IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 October 2024, accepted 19 November 2024, date of publication 2 December 2024, date of current version 11 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3510453

RESEARCH ARTICLE

w2v-SELD: A Sound Event Localization and Detection Framework for Self-Supervised Spatial Audio Pre-Training

ORLEM LIMA DOS SANTOS[1], (Graduate Student Member, IEEE), KAREN ROSERO[2], (Graduate Student Member, IEEE), BRUNO MASIERO[1], (Member, IEEE), AND ROBERTO DE ALENCAR LOTUFO[1], (Member, IEEE)

[1]Department of Computer Engineering and Industrial Automation, University of Campinas, Campinas 13083-970, Brazil
[2]Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson, TX 75080, USA

Corresponding author: Orlem Lima Dos Santos (o211501@dac.unicamp.br)

This work was supported in part by the "Centro Nacional de Processamento de Alto Desempenho em São Paulo (CENAPAD-SP)" and in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–Brasil (CAPES). The work of Bruno Masiero was supported in part by São Paulo Research Foundation (FAPESP), Brasil under Grant 2017/08120-6. The work of Roberto de Alencar Lotufo was supported in part by The Brazilian National Council for Scientific and Technological Development (CNPq) under Grant 313047/2022-7.

ABSTRACT Sound Event Localization and Detection (SELD) is a critical challenge in various industrial applications, such as autonomous systems, smart cities, and audio surveillance, which require accurate identification and localization of sound events in complex environments. Traditional supervised approaches heavily rely on large, annotated multichannel audio datasets, which are expensive and time-consuming to produce. This paper addresses this limitation by introducing the w2v-SELD architecture, a self-supervised model adapted from the wav2vec 2.0 framework to learn effective sound event representations directly from raw, unlabeled 3D audio data in ambisonics format. The proposed model follows a two-stage process: pre-training on large, unlabeled 3D audio datasets to capture high-level features, followed by fine-tuning with a smaller, labeled SELD dataset. Experimental results show that our w2v-SELD method outperforms baseline models on Detection and Classification of Acoustic Scenes and Events (DCASE) challenges, achieving a 66% improvement for DCASE TAU-2019 and a 57% improvement on DCASE TAU-2020 with respect to baseline systems. The w2v-SELD model performs competitively with state-of-the-art supervised methods, highlighting its potential to significantly reduce the dependency on labeled data in industrial SELD applications. The code and pre-trained parameters of our w2v-SELD model are available online.

INDEX TERMS Sound event localization and detection, self-supervised learning, spatial audio, wav2vec 2.0.

I. INTRODUCTION

The Sound Event Localization and Detection (SELD) task is a cutting-edge application to analyze complex acoustic scenes. Inspired by the human auditory system, which processes sound from all our surroundings, the SELD task relies on spatial audio that preserves or recreates the spatial information of natural sound using multichannel Ambisonic recordings. From a computational perspective, SELD encompasses two tasks: Sound Event Detection (SED) and estimation of Direction Of Arrival (DOA), which is also referred to as sound source localization [1]. The SED task involves identifying every sound event's start and end times and the corresponding class. At the same time, the DOA estimation predicts the respective three-dimensional coordinates from the respective sound source. Therefore, the SELD task combines three essential components of sound analysis: temporal identification, spatial localization, and semantic labeling of sound events throughout time.

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

The development of computational solutions to perform SELD in real-time opens up new avenues for innovation and automation in various fields where the accurate analysis of sounds is of utmost importance, such as smart cities [2], autonomous vehicles [3], wildlife monitoring and conservation [4], healthcare and assistive technologies [5], and industrial monitoring [6]. However, the lack of extensive labeled datasets poses a significant challenge for traditional supervised learning approaches applied to the SELD task.

While SED has generally been treated as a multi-label classification problem, the DOA estimation has evolved from parametric approaches into Deep Neural Network (DNN)-based approaches. This tendency responds to the limitations of parametric-based methods in capturing complex and non-linear relationships in the data. In contrast, DNN-based methods have demonstrated the ability to accurately estimate the DOA, even in highly complex and noisy scenarios. The SELDnet [7] was the first DNN model that gained relevance in the field. It performs the SELD task by training a single Convolutional Recurrent Neural Network (CRNN) with two branches: one for SED and the other for DOA estimation. Further works employing supervised learning for SELD include the use of Temporal Convolutional Networks (TCN) [1], ensemble models [8], tailored data augmentation techniques [9], and multi-channel features fusion [10].

Despite the increasing performance of state-of-the-art models for SELD, the limited amount of multichannel recordings contained in spatial audio datasets hinders the models from reaching better metrics. The lack of extensive datasets for the SELD task can be explained by the need for special equipment to record spatial audio, namely microphone arrays, but also because of the arduous labeling process, in which the timestamps, location, and class of every sound event should be accurately annotated to be used on supervised learning approaches.

In contrast, speech recognition has experienced significant advancement with the introduction of Transformer-based models coupled with a self-supervised pre-training framework, especially the wav2vec 2.0 model [11]. This enables the model to learn general representations from vast amounts of unlabeled audio data during pre-training. The pre-training step is crucial as it allows the model to establish a strong foundation of knowledge before being fine-tuned on transcribed speech (labeled data).

The use of pre-trained models for the SELD task has been explored in recent works [12], [13], [14], [15], [16]. However, we have observed that these approaches leverage representations from models pre-trained exclusively on mono-channel audio datasets. In response to this shortfall, we propose utilizing SSL frameworks to extract valuable information from unlabeled spatial audio datasets. We have adapted the wav2vec 2.0 architecture to accommodate a multichannel input and produce outputs containing SED and DOA predictions. We refer to this model as w2v-SELD. Additionally, we explore the gap between using the model pre-trained on single-channel speech data and unlabeled spatial audio. Subsequently, the pre-trained model is fine-tuned into a specific domain using transfer learning. This approach also eliminates the need to train several models independently for a specific task before ensembling them for improved predictions.

Our results demonstrate that modifying w2v-SELD to perform SED and DOA predictions over an entire frame results in a better performance than using the original sequential approach of the model. We also verify a 20% improvement of the $\text{SELD}_{\text{score}}$ metric when using pre-trained weights instead of training the model from scratch. Furthermore, pre-training w2v-SELD on unlabeled spatial audio improves the $\text{SELD}_{\text{score}}$ metric by 40% compared with pre-training on single-channel audio. Lastly, we compare the performance of our w2v-SELD approach with the baseline systems provided for each dataset, and with the state-of-the-art systems. The $\text{SELD}_{\text{score}}$ improved by 66% on DCASE TAU-2019 and 57% on DCASE TAU-2020 with respect to the baseline system performance and reached close state-of-the-art performance using raw spatial audio as input instead of relying on spectrograms, phase, or intensity vectors.

The main contributions of this work are as follows:

- **Development of w2v-SELD**: Adaptation of the wav2vec 2.0 model to spatial audio for SED and DOA estimation, creating the w2v-SELD model;
- **Self-Supervised Learning (SSL) Framework**: Introduction of an SSL approach for SELD using the wav2vec 2.0 pre-training framework, significantly reducing the need for labeled spatial audio data;
- **Multichannel Adaptation:**: Modifying the wav2vec 2.0 architecture to handle multichannel audio inputs and provide frame-level predictions for both SED and DOA tasks;
- **Efficient Transfer Learning and Fine-tuning**: Demonstrated that transfer learning with pre-trained weights improves SELD performance by 66% on DCASE 2019 and 57% on DCASE 2020 compared to baseline systems, especially in data-constrained environments.
- **Toolkit and Pre-trained Weights**: Released an open-source toolkit built upon fairseq and pre-trained weights for both BASE and LARGE configurations of w2v-SELD, tailored for SELD tasks with limited dataset sizes.[1]

The paper is structured as follows: In Section II, we provide a comprehensive literature review concerning the domains of SELD for spatial audio. In Section III, we expound upon the theoretical foundations of SSL. Section IV details the methodology employed in our research, while Section V describes the experimental setup. Section VI presents the outcomes of our experimental endeavors. Finally, critical

[1]The code is available at https://github.com/Orlllem/seld_wav2vec2.git

points are discussed, and conclusive remarks are addressed in Sections VII and VIII, respectively.

II. RELATED WORKS

In this section, we provide a summary of key contributions and advancements in the field of SELD as reported in the literature, highlighting the major trends and challenges encountered in this area of research.

Adavanne et al. [7] introduced SELDnet, the first Neural Network (NN) that jointly handles both SED and DOA estimation without parametric methods, therefore, it is often considered a baseline for the SELD task. SELDnet employs a CRNN with two branches: one for multi-label classification of SED and another for multi-output regression to estimate 3D Cartesian DOA coordinates, as shown in Figure 1. Inspired by SELDnet, several improvements have been proposed. Guirguis et al. [1] replaced RNN layers with TCNs to enhance performance and reduce computational cost, while Rosero et al. [17] combined TCN and RNNs for further optimization, also introducing a Gammatone filterbank in the preprocessing stage.

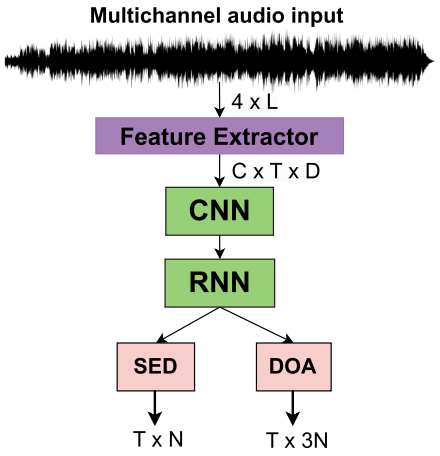


FIGURE 1. Illustration of the SELDnet model. Adapted from [7].

The DCASE challenges have been instrumental in advancing SELD. For DCASE 2019, Kapka et al. [8] used SpecAugment [9] and ensemble SELDnets to achieve state-of-the-art results. In DCASE 2020, Wang et al. [18] utilized ResNet [19] and Xception [20] combined with Gated Recurrent Units (GRUs) or factorized Time Delay Neural Networks (TDNNF), enhancing data augmentation strategies and leveraging spatial audio transformation techniques. In 2021, Shimada et al. [21] introduced ACCDOA, a method that eliminated the traditional SED branch by making the SED task a function of the Cartesian DOA vector's length, which influenced subsequent SELD models by simplifying the architecture. Moreover, a recent approach named FusionNet [10], applies channel-wise convolutional operations on multichannel inputs and fuses these representations into a unified feature space for the SELD task.

