
From-scratch	4	Sim240 + Sim120	0.392	70.6%	12.9°	73.9%	0.255	0.448	63.0%	15.5°	71.2%	0.298
Fine-tune	4	Sim240 + Sim120	0.393	69.3%	12.1°	76.7%	0.250	0.430	64.4%	15.3°	76.4%	0.277
Fine-tune	4	Sim240 + Co120	0.358	72.8%	11.9°	79.9%	0.224	0.414	67.6%	14.0°	79.7%	0.255
Fine-tune	4	Sim120_ov1	0.526	57.9%	13.9°	65.2%	0.343	0.686	36.7%	24.7°	67.3%	0.446
AdapterBit	4	Sim120_ov1	0.480	60.7%	14.1°	63.2%	0.330	0.615	41.1%	23.5°	65.3%	0.420
Fine-tune	1	Sim120_ov1	0.626	43.5%	22.0°	53.3%	0.445	0.809	16.9%	39.7°	50.7%	0.588
AdapterBit	1	Sim120_ov1	0.570	48.4%	22.2°	64.3%	0.392	0.768	22.2%	35.1°	59.5%	0.537

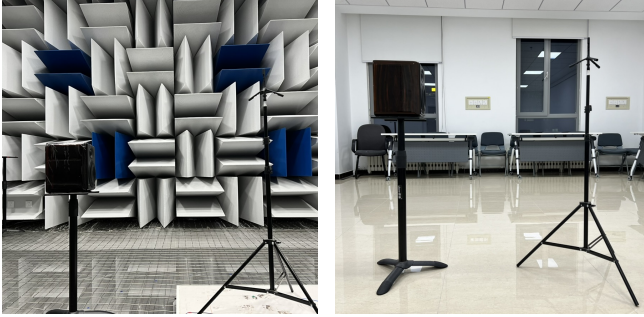


Fig. 9: Recording environments used in real-world experiments, comprising an anechoic chamber and a meeting room.

were randomly placed among these 48 locations for each clip. We also measured the RIRs at 8 supplementary positions in the meeting room to facilitate efficient fine-tuning. The sound event clips from NIGENS [54] are divided into training and test splits, with the training subset for data synthesis and the test subset for playback through the loudspeakers. In total, we collected 60 one-minute audio clips for each recording environment as the evaluation set.

We synthesize four datasets for training: Sim240, Sim120, Sim120_ov1 and Co120. The maximum polyphony of Sim240, Sim120 and Co120 is 2, whereas the maximum polyphony of Sim120_ov1 is 1. Sim240, Sim120 and Sim120_ov1 are generated using simulated SRIRs and contain 240, 120, and 120 one-minute audio clips, respectively, while Co120 is synthesized using previously collected 8 RIRs in the meeting room and also contains 120 one-minute audio clips. The shape of the microphone array in synthetic datasets is consistent with the array used for recording in the real scene.

Same as Sec. VI-C, we compare the *Fine-tune*, *From-scratch*, and *AdapterBit* methods. The top block of Table XI presents the transferability of PSELDNets on Indoor Recordings. We fine-tune PSELDNets on augmented synthetic datasets and evaluate the performance on Indoor Recordings collected from two environments. Notably, the distribution differences in sound event clips and recording locations between these two environments are minimal, excluding acoustic properties, such as noise level and reverberation. We observe that the performance in LR_{CD} between two environments are similar, but the performance difference is mainly in localization, which can be due to differences in acoustic environments. Experimental results demonstrate the effectiveness of PSELD-

Nets. Additionally, replacing Sim120 with Co120 can improve performance in both localization and detection, since collected RIRs carry more information about the acoustic environment and microphone characteristics. On the other hand, we note that LE_{CD} in the ideal acoustic environment, anechoic chamber, reaches approximately 12°, perhaps due to the large radius and unbaffled configuration of the spherical microphone array [63], which leads to incorrect FOA conversion, especially in the high-frequency range. Furthermore, while we transfer pre-trained models to specific scenarios, the considerable variation in distribution between the source domain and the target domain [87], such as the microphone characteristics, puts the prior knowledge at high risk of losing effectiveness [87]–[90].

