Uncertainty Estimation for Sound Source Localization with Deep Learning

Rendong Pi, *Graduate Student Member, IEEE*, Xiang Yu*

***Abstract*—While significant progress has been made in the field of Sound Source Localization (SSL), the confidence and robustness of the localization results still remain low. Conducting uncertainty analysis can effectively alleviate this problem, since it provides a measure of the confidence level in the SSL results. In this work, we propose a novel framework for SSL that not only delivers the state-of-the-art localization performance, but also provides reliable uncertainty estimations. Our framework leverages a novel backbone architecture integrating a multi-head self-attention module to effectively capture spatial features through a self-attention mechanism. Additionally, our approach incorporates subjective theory to associate predictions obtained from the neural network with a Dirichlet distribution. This allows us to model the overall uncertainty by parameterizing the class probabilities of the positions of the sound source. To comprehensively evaluate the performance of the proposed method, extensive experiments were conducted using both simulated and real-world datasets. The results show that the proposed method can improve the SSL accuracy and enhance the neural network's reliability, even out-of-distribution samples can be handled effectively. The obtained accurate sound source positions and uncertainty estimations can be utilized in downstream audio-related tasks, such as enhancing the accuracy and reliability of sound event detection by incorporating uncertainty. This integration can assist robots in making more informed decisions by fusing information from multiple sources. Our code is available at https://github.com/Devin-Pi/uncertainty-estimation-for-ssl.**

***Index Terms*—moving sound source localization, uncertainty estimation, deep learning, attention mechanism, subjective logic theory**

I. Introduction

Sound Source Localization (SSL) refers to the process of determining the location of sound-emitting objects using either the human auditory system or an audio perception machine. It finds widespread applications in many fields, including robot audition [1], [2], speech enhancement [3], video conferencing, and pipe leak detection [4]. To improve the accuracy of SSL tasks, researchers have developed various approaches, which can be broadly classified into two categories: conventional physics-based and deep learning-based methods. Conventional methods primarily focus on estimating the spatial features of sound sources by analyzing multi-channel sound signals collected from the acoustic environment. These methods typically utilize time differences, sound level differences, and phase differences between signals captured by at least two microphones to determine the location of the sound source. Beamforming and other spatial audio processing techniques have also been developed for multi-channel SSL systems [5], [6], [7]. However, traditional methods often struggle with poor prediction accuracy in complex real-world environments due to the limitations of predefined conditions and the presence of noises and reverberations.

In recent years, substantial research efforts have been made to improve the performance of SSL in complex and challenging acoustic environments. Many of these efforts have utilized neural network-based methods. Specifically, neural network models such as Convolutional Recurrent Neural Network (CRNN) [8], Transformer [9], and Long Short-Term Memory (LSTM) [10] have been employed. These methods can be further classified into classification and regression tasks. Although these methods have shown promise in enhancing the SSL performance, the resulting localization estimates usually lack robustness and reliability. This is because these models typically use SoftMax to output the classification results, which tends to produce over-confident predictions, especially for incorrectly classified results [11], [12], [13], [14]. In typical SSL tasks, while acquiring the position of the sound source is important, it is equally important to assess the reliability and robustness of the localization results, particularly in some applications such as robot audition.

Conducting uncertainty estimations can effectively provide insights into the reliability and robustness of the models. Generally, two types of uncertainty estimation models can be considered: Bayesian and non-Bayesian. Bayesian-based methods describe the uncertainty through a distribution over weights. Various Bayesian methods have been developed, such as Markov Chain, Monte Carlo, Laplacian approximation, and other variants. For instance, Schymure [9] employed a Linear Gaussian System to determine the probability of the sound event localization, with an additional branch to output the posterior mean and covariance matrix. However, the output of this approach does not provide a direct measure for estimating the prediction uncertainty in terms of a comparable value. Besides that, Bayesian-based methods are computationally expensive, necessitating the development of new techniques for uncertainty estimation. Recent developments have focused on non-Bayesian methods, including evidential deep learning [11], deep ensembles, and others. These methods have been successfully applied in various domains such as semantic segmentation [15], action recognition [16], multi-view classification [12], etc.