Innovative architectures using attention mechanisms, such as ADRENALINE [22], surpassed SELDnet's DOA performance by capturing long-term dependencies with attention-based deep RNNs, while PILOT [23] introduced transformers to handle temporal dependencies in sequence data of spatial audio. Despite the improvements introduced for the DOA estimation, both approaches omitted the SED task. Expanding the use of transformer-based models for both SED and DOA tasks, CST-former [24] employed dedicated attention mechanisms to process channel, spectral, and temporal information separately. Since these approaches rely solely on annotated datasets, they require significant manual effort for the annotation process, which restricts scalability.

Scheibler et al. [12] adapted a Self-Supervised Audio Spectrogram Transformer (SSAST) [12] for the SELD task. They used the SSAST weights that were pre-trained on monophonic utterances from the Audioset dataset [25], which comprises 4,971 hours of audio. Similarly, Xu et al. [13] employed semi-supervised methods to leverage pre-trained models for the SED task, demonstrating the benefit of using embeddings from neural networks pre-trained on large datasets. Additionally, The ResNet-Conformer architecture, which integrates a ResNet [19] backbone with a Conformer module (a special type of Transformer) [26], has also been adapted for the SELD task. Several studies, including [12], [14], [15], [16], have successfully combined the strengths of ResNet's capability in capturing spatial features and the Conformer module's effectiveness in modeling long-range dependencies through self-attention and convolution mechanisms. These approaches showed promise but remain underexplored for spatial audio pre-training.

Despite these advances, most SELD models rely heavily on labeled datasets, limiting scalability. Our proposed w2v-SELD model introduces a novel SSL-based approach, leveraging both labeled and unlabeled data to enhance SELD system robustness and scalability. This pre-training mechanism on unlabeled spatial audio represents a significant step toward reducing the dependence on costly annotations and improving generalization in real-world audio environments.

III. THEORETICAL BACKGROUND

This section explores the key concepts, techniques, and motivations behind self-supervised Transformers for audio and speech representations. Specifically, we describe the key aspects of the wav2vec 2.0 model and its pre-training methodology, which are adapted for the SELD task in this study.

Self-supervised Transformers learn accurate audio and speech representations by training on large amounts of unlabeled audio data. The learned audio representations have been used in downstream tasks such as speaker verification [27], speech recognition [11], or emotion recognition [28]. Specifically, this approach has achieved state-of-the-art performance

in speech recognition, demonstrating remarkable reductions in the word error rates.

A. WAV2VEC 2.0 MODEL

The wav2vec 2.0 model, proposed by Baevski et al. [11], is one of the current state-of-the-art frameworks for unsupervised speech representations. Through training on a massive amount of audio data, the model learns to extract high-level features, such as phonemes and sub-word units, directly from raw waveform data. The model comprises three main components: 1) feature encoder, 2) context network, and 3) quantization module, which will be briefly described.

The feature encoder includes seven blocks, each containing temporal convolutions used for dimensionality reduction and feature representation. A CNN layer is responsible for modeling relative positional embeddings of the input sequence. As shown in Figure 2, the feature encoder, denoted as $f : X \rightarrow Z$, takes raw audio X as input and creates latent speech representations (embeddings) $z_1, \ldots, z_T$ for T time-steps. In the self-supervised pre-training phase, the output audio embeddings $z_1, \ldots, z_T$ from the feature encoder are discretized into a finite set of speech representations via a quantization module $Z \rightarrow Q$, resulting in discrete representations $\boldsymbol{q}_t$ for each time-step. The quantization step is removed when fine-tuning.

These speech embeddings are then fed into the Transformers block, denoted as $g : Z \rightarrow C$, which produces contextualized representations $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_T$ that capture information from the entire sequence.

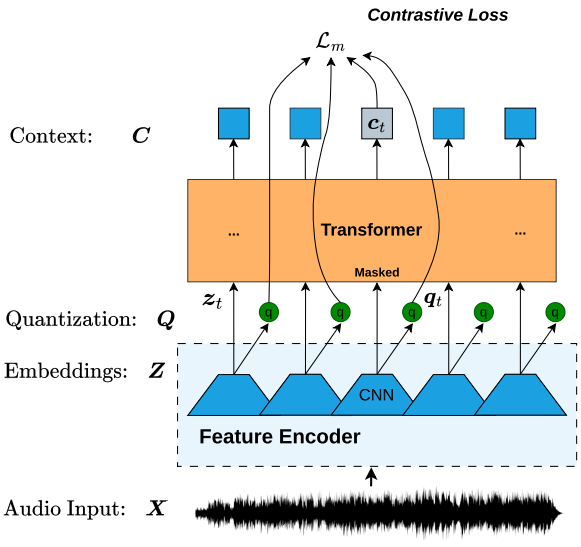


FIGURE 2. Illustration of the wav2vec 2.0 model. Adapted from [11].

B. PRE-TRAINING OBJECTIVE

Pre-training wav2vec 2.0 consists of the joint optimization of a contrastive task $\mathcal{L}_m$ and a diversity task $\mathcal{L}_d$. Then, the objective $\mathcal{L}$ for the whole pre-training stage results in:

$$\mathcal{L} = \mathcal{L}_m + \alpha \mathcal{L}_d, \tag{1}$$

where α is a tuned hyper-parameter.

The contrastive task $\mathcal{L}_m$ measures the discrepancy between the context vectors $\boldsymbol{c}_t$ and the true quantized representation $\boldsymbol{q}_t$ selected from a set of $K + 1$ quantized candidate representations, where K representations are distractors. In contrast, the diversity task $\mathcal{L}_d$ encourages the model to use the codebook entries equally often, increasing the use of quantized codebook representations. Further details about wav2vec 2.0 pre-training objective can be found in [11] and [29].

In this project, we explore the use of wav2vec 2.0 pre-trained with single-channel speech signals, but we also adapt the framework to pre-train the model on multichannel spatial audio.

C. FINE-TUNING WAV2VEC 2.0

Fine-tuning the pre-trained wav2Vec 2.0 model for different speech-related tasks involves adapting the model's representations and hyper-parameters to the specific requirements of the downstream applications. A task-specific head or a linear layer on top of the base model is typically added during fine-tuning. The model learns these new specific weights by training on a smaller and labeled dataset specific to the task.

The fine-tuning loss hardly depends on the task. For example, for Automatic Speech Recognition (ASR), the model may use the Connectionist Temporal Classification (CTC) approach [30] for audio-text alignment, while for speaker identification, fine-tuning might involve contrastive loss functions, where the model learns to discriminate between different speakers. We propose to fine-tune wav2vec 2.0 for both the SED and DOA estimation tasks. We elaborate on the details of the fine-tuning process of our task in Section IV.

IV. METHODOLOGY

In this section, we outline the methodology employed in our project to develop an SSL approach for SELD utilizing the latest version of the wav2vec 2.0 pre-training framework. Our model, named w2v-SELD, adapts both pre-training and fine-tuning frameworks of wav2vec 2.0 to work with spatial audio that contains four audio channels. Pre-training on large amounts of unlabeled 3D audio data takes us a step closer to overcoming the limitations posed by the traditional supervised learning methods, which require large amounts of annotated data to train. By leveraging this pre-training approach, our project aims to develop a highly accurate and robust SELD model fine-tuned on labeled spatial audio datasets with a restricted amount of recordings. In this Section, we describe the pre-training and fine-tuning stages of w2v-SELD and the data augmentation techniques applied for spatial audio.

A. w2v-SELD FOR MULTICHANNEL AUDIO

The feature encoder of our w2v-SELD model takes four channels of raw audio data as input and output feature vectors. The audio recordings follow the Ambisonics B-format for spatial audio [31], are re-sampled into 16 KHz, and standardized to have zero mean and unit variance. We modified the first of the seven convolutional blocks of the feature encoder to accommodate a multichannel input. The remaining blocks follow the architecture of wav2vec 2.0. The output contains a series of feature vectors, where each one represents a segment of 20 ms of input audio. The same feature encoder architecture of w2v-SELD was used for pre-training and fine-tuning.

The Transformer blocks, following the feature encoder, incorporate an encoder Transformer architecture [32] with multi-head self-attention mechanisms. Each attention head is a separate mechanism that learns to attend to different parts of the input data independently, capturing various relationships and dependencies in the data. As in the original wav2vec 2.0, we experiment with two different setups: the BASE model, which contains 12 transformer blocks with 8 attention heads that output an embedding of dimension 768, and a LARGE model with 24 transformer blocks each one with 16 attention heads that output an embedding of dimension 1,024.

The pre-training and fine-tuning stages of the w2v-SELD model are presented in the Figure 3. The pre-training stage is composed of the unsupervised contrastive learning of the w2v-SELD backbone (Feature Encoder and the Transformer blocks) using the unlabeled 3D audio. On the other hand, the fine-tuning stage of w2v-SELD is a supervised step where the SED and DOA branches are trained with a multi-task objective of multi-label classification and multi-output regression.

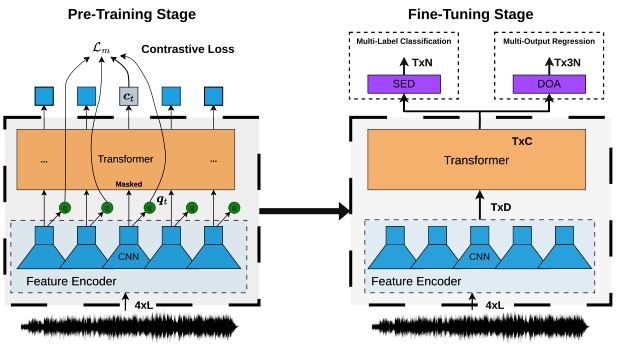


FIGURE 3. Pre-training and fine-tuning Stages of w2v-SELD. On the pre-training stage the w2v-SELD backbone (Feature encoder and the transformer blocks) are trained with the contrastive loss using the unlabeled 3D audio. On the fine-tuning Stage both w2v-SELD and the SED and DOA branches are trained with a multi-task objective of multi-label classification and multi-output regression.