The bottom block of Table XI illustrates the efficacy of the data-efficient fine-tuning method on Indoor Recordings. The results indicate that the model fine-tuned on FOA signals performs better than on pseudo-FOA signals by a large margin, which can be attributed to substantial differences in microphone array configurations and acoustic environments between *synthetic-training-set* and the synthetic and recorded datasets used in this section. Additionally, when using multi-channel or monophonic signals, AdapterBit tuning performs better than the Fine-tune method. Notably, it suggests that AdapterBit is particularly effective when only using monophonic signals. We hypothesize that AdapterBit prevents catastrophic interference [33], thereby decreasing the likelihood that the model forgets previously learned knowledge when adapting to new scenarios.

VII. CONCLUSION

This paper has built sound event localization and detection (SELD) foundation models by introducing pre-trained SELD networks (PSELDNets) on a large-scale synthetic dataset. The synthetic dataset encompasses 1,167 hours of audio recordings with an ontology of 170 sound classes. To enhance the adaptability of PSELDNets to specific scenarios with low-resource data, we have presented AdapterBit, a data-efficient fine-tuning technique. We evaluate PSELDNets on *synthetic-test-set* and achieve satisfactory performance. We transfer PSELDNets to several downstream datasets that are publicly available, as well as to our own collected recordings, Indoor Recordings. The experimental results demonstrate superior performance compared to previous state-of-the-art systems. Moreover, incorporating AdapterBit into PSELDNets enhances the efficiency of the transferability for low-resource data, including both limited multi-channel and monophonic audio clips. Future work will focus on model compression and computational efficiency to enable real-time applications.