During the pre-training stage, we leverage the pre-trained weights of both the BASE and LARGE models of wav2vec 2.0. This approach allows us to benefit from transfer learning, avoiding the need to initialize the model from scratch with random parameters.

In the fine-tuning stage, we diverged from the original wav2vec 2.0 by replacing both the classification head and the CTC loss typically used for speech recognition. Instead, we introduced SED and DOA branches, incorporating a multi-task objective that includes Binary Cross-Entropy (BCE) for the multi-label classification task and the Logarithm of the hyperbolic cosine (LogCosh) for the multi-output regression task.

B. PRE-TRAINING

We compare the performance of the w2v-SELD model for the SELD task by using the weights pre-trained on spatial or non-spatial audio. As for the non-spatial audio pre-training, the BASE wav2vec 2.0 model was pre-trained with the LibriSpeech (LS-960) [33] dataset which contains 960 hours, while the LARGE wav2vec 2.0 model was pre-trained with LibriVox (LV-60k) [33], comprising a vast 60,000 hours of audio. These datasets encompass English speech recordings featuring American, Canadian, and British accents. For the pre-training on spatial audio, we employed the TAU Spatial Sound Events datasets and the Learning 3D Audio Sources (L3DAS) datasets, which will be further detailed in Section V-A. It is noteworthy that, as unsupervised pre-training does not require annotations, we could use spatial audio datasets annotated at different temporal resolutions.

During pre-training, the model acquired an understanding of audio data structure by predicting masked frames within the input. This masking process involves the random selection and concealment of a subset of audio frames. In essence, certain segments of the input audio remain hidden from the model. To determine which time-steps to mask within the input sequence, each latent speech representation within an utterance is considered as a potential starting time-step with a probability of p, where M denotes the length of each masked span from the respective time step. We adhere to the p and M values as described in [11] for pre-training w2v-SELD.

By using single-channel audio, the model is expected to capture various aspects of audio structure such as phonemes, prosody, and background noise, without the need for explicit phonetic transcriptions or annotations. Conversely, when employing spatial audio for pre-training, the model is expected to learn additional spatial information from the four input channels.

C. w2v-SELD FINE-TUNING

The different pre-trained versions of w2v-SELD are fine-tuned by incorporating two randomly initialized dense layer blocks within the Transformer output—one dedicated to each sub-task of SELD. The SED branch is responsible for multi-label classification, while the DOA branch performs a multi-output regression task. The multi-label classification between the predictions of the w2v-SELD and the SED reference is obtained with the BCE loss. In the multi-output regression the LogCosh is used to estimate the DOA values by regressing the (x, y, z) of the active sounds obtained by the

SED branch.

$$\mathcal{L}_{SELD} = w_1 \times \mathcal{L}_{BCE}(\hat{y}_{SED}, y_{SED}) + w_2 \times \mathcal{L}_{LogCosh}(\hat{y}_{DOA}[mask_{SED}], y_{DOA}[mask_{SED}])$$
$$mask_{SED} = \sigma(\hat{y}_{SED}) > \gamma \quad (2)$$

where:

- w_1 and w_2 are tuned hyper-parameters of weights losses;
- $\hat{y}_{SED}$ and y_{SED} are SED predictions and references;
- $\hat{y}_{DOA}$ and y_{DOA} are DOA estimations and references;
- σ is the sigmoid activation;
- γ is the threshold for active sound obtained at validation.

The layer-wise learning rate is employed for the components: the w2v-SELD Transformer, SED, and DOA branches, with two distinct rates for the Transformer and the branches. The layer rate used in the w2v-SELD Transformer is ten times smaller than the SED and DOA branches. The main idea is to take better advantage of the pre-training with the small learning rate and not forget what was learned during the pre-training.

Adhering to the minimum temporal resolution of 100 ms established for the DCASE Challenges focused on SELD, our model is designed to estimate SED classifications and DOA estimations at intervals of 100 ms or shorter. For the fine-tuning process of w2v-SELD predictions, we explore two distinct approaches: sequence-based prediction (w2v-SELD-SegPred) and frame-based prediction (w2v-SELD-FramePred). These methodologies will be further elaborated below.

1) SEGMENT-BASED PREDICTION (w2v-SELD-SegPred)

In this approach, each spatial audio recording is segmented into 100 ms intervals. The w2v-SELD model is then fed four channels of 100 ms each and outputs a single SED and DOA prediction, as illustrated in Figure 4. The output from the Transformer blocks forms a 2D vector, which passes by a mean-pooling layer applied to the temporal dimension. Subsequently, the resulting 1D vector feeds two separate dense layers, each one for SED and DOA tasks. The dimension of the SED vector, denoted as N, corresponds to the number of sound classes present in the dataset. On the other hand, the DOA vector's dimension is $3N$, representing the Cartesian coordinates associated with each sound class.

2) FRAME-BASED PREDICTION (w2v-SELD-FramePred)

The second approach, named w2v-SELD-FramePred, receives a segment of 3D audio of 2.97 s, following the same window size used by SELDnet, and generates predictions on a per-frame basis. Unlike the w2v-SELD-SegPred method, w2v-SELD-FramePred retains the original temporal dimension T of the embedding. It predicts a vector for each time step $t = 1, \ldots, T$. The choice of input audio duration and frame-wise prediction draws inspiration from SELDnet [7]. However, our w2v-SELD-FramePred approach improves the time resolution of predictions to 20 ms, which results from the CNN blocks within the feature encoder of the w2v-SELD model. This increased temporal resolution serves as an advantage as the model predicts outputs within smaller time windows, adhering to the minimum temporal resolution specified in the DCASE challenges.

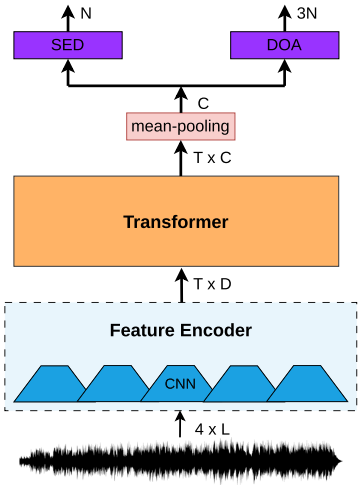


FIGURE 4. Illustration of the w2v-SELD-SegPred approach. *N* denotes the number of SED classes, *T* represents the number of time-steps, and *C* denotes the embedding dimension of the w2v-SELD model.

As depicted in Figure 5, the 2D vector embedding of the w2v-SELD model is fed into both the SED and DOA branches. Subsequently, the SED branch produces a 2D vector with dimensions $T \times N$, while the DOA branch yields a 2D vector of dimension $T \times 3N$. This approach adeptly captures the temporal correlations among sound events within the input segment.

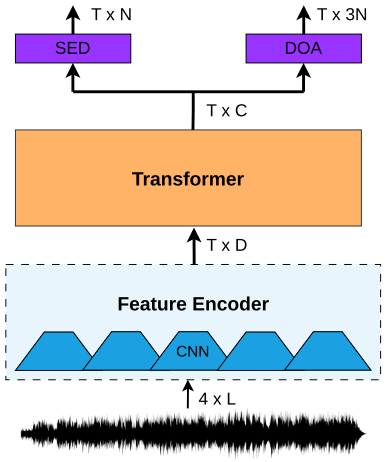


FIGURE 5. Illustration of the w2v-SELD-FramePred. *N* denotes the number of SED classes, *T* represents the number of time-steps, and *C* denotes the embedding dimension of the w2v-SELD model.

D. DATA AUGMENTATION FOR SPATIAL AUDIO

Data augmentation techniques for audio signals are used to create variations in audio data, enhancing the robustness and

generalization of deep learning models. Beyond conventional data augmentation methods, spatial audio benefits from specific techniques tailored for multichannel inputs and the spatialization encoded in the Ambisonics B-format. Augmenting our training samples plays a crucial role, especially during the fine-tuning phase when dealing with limited datasets of 3D audio. The data augmentation techniques applied during the fine-tuning stage of this study are described below.

1) TRADITIONAL AUDIO AUGMENTATIONS TECHNIQUES

We apply traditional techniques individually to each channel of spatial audio. Leveraging the Torch-Audiomentations library [34], we use methods such as random gain (Gain Augmentation) and colored noise addition (Noise Augmentation). These modifications are directly applied to the audio waveform, while the SED and DOA annotations are unmodified.

- Gain Augmentation: Gain applies random gain in decibels (dB), between the range defined by *min_gain_in_db* and *max_gain_in_db*.
- Noise Augmentation: AddColoredNoise adds noise to the audio with a specified signal-to-noise ratio (SNR) range.

2) TIME AND FREQUENCY MASKING (SpecAugment)

We employ the SpecAugment method [9] for time and frequency masking. SpecAugment originally operated on a spectrogram representation, masking random time frames and frequency bins. However, in our case, we implement this technique on the output of the feature encoder. The output embedding has both time and frequency dimensions in our case 512 channels from CNNs blocks. This process is illustrated in Figure 6, where a multichannel audio waveform is processed by the feature encoder to obtain an embedding representation (Figure 6b) where random time and frequency masking is applied. Figure 6c shows the resulting masked embedding, which is then fed into the Transformer blocks.

3) SOUND FIELD ROTATION (CHANNEL FOA SWAPPING)

The B-format is a standardized representation of Ambisonics audio that encodes information about the acoustic scene in all directions. The multichannel FOA representation comprises one omnidirectional channel (W) and three channels– X, Y, and Z–encoding pressure differences between hemispheres, expressed in Cartesian coordinates.

Channel swapping is a technique that rotates the sound field in spatial audio recordings. This technique alters the perceived localization of sounds by swapping audio channels and reversing their signs. These modifications are used to transform the DOA annotations, ensuring congruence with the sound field rotation. Eight types of transformations, as outlined in [35], are randomly applied during training in the fine-tuning stage (Table 1).

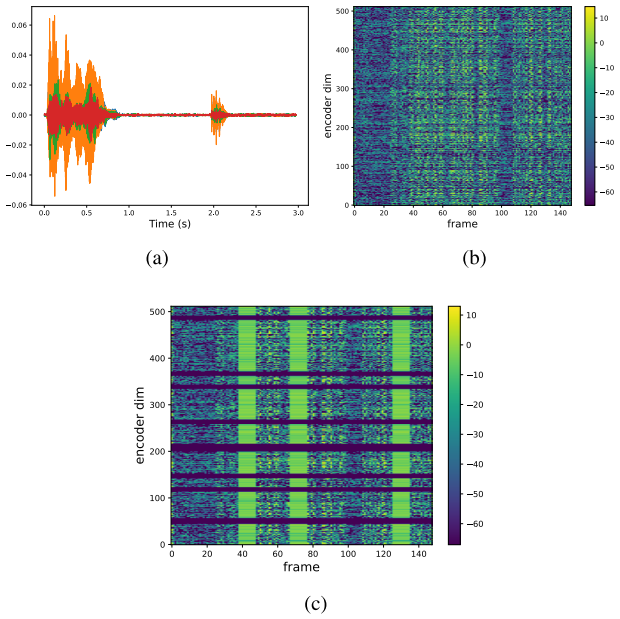


FIGURE 6. SpecAugment on w2v-SELD encoder output. (a) Raw audio signal. (b) Feature encoder output. (c) Feature encoder output with SpecAugment.

TABLE 1. Channel swapping eight transformations applied to Azimuth ϕ and elevation θ, the original arrangement is (C_1, C_2, C_3, C_4). Adapted from [35].

DOA Transformation	FOA
$\phi = \phi, \theta = \theta$	(C_1, C_2, C_3, C_4)
$\phi = -\phi - \frac{\pi}{2}, \theta = \theta$	$(C_1, C_{-4}, C_3, C_{-2})$
$\phi = -\phi + \frac{\pi}{2}, \theta = \theta$	(C_1, C_4, C_3, C_2)
$\phi = \phi + \pi, \theta = \theta$	$(C_1, C_{-2}, C_3, C_{-4})$
$\phi = \phi - \frac{\pi}{2}, \theta = -\theta$	$(C_1, C_{-4}, C_{-3}, C_2)$
$\phi = \phi + \frac{\pi}{2}, \theta = -\theta$	(C_1, C_{-4}, C_3, C_2)
$\phi = -\phi, \theta = -\theta$	$(C_1, C_{-2}, C_{-3}, C_4)$
$\phi = -\phi + \pi, \theta = -\theta$	$(C_1, C_2, C_{-3}, C_{-4})$

V. EXPERIMENTAL SETUP

The w2v-SELD model was pre-trained on an A100 GPU for 3 days for the BASE version and 7 days for the LARGE version. For fine-tuning, any GPU with at least 16 GB of RAM can be employed. In this project, we fine-tuned the w2v-SELD model using an NVIDIA TITAN V GPU, which typically requires approximately 10 hours to complete.

The growing interest in 3D audio analysis has been accompanied by an increase in the number of spatial audio datasets. This expansion has been facilitated by advancements in recording and playback devices able to handle spatial audio. However, the limited availability of datasets result from of

the challenge of collecting accurate annotations for SED and DOA. In this study, we combine multiple 3D audio datasets for pre-training the model, while two smaller datasets were used for fine-tuning. The following section provides a brief overview of the characteristics of the spatial audio datasets considered for this study.

A. SPATIAL AUDIO DATASETS

The spatial audio datasets employed in the pre-training stage of the w2v-SELD model are delineated as follows. The data processing pipeline within the project is orchestrated through the utilization of Data Version Control (DVC) [36], a tool selected to facilitate ease of reproducibility. This systematic approach ensures a robust and transparent methodology in the handling and processing of spatial audio data, aligning with best practices in the field. The use of DVC contributes to the overall reliability of the research, fostering replicability and a comprehensive understanding of the undertaken procedures.

In the pre-training phase we follow the same idea employed in NLP (Natural Processing Language) pre-training such as BERT (Bidirectional Encoder Representations for Transformers) [37], GPT (Generative Pre-trained Transformer) [38] and wav2vec 2.0 where we train on a massive amount of unlabeled data. Thus, from all these pre-training datasets the idea is to source all samples and crop them to four seconds. Two challenge datasets are selected: L3DAS and the Tampere University dataset. A total of six datasets were selected, two from L3DAS: L3DAS21, L3DAS22; and four from Tampere University: TUT-2018, TAU-2019, TAU-2020, and TAU-2021. The summary of the characteristics of the datasets is illustrated in Table 2.

TABLE 2. Summary of datasets of pre-training stage.

Dataset	N^o of samples	Duration (h)	Sample Rate (kHz)
L3DAS21	89946	65	16
L3DAS22	128133	94	16
TUT-2018	2700	22.5	44.1
TAU-2019	500	8	48
TAU-2020	800	13	24
TAU-2021	600	10	24

The L3DAS21 dataset is utilized for pre-training the BASE model, while the L3DAS22 dataset is used for the LARGE model, following the wav2vec 2.0 framework, where the LARGE model is trained on a larger dataset.

1) L3DAS DATASETS

The L3DAS21 [39] and L3DAS22 [40] challenges focus on 3D speech enhancement (SE) and 3D SELD. These datasets were created by convolving single-channel sound events from FSD50K [41] or speech from LibriSpeech with impulse responses (IRs) captured using two First-Order Ambisonic (FOA) microphones. Each dataset contains two subsets: one designed for SE and the other for SELD. In the SE subset, each recording contains speech, whereas, in the SELD subset, voiced sounds including speech may be present. While only the subsets for SELD provide annotations for SED and DOA, the SE subsets can be used along with the SELD subset during the unsupervised pre-training stage of our model, as annotations are unnecessary for this stage; only Ambisonic recordings are required.

The L3DAS21 dataset comprises approximately 65 hours of Ambisonics audio recordings, distributed into 50 hours dedicated to SE and 15 hours for SELD. To synthesize the 3D audio, 14 types of sounds typically encountered in an office environment were convolved with the IRs. Clean sound samples were sourced from the LibriSpeech and FSD50K [41] datasets, while four types of office-like background noises were chosen from FSD50K. The sampling frequencies for the SE and SELD subsets are 16 kHz and 32 kHz, respectively.

In contrast, L3DAS22, an extension of the L3DAS21 dataset, offers an expanded dataset with an increased volume of data. It comprises approximately 94 hours of audio recordings, with 86 hours allocated for the SE task and 7.5 hours for SELD. The type of sounds, background noises, and recording strategies remain consistent with those of the L3DAS21 dataset.

2) TAMPERE UNIVERSITY DATASETS

The Tampere University of Technology (TUT) released a series of SELD datasets specifically designed for use in the DCASE competitions. All these datasets contain recordings in the FOA B-format for spatial audio, alongside annotations for SED and DOA. In this study, we consider the datasets released annually from 2018 to 2021.

TUT-2018 [7]: This dataset incorporates stationary sound events placed synthetically in specific spatial coordinates. For this study, we employ the ANSYN and REAL datasets from TUT-2018, which were synthesized using artificial and real IRs, respectively. The ANSYN dataset simulates an anechoic environment without reverberation, while the REAL dataset uses IRs recorded using a spherical microphone array in a university corridor surrounded by classrooms. Sound events isolated from the urbansound8k dataset [42] were used in both datasets. REAL comprises 10 types of sounds, while ANSYN considers 11 classes. Each dataset contains up to three overlapping sound events in the recordings. There are 2,700 spatial audio files in each dataset, sampled at 44.1 kHz for 30 s, totaling 22.5 hours in each dataset.

TAU-2019 [43]: This dataset was synthesized by convolving 11 types of static sound events with real-life IRs collected from five rooms with different reverberant characteristics at multiple spatial coordinates. The sound events remain stationary with fixed locations throughout their duration. The dataset comprises two sub-datasets: TAU Spatial Sound Events 2019 - Ambisonic (FOA), and TAU Spatial Sound Events 2019 - Microphone Array (MIC). The FOA follows

the B-format of Ambisonics, while MIC provides the directional microphone recordings of a tetrahedral microphone array (the A-format). Each sub-dataset's development set includes 400 one-minute-long recordings sampled at 48 kHz, and the evaluation set consists of 100 one-minute recordings. Approximately 8 hours of spatial audio are available in this dataset.

TAU-2020 [44]: This dataset introduced diversified acoustical conditions and sound event trajectories. It contains 714 sound examples of 14 categories convolved with recorded IRs, encompassing static and dynamic trajectories. Similar to TAU-2019, FOA and MIC sub-datasets were released. Each sound event in the sound scene is linked to its DOA trajectory, onset, and offset times. The development set includes 600 one-minute-long recordings sampled at 24 kHz, and the evaluation set comprises 200 one-minute recordings. The dataset offers around 13 hours of spatial audio.

TAU-2021 [44]: Released for the DCASE2021 Challenge, this dataset features moving sound sources, directional interference events, and an additional layer of background noise in all samples. It includes 12 classes of sound events and realistic spatialization and reverberation achieved through IRs collected in 13 different enclosures. The development dataset consists of 500 one-minute-long audio samples at a sampling rate of 24 kHz, while the testing set contains 100 one-minute-long audio samples. This dataset comprises 10 hours of spatial audio.

B. PRE-TRAINING HYPER-PARAMETERS

Adhering to the pre-training methodology of wav2vec 2.0, wherein the LARGE model employs a more extensive dataset compared to the BASE model, we integrate the L3DAS22 dataset to augment the dataset size for training our LARGE w2v-SELD model. Generally, we combine unlabeled 3D audio data for pre-training the w2v-SELD model as follows:

- **L3DAS21-SELD**: Combination of L3DAS21, TUT-2018, TAU-2019, TAU-2020, and TAU-2021 datasets.
- **L3DAS22-SELD**: Combination of L3DAS22, TUT-2018, TAU-2019, TAU-2020, and TAU-2021 datasets.

We downsample the 3D audio signals to a 16 kHz frequency and subsequently divide them into frames lasting 4 s. The BASE w2v-SELD model was trained using the L3DAS21-SELD set, resulting in 179,442 utterances, while the LARGE w2v-SELD model utilized the L3DAS22-SELD set, comprising 227,797 utterances after preprocessing.

The original wav2vec 2.0 pre-training stage uses audio segments lasting 15.6 s. However, we have reduced the duration of our segments. This adjustment was made due to the heightened complexity of learning unsupervised representations for SED and DOA from multichannel audio, which presents a more challenging task compared to the ASR goal of wav2vec 2.0.