REFERENCES

- [1] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2019.
- [2] Y. Cao, T. Iqbal, Q. Kong, Y. Zhong, W. Wang, and M. D. Plumbley, "Event-independent network for polyphonic sound event localization and detection," in *Proc. Detect. Classification Acoust. Scenes Events Workshop*, 2020, pp. 11–15.
- [3] Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang, and M. D. Plumbley, "An improved event-independent network for polyphonic sound event localization and detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 885–889.
- [4] J. Hu, Y. Cao, M. Wu, Q. Kong, F. Yang, M. D. Plumbley, and J. Yang, "A track-wise ensemble event independent network for polyphonic sound event localization and detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 9196–9200.
- [5] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 915–919.
- [6] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji, "Multi-ACCDOA: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 316–320.
- [7] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," in *Proc. Detect. Classification Acoust. Scenes Events Workshop*, 2019, pp. 30–34.
- [8] T. T. N. Nguyen, K. N. Watcharasupat, K. N. Nguyen, D. L. Jones, and W.-S. Gan, "SALSA: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1749–1762, 2022.
- [9] J. Hu, Y. Cao, M. Wu, Q. Kong, F. Yang, M. D. Plumbley, and J. Yang, "Sound event localization and detection for real spatial sound scenes: Event-independent network and data augmentation chains," in *Proc. Detect. Classification Acoust. Scenes Events Workshop*, 2022, pp. 46–50.
- [10] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A four-stage data augmentation approach to ResNet-Conformer based acoustic modeling for sound event localization and detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1251–1264, 2023.
- [11] Q. Wang, Y. Jiang, S. Cheng, M. Hu, Z. Nian, P. Hu, Z. Liu, Y. Dong, M. Cai, J. Du, and C.-H. Lee, "The NERC-SLIP system for sound event localization and detection of DCASE 2023 challenge," *DCASE2023 Challenge*, Tech. Rep., 2023.
- [12] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Proc. Detect. Classification Acoust. Scenes Events Workshop*, 2019, pp. 10–14.
- [13] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Proc. Detect. Classification Acoust. Scenes Events Workshop*, 2020, pp. 165–169.
- [14] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," in *Proc. Detect. Classification Acoust. Scenes Events Workshop*, 2021, pp. 125–129.
- [15] E. Guizzo, R. F. Gramaccioni, S. Jamili, C. Marinoni, E. Massaro, C. Medaglia, G. Nachira, L. Nucciarelli, L. Paglialunga, M. Pennese, S. Pepe, E. Rocchi, A. Uncini, and D. Comminiello, "L3DAS21 challenge: Machine learning for 3d audio signal processing," in *Proc. IEEE Int. Workshop Mach. Learning Signal Process.*, 2021, pp. 1–6.
- [16] E. Guizzo, C. Marinoni, M. Pennese, X. Ren, X. Zheng, C. Zhang, B. Masiero, A. Uncini, and D. Comminiello, "L3DAS22 Challenge: Learning 3D audio sources in a real office environment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 9186–9190.
- [17] C. Marinoni, R. F. Gramaccioni, C. Chen, A. Uncini, and D. Comminiello, "Overview of the l3das23 challenge on audio-visual extended reality," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–2.
- [18] J. Hu, Y. Cao, M. Wu, Q. Kong, F. Yang, M. D. Plumbley, and J. Yang, "Selective-memory meta-learning with environment representations for sound event localization and detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 4313–4327, 2024.
- [19] A. Politis, K. Shimada, P. Sudarsanam, S. Adavanne, D. Krause, Y. Koyama, N. Takahashi, S. Takahashi, Y. Mitsufuji, and T. Virtanen, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Proc. Detect. Classification Acoust. Scenes Events Workshop*, 2022, pp. 161–165.
- [20] K. Shimada, A. Politis, P. Sudarsanam, D. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, T. Virtanen, and Y. Mitsufuji, "STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 72 931–72 957.
- [21] K. Shimada, K. Uchida, Y. Koyama, T. Shibuya, S. Takahashi, Y. Mitsufuji, and T. Kawahara, "Zero- and few-shot sound event localization and detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 636–640.
- [22] J. Hu, Y. Cao, M. Wu, Z. Yu, F. Yang, W. Wang, M. D. Plumbley, and J. Yang, "Meta-SELD: Meta-learning for fast adaptation to the new environment in sound event localization and detection," in *Proc. Detect. Classification Acoust. Scenes Events Workshop*, 2023, pp. 51–55.
- [23] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2880–2894, 2020.
- [24] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proc. Interspeech*, 2021, pp. 571–575.
- [25] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Proc. Interspeech*, 2022, pp. 2753–2757.
- [26] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 646–650.
- [27] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 776–780.
- [28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.
- [30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10 012–10 022.
- [31] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613 – 2617.
- [32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
- [33] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "AdaptFormer: Adapting vision transformers for scalable visual recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 16 664–16 678.
- [34] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations*, 2022.
- [35] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig, "Towards a unified view of parameter-efficient transfer learning," in *Proc. Int. Conf. Learn. Representations*, 2022.
- [36] E. Ben Zaken, Y. Goldberg, and S. Ravfogel, "BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2022, pp. 1–9.
- [37] N. Houlsby, A. Giurghi, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *Proc. Int. Conf. Mach. Learn.*, vol. 97, 2019, pp. 2790–2799.
- [38] T. Yang, Y. Zhu, Y. Xie, A. Zhang, C. Chen, and M. Li, "AIM: Adapting image models for efficient video action recognition," in *Proc. Int. Conf. Learn. Representations*, 2023.
- [39] D. Yin, Y. Yang, Z. Wang, H. Yu, K. Wei, and X. Sun, "1% VS 100%: Parameter-efficient low rank adapter for dense predictions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 20 116–20 126.

- [40] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 709–727.
- [41] Y. Liang, H. Lin, S. Qiu, and Y. Zhang, "AAT: Adapting audio transformer for various acoustics recognition tasks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 1361–1365.
- [42] T. Rolland and A. Abad, "Exploring adapters with conformers for children's automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 12 747–12 751.
- [43] M. Sang and J. H. Hansen, "Efficient adapter tuning of pre-trained speech models for automatic speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 12 131–12 135.
- [44] K. C. Sim, Z. Huo, T. Munkhdalai, N. Siddhartha, A. Stooke, Z. Meng, B. Li, and T. Sainath, "A comparison of parameter-efficient ASR domain adaptation methods for universal speech and language models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 6900–6904.
- [45] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 829–852, 2022.
- [46] A. Politis, S. Adavanne, and T. Virtanen, "TAU Spatial Room Impulse Response Database (TAU-SRIR DB)," 2022. [Online]. Available: <https://doi.org/10.5281/zenodo.6408611>
- [47] A. Politis, A. Deleforge, S. Adavanne, P. Srivastava, D. Krause, and T. Virtanen, "[DCASE2021 task 3] sound event localization and detection with directional interference," 2021. [Online]. Available: <https://dcase.community/challenge2021>
- [48] K. Shimada, N. Takahashi, Y. Koyama, S. Takahashi, E. Tsunoo, M. Takahashi, and Y. Mitsufuji, "Ensemble of ACCDOA- and EINV2-based systems with D3Nets and impulse response simulation for sound event localization and detection," DCASE2021 Challenge, Tech. Rep., 2021.
- [49] A. Politis, K. Shimada, Y. Mitsufuji, T. Virtanen, S. Adavanne, P. Sudarsanam, D. Krause, N. Takahashi, S. Takahashi, Y. Koyama, K. Uchida, and A. Hakala, "[DCASE2023 task 3] sound event localization and detection evaluated in real spatial sound scenes," 2023. [Online]. Available: <https://dcase.community/challenge2023>
- [50] E. Fonseca, M. Plakal, D. P. W. Ellis, F. Font, X. Favory, and X. Serra, "Learning sound event classifiers from web audio with noisy labels," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 21–25.
- [51] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, and X. Serra, "Audio tagging with noisy labels and minimal supervision," in *Proc. Detect. Classification Acoust. Scenes Events Workshop*, 2019, pp. 69–73.
- [52] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. ACM Int. Conf. Multimedia*, 2015, p. 1015–1018.
- [53] M. Cartwright, A. Cramer, A. E. M. Mendez, Y. Wang, H.-H. Wu, V. Lostanlen, M. Fuentes, G. Dove, C. Mydlarz, J. Salamon *et al.*, "SONYC-UST-V2: An urban sound tagging dataset with spatiotemporal context," in *Proc. Detect. Classification Acoust. Scenes Events Workshop*, 2020, pp. 16–20.
- [54] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, "NIGENS general sound events database," 2019. [Online]. Available: <https://doi.org/10.5281/zenodo.2535878>
- [55] P. Foster, S. Sigita, S. Krstulovic, J. Barker, and M. D. Plumbley, "Chime-home: A dataset for sound source recognition in a domestic environment," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2015, pp. 1–5.
- [56] S. Hershey, D. P. W. Ellis, E. Fonseca, A. Jansen, C. Liu, R. Channing Moore, and M. Plakal, "The benefit of temporally-strong labels in audio event classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 366–370.
- [57] B. Frenay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, 2014.
- [58] A. Shah, A. Kumar, A. G. Hauptmann, and B. Raj, "A closer look at weak label learning for audio events," *arXiv preprint arXiv:1804.09288*, 2018.
- [59] K. Nagatomo, M. Yasuda, K. Yatabe, S. Saito, and Y. Oikawa, "Wearable SELD dataset: Dataset for sound event localization and detection using wearable devices around head," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 156–160.
- [60] Q. Wang, Y. Dong, H. Hong, R. Wei, M. Hu, S. Cheng, Y. Jiang, M. Cai, X. Fang, and J. Du, "The NERC-SLIP system for sound event localization and detection with source distance estimation of DCASE 2024 challenge," DCASE2023 Challenge, Tech. Rep., 2023.