In the w2v-SELD pre-training, the model is optimized by initially warming up the learning rate for the first 32,000 updates, reaching a peak of 5×10^{-4} for both BASE and LARGE models, followed by a linear decay. The BASE model was pre-trained using the L3DAS21-SELD dataset for 400k updates, while the LARGE model was pre-trained on the L3DAS22-SELD dataset for 600k updates.

C. FINE-TUNING HYPER-PARAMETERS

A tri-stage learning rate schedule is implemented to cautiously adapt to the new dataset without compromising the knowledge gained during pre-training. To summarize, the following hyper-parameters are configured:

- The tri-stage schedule comprises 10% warm-up, 30% constant learning, and 60% decay phases.
- The learning rates are set as follows: 5×10^{-5} for the w2v-SELD model until the Transformer's output, 5×10^{-4} for SED and DOA branches.

As observed, the learning rate for the w2v-SELD model until the Transformer's output is lower than that of the SED and DOA branches. This intentional discrepancy aims to leverage the knowledge acquired during the pre-training stage.

D. EVALUATION METRICS

The evaluation metrics used for SELD gauge the performance of systems based on their ability to accurately detect and localize sound events. We have adopted segmented-based metrics as outlined in [45], commonly referred to as frame-based metrics. For SED, the metrics used are F-score (F1) and Error Rate (ER). Regarding the DOA metrics, we use the $\text{DOA}_{\text{error}}$ to measure the difference between the estimated and reference DOA angle, and the Frame-Recall (FR) defined as the relation between true positive and true negative estimations.

The overall score, named $\text{SELD}_{\text{score}}$, combines both SED and DOA metrics as follows:

$$\text{SELD}_{\text{score}} = \frac{\text{SED}_{\text{score}} + \text{DOA}_{\text{score}}}{2}, \tag{3}$$

where

$$\text{SED}_{\text{score}} = \frac{\text{ER} + (1 - \text{F1})}{2}, \tag{4}$$

$$\text{DOA}_{\text{score}} = \frac{\text{DOA}_{\text{score}}/180 + (1 - \text{FR})}{2}. \tag{5}$$

In an ideal scenario, the metrics for a perfect SELD model would be $\mathbf{F1\%} = 100$, $\mathbf{ER} = 0$, $\mathbf{DOA_{error}} = 0$, and $\mathbf{FR\%} = 100$, resulting in a $\mathbf{SELD_{score}}$ of 0. In the DCASE2020 Challenge, the SED metrics (i.e., F1 and ER) consider a frame prediction as a true positive if the spatial error, calculated as the angular distance between reference and predicted DOA, is less than 20° [44].

VI. RESULTS

In the following sections, we present the results of our proposed w2v-SELD approach fine-tuned using the

TABLE 3. Comparison between prediction strategies: w2v-SELD-FramePred and w2v-SELD-SegPred fine-tuned using the TAU-2019 dataset.

Strategy	w2v-SELD-SegPred		w2v-SELD-FramePred	
Pre-training Set	LS-960	L3DAS21-SELD	LS-960	L3DAS21-SELD
↓ER	1.12	0.94	0.14	**0.10**
↑F1%	54.90	58.83	91.50	**94.20**
$\downarrow DOA_{error}$	17.08	17.06	**4.80**	4.88
↑FR%	72.84	74.04	91.66	**93.12**
$\downarrow SELD_{score}$	0.484	0.427	0.084	**0.063**

*The arrows indicate whether the metric improves with an increase (↑) or decrease (↓) in its value.

TAU-2019 and TAU-2020 datasets. Initially, we examine the influence of employing the w2v-SELD-FramePred approach in contrast to the w2v-SELD-SegPred approach (Section VI-A). Subsequently, in Section VI-B, we assess the effects of pre-training our w2v-SELD model using various unlabeled datasets. This section offers a comparison between model performance during linear evaluation with frozen parameters and when parameters are updated during fine-tuning. Moving forward, Section VI-C compares our approaches against baseline systems and state-of-the-art techniques. Finally, we evaluate the impact of data augmentation on fine-tuning the SELD task (Section VI-D).

A. SEQUENCE CLASSIFICATION VERSUS AUDIO FRAME CLASSIFICATION

The primary objective of this experiment is to compare the performance between the w2v-SELD-FramePred and w2v-SELD-SegPred approaches, regardless of the type of audio used for pre-training. To accomplish this, we implemented both strategies on two models: the wav2vec 2.0 BASE model, pre-trained on single-channel audio data, and our w2v-SELD model, pre-trained on spatial audio data. Subsequently, all models underwent fine-tuning using the TAU-2019 dataset.

The results presented in Table 3 demonstrate the advantages of employing the w2v-SELD-FramePred approach across both pre-training sets. The adoption of w2v-SELD-FramePred technique led to a significant enhancement in the $SELD_{score}$ by 83% and 86% for the LS-960 and L3DAS21-SELD pre-training sets, respectively. This remarkable improvement emphasizes the significance of leveraging temporal dependencies within the input representation. Despite using short frames lasting 2.97 s for fine-tuning, the w2v-SELD-FramePred approach effectively utilizes previous information to enhance future predictions. Conversely, the wav2vec2-SeqPred approach loses temporal relationships among samples, resulting in a detrimental impact on the SELD system's performance. Although wav2vec2-SeqPred might seem the more intuitive choice for the SELD task, it consistently fails to yield optimal results, regardless of the pre-training data employed.

TABLE 4. Linear evaluation of the frozen parameters of the w2v-SELD BASE model pre-trained various datasets. Evaluation conducted on the TAU-2019 dataset.

Pre-training Set	-	LS-960	L3DAS21-SELD
↓ER	0.90	0.81	**0.56**
↑F1%	26.67	41.98	**65.19**
$\downarrow DOA_{error}$	**70.74**	72.15	80.06
↑FR%	50.27	52.96	**58.45**
$\downarrow SELD_{score}$	0.631	0.565	**0.442**

*The arrows indicate whether the metric improves with an increase (↑) or decrease (↓) in its value.

TABLE 5. Performance results of the w2v-SELD BASE model pre-trained using various datasets and fine-tuned on the TAU-2019 dataset.

Pre-training Set	-	LS-960	L3DAS21-SELD
↓ER	0.17	0.14	**0.10**
↑F1%	89.45	91.50	**94.20**
$\downarrow DOA_{error}$	4.97	**4.80**	4.88
↑FR%	92.01	91.66	**93.12**
$\downarrow SELD_{score}$	0.096	0.084	**0.063**

*The arrows indicate whether the metric improves with an increase (↑) or decrease (↓) in its value.

B. IMPACT OF PRE-TRAINING ON SPATIAL AUDIO

Upon evaluating our prior analyses, the audio frame prediction approach (i.e., w2v-SELD-FramePred) emerged as the most effective method for both sound event classification and localization. Consequently, this approach is employed in the subsequent experiments. To further assess the influence of pre-training on the SELD task, a series of fine-tuning experiments were conducted on the TAU-2019 dataset using weights acquired from various pre-training configurations of our w2v-SELD model. During fine-tuning, the parameters of w2v-SELD are not back-propagated, similar to a model in inference mode. With our focus on evaluating the representations learned during pre-training using a linear evaluation, all layers of the w2v-SELD model before the SED and DOA branches are frozen. This ensures that only the SED and DOA layers' weights are updated during fine-tuning, while the preceding layers remain isolated from this process.

We experiment with three configurations: 1) no pre-training, for which the w2v-SELD model's weights are randomly initialized and frozen during fine-tuning, 2) pre-training on LS-960 single-channel speech data, and 3) pre-training on the L3DAS21-SELD set of unlabeled spatial audio data. It is essential to note that across these configurations, the weights of the w2v-SELD model before the SED and DOA branches remain frozen, and only the SED and DOA layers are updated during fine-tuning. These experiments aim to gain a deeper understanding of the impact of the pre-training stage on the performance of the SELD task and to identify any potential areas for improvement.

Table 4 exhibits the results for the mentioned configurations. These metrics are not expected to reach optimal

TABLE 6. Comparison of w2v-SELD-FramePred performance against baseline and state-of-the-art methods.

Dataset	Model	Unlabeled Data	↓ER	↑F1 %	↓DOA_{error}	↑FR %	↓$SELD_{score}$
TAU-2019	CRNN (baseline) [43]	-	0.28	85.40	24.60	85.40	0.177
	Ensemble-CRNN [8]	-	*0.08*	*94.70*	*3.70*	*96.80*	*0.046*
	This work						
	BASE	LS-960	0.142	91.50	4.80	91.66	0.080
	LARGE	LV-60k	0.139	92.04	4.84	92.77	0.079
	BASE	L3DAS21-SELD	0.099	94.20	4.88	**93.12**	0.063
	LARGE	L3DAS22-SELD	**0.096**	**94.66**	**4.67**	93.05	**0.061**
TAU-2020	CRNN (baseline) [44]	-	0.69	41.30	23.10	62.40	0.445
	Ensemble-CRNN [18]	-	*0.20*	*84.90*	*6.00*	*88.50*	*0.125*
	This work						
	BASE	LS-960	0.363	72.56	9.06	76.43	0.231
	LARGE	LV-60k	0.378	70.77	10.68	76.43	0.241
	BASE	L3DAS21-SELD	0.307	76.96	**8.37**	79.96	0.196
	LARGE	L3DAS22-SELD	**0.301**	**77.17**	8.51	**80.83**	**0.192**

*The arrows indicate whether the metric improves with an increase (↑) or decrease (↓) in its value. Additionally, we have highlighted in bold the metrics of the best approach observed among the experiments conducted in this study for each dataset. In italic is highlighted the state-of-art metric obtained in each dataset.