- [61] L. McCormack, A. Politis, R. Gonzalez, T. Lokki, and V. Pulkki, "Parametric ambisonic encoding of arbitrary microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2062–2075, 2022.
- [62] M. Heikkinen, A. Politis, and T. Virtanen, "Neural ambisonics encoding for compact irregular microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 701–705.
- [63] B. Rafaely, *Fundamentals of Spherical Array Processing*. Springer, 2015.
- [64] A. Politis, "Microphone array processing for parametric spatial audio techniques," Ph.D. dissertation, Aalto University, 2016.
- [65] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.
- [66] A. Politis and H. Gamper, "Comparing modeled and measurement-based spherical harmonic encoding filters for spherical microphone arrays," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 224–228.
- [67] Y. Koyama, K. Shigemitsu, M. Takahashi, K. Shimada, N. Takahashi, E. Tsunoo, S. Takahashi, and Y. Mitsufuji, "Spatial data augmentation with simulated room impulse responses for sound event localization and detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 8872–8876.
- [68] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020, pp. 5036 – 5040.
- [69] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.
- [70] W. Gao, F. Wan, X. Pan, Z. Peng, Q. Tian, Z. Han, B. Zhou, and Q. Ye, "TS-CAM: Token semantic coupled attention map for weakly supervised object localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2886–2895.
- [71] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Representations*, 2018.
- [72] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, 2020, pp. 13 001–13 008.
- [73] L. Mazzon, Y. Koizumi, M. Yasuda, and N. Harada, "First order ambisonics domain spatial augmentation for DNN-based direction of arrival estimation," in *Proc. Detect. Classification Acoust. Scenes Events Workshop*, 2019, pp. 154–158.
- [74] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 351–355.
- [75] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2018.
- [76] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 333–337.
- [77] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in DCASE 2019," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 684–698, 2021.
- [78] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Appl. Sci.*, vol. 6, no. 6, p. 162, 2016.
- [79] T. N. T. Nguyen, K. Watcharasupat, N. K. Nguyen, D. L. Jones, and W. S. Gan, "DCASE 2021 task 3: Spectrotemporally-aligned features for polyphonic sound event localization and detection," DCASE2021 Challenge, Tech. Rep., 2021.
- [80] L. Xue, H. Liu, and Y. Zhou, "Attention mechanism network and data augmentation for sound event localization and detection," DCASE2023 Challenge, Tech. Rep., 2023.
- [81] A. Politis, "[DCASE2022 Task 3] Synthetic SELD mixtures for baseline training," 2022. [Online]. Available: <https://doi.org/10.5281/zenodo.6406873>
- [82] S. Cheng, J. Du, Q. Wang, Y. Jiang, Z. Nian, S. Niu, C.-H. Lee, Y. Gao, and W. Zhang, "Improving sound event localization and detection with class-dependent sound separation for real-world scenarios," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2023, pp. 2068–2073.
- [83] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, 2021.
- [84] A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," in *Proc. Int. Conf. Learn. Representations*, 2018.

- [85] K. Tyagi, S. Nguyen, R. Rawat, and M. Manry, "Second order training and sizing for the multilayer perceptron," *Neural Process. Lett.*, vol. 51, pp. 963–991, 2020.
- [86] Y. Zhang, M. Lin, C.-W. Lin, J. Chen, Y. Wu, Y. Tian, and R. Ji, "Carrying out CNN channel pruning in a white box," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 7946–7955, 2022.
- [87] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, 2020.
- [88] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *J. Acoust. Soc. Amer.*, vol. 152, no. 1, pp. 107–151, 2022.
- [89] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.
- [90] W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 766–785, 2019.