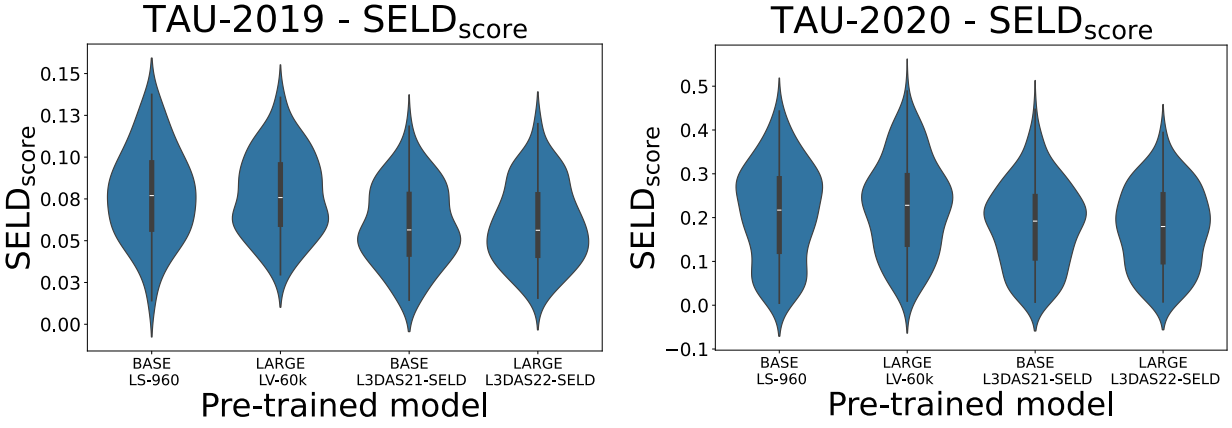


FIGURE 7. Violin plots of the $SELD_{score}$ per-record for w2v-SELD models LARGE/L3DAS22-SELD, BASE/L3DAS21-SELD, LARGE/LV-60k, and BASE/LS-960 for the TAU-2019 and TAU-2020 datasets.

values, as our objective here is to evaluate the pre-training representations rather than optimal performance. Notably, we observed a consistent trend of improvement with pre-training, even when utilizing single-channel data, compared to fine-tuning the SED and DOA branches from scratch. Pre-training the model on the LS-960 unlabeled speech dataset improved the $SELD_{score}$ by 10%. The use of spatial audio for pre-training led to a 23% $SELD_{score}$ improvement compared to single-channel data. Moreover, significant improvements in SED metrics were observed, with a 55.28% increase in F1 when pre-training on L3DAS21-SELD data compared to the LS-960 dataset, and a 144.43% improvement compared to the model without pre-training. Additionally, it is evident that the DOA metrics are influenced by updating the parameters of the w2v-SELD model during fine-tuning, which significantly contributes to their improvement.

Subsequently, we fine-tuned the same pre-training configurations but unfreezing the w2v-SELD model parameters. This allows the initialized pre-trained weights to be updated during fine-tuning for improved adaptation to the SELD task. Table 5 highlights the advantageous outcomes of the proposed w2v-SELD-FramePred approach, showcasing a 20% $SELD_{score}$ enhancement when pre-trained on the LS-960 dataset and a 40% improvement using the L3DAS21-SELD spatial audio set for pre-training, compared to no pre-trained weights.

C. COMPARISON WITH STATE-OF-THE-ART MODELS

The proposed w2v-SELD model, implemented with the w2v-SELD-FramePred approach, was fine-tuned and subsequently evaluated on both the TAU-2019 and TAU-2020

datasets. Our approach was compared against the baseline models provided for each dataset, as well as the current state-of-the-art models. We present the summarized results for both our BASE and LARGE w2v-SELD models, pre-trained on the L3DAS21-SELD and L3DAS22-SELD datasets, respectively, in Table 6.

On TAU-2019, all models trained in this study outperformed the baseline system [43]. Notably, our best model (LARGE/L3DAS22-SELD) exhibited a 66% improvement in $\text{SELD}_{\text{score}}$ compared to the baseline. Moreover, our top-performing model achieved an $\text{SELD}_{\text{score}}$ very close to the state-of-the-art system for TAU-2019 [8], trailing behind by only 1.64%.

Moving to TAU-2020, the SED metrics were adjusted to align with the DCASE2020 Challenge criteria, where a spatial error of less than 20° was required to classify as a true positive. Once again, all our experiments surpassed the baseline system for TAU-2020 [44] in both SED and DOA metrics. Our best model (LARGE/L3DAS22-SELD) exhibited a 57% improvement in overall $\text{SELD}_{\text{score}}$ compared to the baseline system. However, there remains room for enhancement in comparison to the state-of-the-art system proposed in [18], where our best approach lags by 32% in the $\text{SELD}_{\text{score}}$.

We highlight the fact that the significant performance improvements in our approach primarily stemmed from the pre-training stage using unlabeled multichannel audio. In contrast, the state-of-the-art system proposed in [18] heavily relies on manually synthesizing new Ambisonics audio samples by mixing non-overlapping samples and transforming them into spatial audio format. Given that synthesized samples require manual labeling, this process is time-consuming and challenging to replicate. Additionally, the system proposed in [18] does not provide either a repository or the data used for training the model. Regarding computational efficiency, the w2v-SELD models demonstrate their per-sample inference time, as shown in Table 7. These results are compared with traditional SELD approaches referenced in [1].

We emphasize the availability of a repository of our proposed approach, where the trained models can be utilized for SELD task inference or the source code can serve as a foundational framework for further enhancements.

D. IMPACT OF DATA AUGMENTATION FOR SPATIAL AUDIO

As evinced in [18], the use of data augmentation holds significant importance in enhancing the SELD task, particularly in reducing DOA metrics. In light of this, we conducted a further investigation to assess the impact of employing data augmentation techniques, detailed in Section IV-D, on the performance of our best approach (LARGE/L3DAS22-SELD). The results comparing the SELD metrics with and without data augmentation for the best models on both TAU-2019 and TAU-2020 are presented in Table 8.

The inclusion of data augmentation improved the overall $\text{SELD}_{\text{score}}$ for both datasets. Specifically, enhancements of 1.6% and 16.16% were observed for TAU-2019 and TAU-2020, respectively. This demonstrates the relevance of augmenting labeled spatial audio datasets for effective fine-tuning purposes. While we implemented well-established data augmentation techniques, we acknowledge the potential for further exploration of additional methods, especially considering the limited DOAs in the development set. Exploring novel augmentation techniques might yield greater improvements in DOA metrics. However, we highlight the fact that the improvement in metrics introduced by our top-performing approaches is not solely dependent on data augmentation techniques during fine-tuning. Instead, our focus lies on leveraging unlabeled spatial audio, a more feasible approach when manual annotations are unavailable.

TABLE 7. Inference time of w2v-SELD models against traditional SELD methods.

Models	Inference Time (s)
SELDnet	0.384
SELD-TCN	0.012
w2v-SELD (BASE)	0.017
w2v-SELD (LARGE)	0.024

TABLE 8. Comparison of top-performing w2v-SELD-FramePred model with and without data augmentation (DA).

Dataset	TAU-2019		TAU-2020	
DA	×	✓	×	✓
↓ER	0.106	**0.096**	0.368	**0.301**
↑F1%	94.29	**94.66**	72.29	**77.17**
$\downarrow\text{DOA}_{\text{error}}$	5.57	**4.67**	11.18	**8.51**
↑FR%	**94.55**	93.05	78.97	**80.83**
$\downarrow\text{SELD}_{\text{score}}$	0.062	**0.061**	0.229	**0.192**

*The arrows indicate whether the metric improves with an increase (↑) or decrease (↓) in its value.

E. PER-RECORD ANALYSIS OF w2v-SELD

We conduct a per-record analysis on the test-set samples of the TAU-2019 and TAU-2020 datasets to gain a more granular understanding of the model's behavior on specific subsets of inputs. In this analysis, we compute the SELD metrics (ER, F1%, $\text{DOA}_{\text{error}}$, FR%, and $\text{SELD}_{\text{score}}$) for each record in both datasets, using the pre-trained w2v-SELD models introduced in Section VI-C. This per-record evaluation enables us to identify which metrics are more challenging to improve.

Figure 7 shows the violin plots for the $\textbf{SELD}_{\textbf{score}}$, which is the metric that reflects the overall performance of the models. Due to the increased challenging conditions of TAU-2020 dataset, we observe an increased average $\textbf{SELD}_{\textbf{score}}$ across all models compared to TAU-2019. Also, the increased dispersity in the model's distribution demonstrates an increased uncertainty in the predictions. Moreover, we evince an improved and more condensed $\textbf{SELD}_{\textbf{score}}$ for both datasets when pre-training the models with spatial audio. Additionally, the increased variability in the model

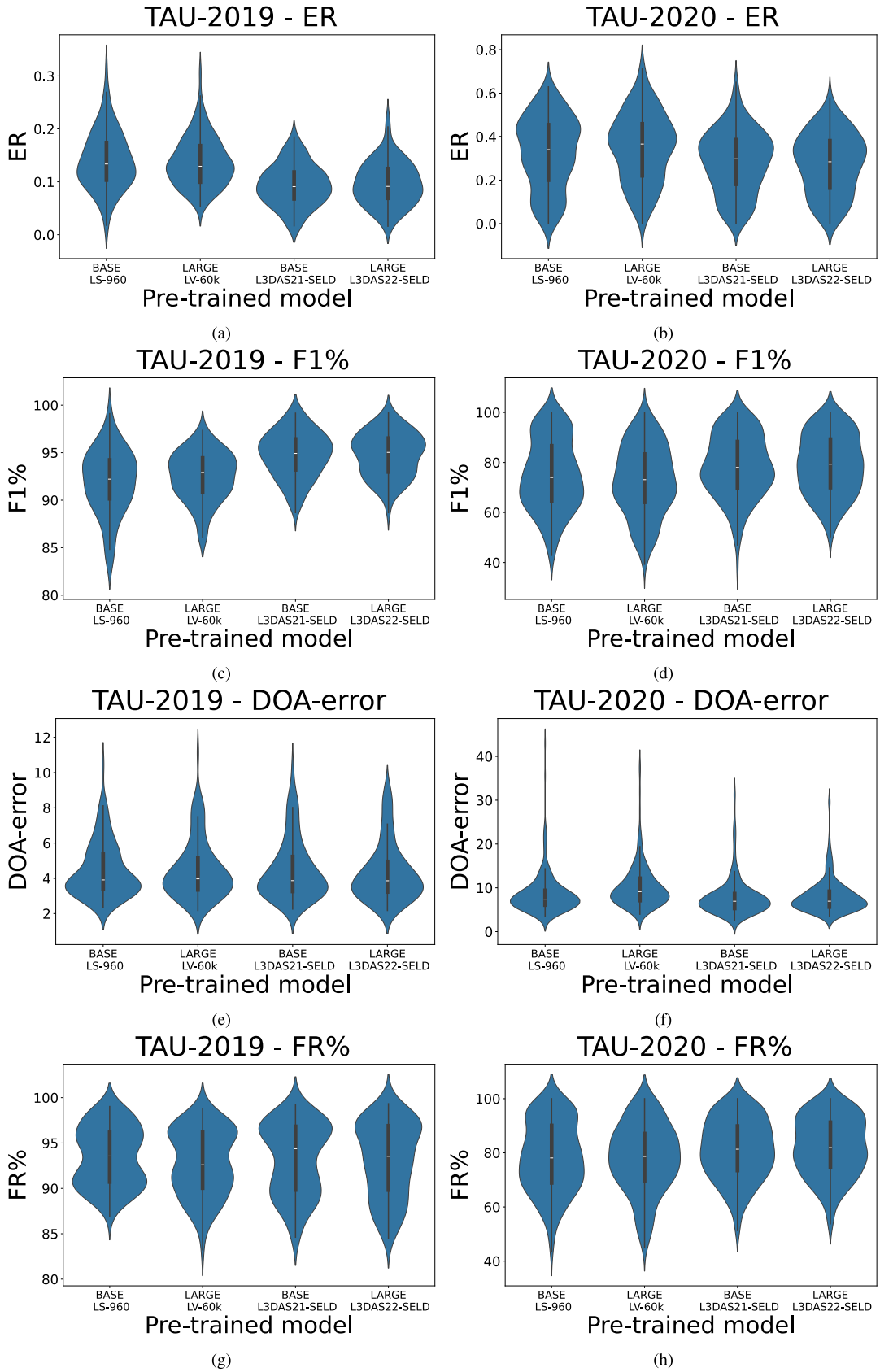
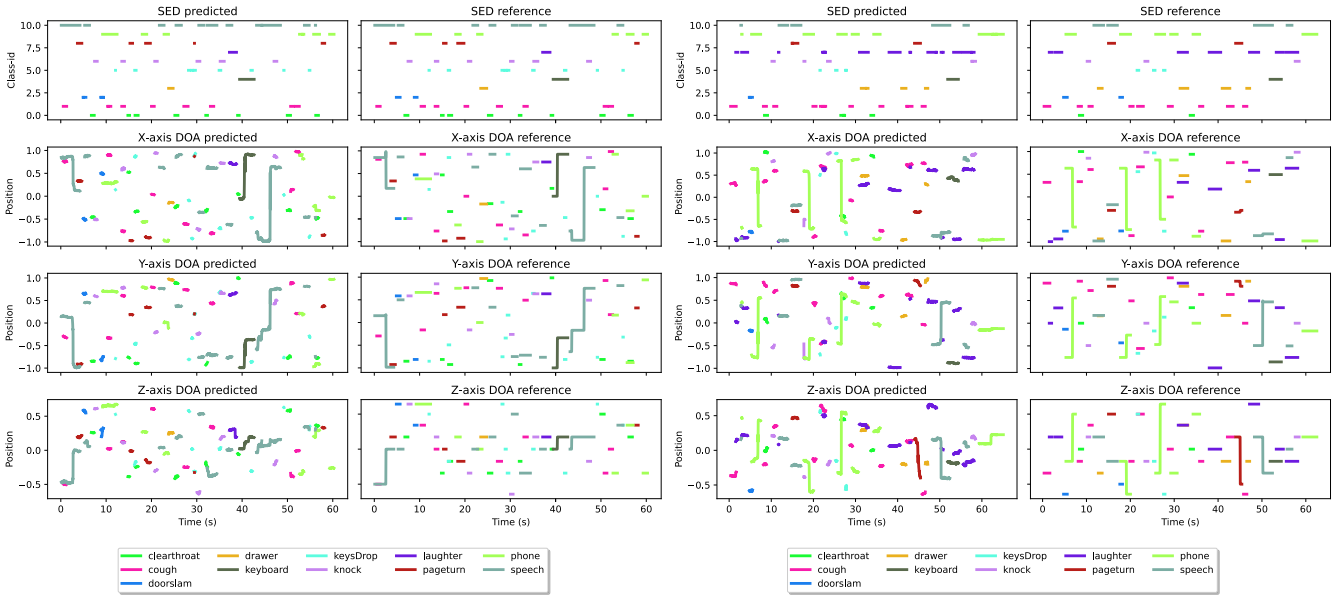


FIGURE 8. Violin plot of the SELD metrics (ER, F1%, DOA_{error}, FR%) per-record for w2v-SELD models LARGE/L3DAS22-SELD, BASE/L3DAS21-SELD, LARGE/LV-60k, and BASE/LS-960.

TABLE 9. SELD metrics (mean ± std) per-record of the w2v-SELD model (LARGE/L3DAS22-SELD) on TAU-2019 and TAU-2020 under OV1 and OV2 conditions.

Dataset	TAU-2019†		TAU-2020†	
Metric	OV1 (50 records)	OV2 (50 records)	OV1 (100 records)	OV2 (100 records)
↓ER	**0.09 ± 0.04**	0.1 ± 0.04	**0.20 ± 0.13**	0.35 ± 0.10
↑F1%	**95.42 ± 2.05**	94.15 ± 2.43	**85.03 ± 11.61**	73.16 ± 9.37
↓DOA_{error}	**3.18 ± 0.52**	5.48 ± 1.60	**6.27 ± 2.20**	9.88 ± 4.36
↑FR%	**96.52 ± 1.62**	89.74 ± 2.73	**86.41 ± 11.24**	77.96 ± 8.70
↓$SELD_{score}$	**0.05 ± 0.02**	0.07 ± 0.02	**0.13 ± 0.09**	0.22 ± 0.07

*The arrows indicate whether the metric improves with an increase (↑) or decrease (↓) in its value.
† TAU-2019 has 50 records of OV1 and 50 records of OV2. Also, TAU-2020 has 100 records of OV1 and 100 records of OV2.



(a) (b)

FIGURE 9. w2v-SELD predictions versus reference on two examples of TAU-2019. (a) Best OV2 record with $SELD_{score} = 0.025$. (b) Worst OV2 record with $SELD_{score} = 0.120$.

distributions indicates a greater uncertainty in predictions. However, we observe a significant reduction in score variability and improved $SELD_{score}$ for both datasets when the models are pre-trained with spatial audio.

Violin plots for the individual SELD task metrics are shown in Figure 8. A clear improvement is observed in the ER and F1% metrics, with reduced variability in the models pre-trained with spatial audio (Figure 8a-b). Localization-related metrics (DOA_{error}, FR%) remain challenging but show modest improvement with spatial audio pre-training across both datasets. This analysis supports our hypothesis that pre-training models with spatial audio are critical for enhancing performance, as opposed to relying solely on single-channel recordings, as done in previous works.

F. w2v-SELD ERROR ANALYSIS

We conduct a statistical analysis on both the TAU-2019 and TAU-2020 datasets to identify scenarios where the w2v-SELD model underperforms. Specifically, we analyze the best-performing model (LARGE/L3DAS22-SELD) for each dataset. The most frequent failure cases are associated with the number of overlapping sounds, referred to as OV, which can be either OV1 (no overlapping sound) or OV2 (two overlapping sounds). Records containing OV2 sounds tend to show reduced performance compared to OV1, highlighting the increased difficulty in identifying multiple sound sources, as shown in Table 9. In both datasets, we observe that the w2v-SELD model performs better on OV1 records compared to OV2.

Figure 9 illustrates sound class identification over time and sound localization in Cartesian coordinates for two TAU-2019 records. The predictions were made using our best-performing model (LARGE/L3DAS22-SELD), with the ground truth values shown to the right of each example. We selected two contrasting examples: Figure 9a presents a case with top metric predictions and highly accurate

estimates, while Figure 9b demonstrates a less accurate prediction with more errors. These visualizations aim to provide an intuitive understanding of the challenges posed by the SELD task, where sound sources are short in duration, some overlap with other sounds, and their Cartesian positions change rapidly within the acoustic environment.

Even in the more challenging example (Figure 9b), our model demonstrates a reasonable ability to detect subtle sounds, such as a page turning or someone clearing their throat, while also providing a decent estimation of the sound source's location. Notably, when comparing our model's performance on OV2 records with human listening capabilities, our approach achieves an FR% of 89.74 ± 2.73. This is in contrast to human listeners, who, according to [46], achieve an accuracy range of 42–84% for identifying and localizing two simultaneous overlapping sound sources.

VII. DISCUSSION

While there remains potential for further improvements in the DOA metrics ($\text{DOA}_{\text{error}}$ and FR) for both datasets, it is worth highlighting that our study has achieved promising localization metrics solely by utilizing raw audio as input. Significantly, we achieved this without reliance on spectrograms, phase, or intensity vectors. To the best of the authors' knowledge, this constitutes the first instance of such results in the field. Furthermore, our models were pre-trained on relatively small datasets, approximately 65 and 94 hours for the BASE and LARGE models, respectively, in contrast to speech recognition where the original wav2vec 2.0 was pre-trained with datasets of 960 and 60,000 hours.

The achievements of our study underline the potential for successful localization without intricate audio pre-processing techniques, due to the self-supervised pre-training stage. These findings open new research avenues and emphasize the significance of our contribution to the field.

In terms of computational efficiency, both the BASE and LARGE versions of the w2v-SELD models demonstrated satisfactory performance. They maintained a competitive runtime, without significant slowdown, when compared to traditional CRNN models. This balance of performance and efficiency highlights the potential of w2v-SELD models for practical applications in SELD tasks.

The wav2vec 2.0 framework demonstrates its task-agnostic nature concerning audio and speech data, extending beyond exclusive use in the speech recognition domain. This suggests that the pre-task learns comprehensive representations of audio data, transcending specific task boundaries.

It was observed that the w2v-SELD has some limitations throughout the process: 1) The pre-training dataset needs to be large enough to provide diverse samples, enabling the model to extract high-level feature representations effectively. Datasets like L3DAS21 (65 hours) and L3DAS22 (94 hours) are relatively small compared to typical wav2vec 2.0 pre-training datasets, such as LS-960 (960 hours) and LV-60k (60,000 hours). 2) During fine-tuning, it is necessary to experiment with different hyperparameter weightings for the multi-task objective to optimize $\textbf{SELD}_{\textbf{score}}$.

VIII. CONCLUSION

In this paper, we introduced a novel approach to SELD by leveraging an SSL pre-training framework based on wav2vec 2.0. Our method demonstrated that robust SELD models can be developed without heavy reliance on labeled spatial audio data. We tailored the fine-tuning process of our w2v-SELD model to accurately predict SED and DOA estimation at the frame level, enhancing both precision and performance.

Through the evaluation of two DCASE challenges, we showcased the adaptability and effectiveness of our w2v-SELD model across different pre-training configurations. Pre-training with diverse and unlabeled spatial audio datasets resulted in improvements in the $\text{SELD}_{\text{score}}$, including a 20% increase when using pre-trained weights compared to training from scratch, and a 40% improvement when pre-training with spatial audio instead of single-channel audio. Our model achieved a 66% improvement on DCASE 2019 and 57% on DCASE 2020 over baseline systems, performing close to state-of-the-art levels while utilizing raw spatial audio input without relying on spectrograms or other feature-engineered inputs. Also, the proposed w2v-SELD model can work alongside known SELD data augmentations.

Furthermore, we introduced a toolkit specifically designed for SSL pre-training in spatial audio applications, providing a valuable resource for future research in this area. Our work demonstrates advancements in the field of SELD and opens avenues for further innovations in audio processing and machine learning. In future research, we plan to pursue the following directions:

- **Integration of Conformer-Based Models**: Investigate the application of the Conformer architecture to enhance performance in SELD tasks.
- **SSL with Phase Representations**: Pre-train the w2v-SELD model using phase-based inputs, such as intensity vectors, which provide richer and more informative data compared to raw audio signals.
- **Dynamic Task Weighting in Multi-Task Learning**: Implement dynamic task weighting strategies to automatically adjust the importance of SED and DOA tasks during training. This will allow the model to adapt to task difficulty or uncertainty without requiring manual tuning.
- **Knowledge Distillation for Model Efficiency**: Leverage knowledge distillation from large pre-trained models to create lightweight SELD models, optimized for deployment in low-power environments such as mobile and wearable devices.

REFERENCES

[1] K. Guirguis, C. Schorn, A. Guntoro, S. Abdulatif, and B. Yang, "SELD-TCN: Sound event localization & detection via temporal convolutional networks," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 16–20.

[2] M. Sammarco, T. Z. Stellantis, L. Gantert, and M. E. M. Campista, "Sound event detection via pervasive devices for mobility surveillance in smart cities," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops other Affiliated Events (PerCom Workshops)*, Mar. 2024, pp. 581–586.

[3] I. Marques, J. Sousa, B. Sá, D. Costa, P. Sousa, S. Pereira, A. Santos, C. Lima, N. Hammerschmidt, S. Pinto, and T. Gomes, "Microphone array for speaker localization and identification in shared autonomous vehicles," *Electronics*, vol. 11, no. 5, p. 766, Mar. 2022.

[4] S. Schneider and P. W. Dierkes, "Localize animal sound events reliably (LASER): A new software for sound localization in zoos," *J. Zoological Botanical Gardens*, vol. 2, no. 2, pp. 146–163, Apr. 2021.

[5] H. Liang, J. Chen, F. Khan, G. Srivastava, and J. Zeng, "Audio-visual event localization using multi-task hybrid attention networks for smart healthcare systems," *ACM Trans. Internet Technol.*, Mar. 2024.

[6] Z. Ahmad, T.-K. Nguyen, A. Rai, and J.-M. Kim, "Industrial fluid pipeline leak detection and localization based on a multiscale mann-Whitney test and acoustic emission event tracking," *Mech. Syst. Signal Process.*, vol. 189, Apr. 2023, Art. no. 110067.

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

[8] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of CRNN models," 2019, *arXiv:1908.00766*.

[9] 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," 2019, *arXiv:1904.08779*.

[10] S. V. and S. G. Kooolagudi, "Polyphonic sound event localization and detection using channel-wise FusionNet," *Appl. Intell.*, vol. 54, no. 6, pp. 5015–5026, Mar. 2024.

[11] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 12449–12460.

[12] R. Scheibler, T. Komatsu, Y. Fujita, and M. Hentschel, "Sound event localization and detection with pre-trained audio spectrogram transformer and multichannel separation network," *OMNI (1ch)*, vol. 13, pp. 5–13, Nov. 2022.

[13] L. Xu, L. Wang, S. Bi, H. Liu, and J. Wang, "Semi-supervised sound event detection with pre-trained model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[14] 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, Language Process.*, vol. 31, pp. 1251–1264, 2023.

[15] 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. (ICASSP)*, May 2022, pp. 9196–9200.

[16] S. Niu, J. Du, Q. Wang, L. Chai, H. Wu, Z. Nian, L. Sun, Y. Fang, J. Pan, and C.-H. Lee, "An experimental study on sound event localization and detection under realistic testing conditions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[17] K. G. R. Jacome, F. L. Grijalva, and B. S. Masiero, "Sound events localization and detection using bio-inspired gammatone filters and temporal convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2314–2324, 2023.

[18] Q. Wang, H. Wu, Z. Jing, F. Ma, Y. Fang, Y. Wang, T. Chen, J. Pan, J. Du, and C.-H. Lee, "The USTC-IFLYTEK system for sound event localization and detection of DCASE 2020 challenge," Tech. Rep. DCASE 2020 Challenge, pp. 1–2, 2020. [Online]. Available: https://dcase.community/documents/challenge2020/technical_reports/DCASE2020_Du_110.pdf

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[20] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[21] 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. (ICASSP)*, Jun. 2021, pp. 915–919.

[22] C. Schymura, T. Ochiai, M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, and D. Kolossa, "Exploiting attention-based sequence-to-sequence architectures for sound event localization," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 231–235.

[23] C. Schymura, B. Bönninghoff, T. Ochiai, M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, and D. Kolossa, "PILOT: Introducing transformers for probabilistic sound event localization," in *Proc. Interspeech*, Aug. 2021, pp. 2117–2121.

[24] Y. Shul and J.-W. Choi, "CST-former: Transformer with channel-spectro-temporal attention for sound event localization and detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 8686–8690.

[25] J. F. Gemmeke, D. P. W. 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. (ICASSP)*, Mar. 2017, pp. 776–780.

[26] 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," 2020, *arXiv:2005.08100*.

[27] Z. Chen, S. Chen, Y. Wu, Y. Qian, C. Wang, S. Liu, Y. Qian, and M. Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6147–6151.

[28] L.-W. Chen and A. Rudnicky, "Exploring Wav2vec 2.0 fine tuning for improved speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[29] S. Dieleman, A. van den Oord, and K. Simonyan, "The challenge of realistic music generation: Modelling raw audio at scale," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 3–4.

[30] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 369–376.

[31] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Cham, Switzerland: Springer, 2019.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–7.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.

[34] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: The PyTorch-based audio source separation toolkit for researchers," 2020, *arXiv:2005.04132*.

[35] D. Huang and R. F. Perez, "SSELDNet: A fully end-to-end sample-level framework for sound event localization and detection," in *Proc. DCASE*, 2021, p. 3.

[36] R. Kuprieiev, "DVC: Data version control-git for data & models (2.3. 0)," Iterative, San Francisco, CA, USA, Tech. Rep. 4892897, 2021. [Online]. Available: https://zenodo.org/record/4892897

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[38] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, San Francisco, CA, USA, Tech. Rep. 2018-01, 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[39] 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 31st Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Oct. 2021, pp. 1–6.

[40] 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. (ICASSP)*, May 2022, pp. 9186–9190.

[41] 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, Language Process.*, vol. 30, pp. 829–852, 2022.
[42] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. 22nd ACM Int. Conf. Multimedia (ACM-MM)*, Orlando, FL, USA, Nov. 2014, pp. 1041–1044.
[43] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," 2019, *arXiv:1905.08546*.
[44] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," 2020, *arXiv:2006.01919*.
[45] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Appl. Sci.*, vol. 6, no. 6, p. 162, May 2016.
[46] X. Zhong and W. A. Yost, "How many images are in an auditory scene?" *J. Acoust. Soc. Amer.*, vol. 141, no. 4, pp. 2882–2892, Apr. 2017.

ORLEM LIMA DOS SANTOS (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from the Federal University of Roraima, Roraima, Brazil, in 2017, and the M.Sc. degree in electrical engineering from the School of Electrical and Computer Engineering, University of Campinas (UNICAMP), Campinas, Brazil, in 2021. He is currently pursuing the Ph.D. degree with the University of Campinas. His research interests include spatial audio, deep learning, and self-supervised learning.

KAREN ROSERO (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering and telecommunications from the Army Polytechnic School, Sangolqui, Ecuador, in 2020, and the M.Sc. degree in electrical engineering from the School of Electrical and Computer Engineering, University of Campinas (UNICAMP), Campinas, Brazil, in 2022. She is currently pursuing the Ph.D. degree with The University of Texas at Dallas. Her research interests include spatial audio, deep learning, music information retrieval, affective computing, and multimodal signal processing.

BRUNO MASIERO (Member, IEEE) received the B.S. and M.Sc. degrees in electrical engineering from the University of São Paulo, Brazil, in 2005 and 2007, respectively, and the Ph.D. degree in engineering by the RWTH Aachen University, Germany, in 2012. He is an Assistant Professor with the School of Electrical and Computer Engineering (FEEC), University of Campinas (UNICAMP), Campinas, Brazil. He is currently the Associate Dean of FEEC-UNICAMP. His research focuses on the application of modern digital signal processing techniques in audio and acoustic applications. He served as a member of the Board of the International Commission for Acoustics, from 2019 to 2022.

ROBERTO DE ALENCAR LOTUFO (Member, IEEE) received the B.S. degree in electrical engineering from the Instituto Tecnologico de Aeronautica, Brazil, in 1978, and the Ph.D. degree in electrical engineering from the University of Bristol, U.K., in 1990. He has been with the School of Electrical and Computer Engineering, University of Campinas (UNICAMP), Brazil, since 1981, where he is currently a Full Professor. His principal research interests include image processing and analysis, pattern recognition, and machine learning. He has authored more than 150 refereed international journals and full conference papers. He was awarded the Innovation Personality, in 2008, and the Zeferino Vaz Academic Recognition from UNICAMP, in 2011.