In summary, existing approaches to SSL tasks, whether based on physics or neural networks, often lack robust uncertainty estimations. To address this challenge, we propose

The authors are with the Department of the Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong, China (e-mail:devin.pi@connect.polyu.hk; lucien.yu@polyu.edu.hk).

Manuscript received xxx, 2024; revised xxx, 2024.

a novel framework in this study, aiming to achieve the state-of-the-art localization performance while simultaneously providing uncertainty estimates for predicted sound source positions. Unlike previous deep learning-based approaches which directly output the localization results, our proposed method incorporates *uncertainty* into the neural network's output. Specifically, we employ the Dempster-Shafer theory [17] and Subject Logic theory [18] to generate final localization probabilities and uncertainties for different Direction of Arrival (DOA) candidates. The distribution of possibilities for different classes of SSL is parameterized using the Dirichlet distribution. Additionally, we introduce new backbones for SSL that utilize the multi-head self-attention mechanism, achieving state-of-the-art performance compared to previous methods. The main contributions of this paper are summarized as follows:

- We propose a novel framework that addresses the critical need for uncertainty estimation in SSL. This framework enables simultaneous prediction of sound source locations and reliable quantification of uncertainty for those predictions.
- Two novel backbones by leveraging the multi-head self-attention mechanism are proposed, namely TCRNN and TLSTM. Compared with previous models, these proposed backbones learn the relationship between the binaural sound signals and DOA in an attention-based manner, effectively emphasizing critical features for accurate SSL.
- Through extensive experiments on simulated datasets and the publicly available LOCATA dataset, we demonstrate the feasibility and generalization ability of our proposed method. The results show that our approach not only achieves the state-of-the-art performance for SSL but also provides valuable uncertainty estimations, enhancing the reliability and robustness of the predicted results.

The rest of the paper is organized as follows: Section II reviews the related work and defines the research gap. Section III details the proposed method. Experimental results and discussions are presented in Section IV. Finally, the conclusions are drawn in Section V.

II. Related Work

In this section, we first review the existing methods related to SSL and uncertainty estimation tasks.

A. Deep Learning-Based SSL

Accurately determining the position of the sound source is crucial for various SSL-related tasks. Many deep learning-based approaches have been proposed to address SSL, as summarized in Table I. Broadly speaking, the task of determining the position of sound sources can be divided into two categories: classification and regression. Classification-based methods aim to predict the DOA of sound sources, which is usually represented by azimuth and elevation angles typically divided into continuous degree candidates ranging from $0^\circ \sim 180^\circ$ [8], [10]. As for the regression-based approaches, they aim to obtain the exact position of sound sources in the coordinate systems such as Cartesian coordinates, spherical coordinates, and cylindrical coordinates [9], [10], [31]. The coordinates of the sound sources can also be determined by the distance between the sound source and the microphone array and the DOA [25]. In addition, there also exist methods that estimate the localization of sound sources by using intermediate features, such as inter-channel phase differences (IPD) [8], [10], [20], [26], ray space transform [22], and spherical maps [31]. Regardless of the specific SSL task, the input features are typically extracted from multi-channel sound recording signals. These input features can include the phase and magnitude spectrum [19] or short-time spatial pseudo-spectrum [23], typical spectrogram involving phase and magnitude [24] and its variations, like logarithmic magnitude and phase spectrogram [8], [28], the real and imaginary parts of the short time Fourier transform (STFT) coefficients [10], [9], [27], and relatively high-level features, such as generalized cross-correlation with phase transform (GCC-PHAT) [21], [22], steered response power with phase transform spectrum (SRP-PHAT) [25], [31], [33], icosahedral SRP-PHAT map [32], circular harmonic features [29], and spherical beamforming map [30]. The commonly used neural network architectures for SSL include convolutional neural network (CNN) [21], [22], [23], [24], [27], [29], convolutional recurrent neural network (CRNN) [19], [8], [28], long short-term memory (LSTM) neural network [10], [26], fully-connected neural network [20], etc [30], [31], [32], [33], [9], [25].

In sum, the existing deep learning-based methods have shown promising results for accurate SSL. However, the reliability and robustness of the SSL results are still limited. Specifically, these approaches lack proper consideration of uncertainty estimation. To bridge this gap, our aim is to propose an end-to-end method that not only delivers accurate localization results, but also provides reliable uncertainty estimations for the predicted results. By incorporating uncertainty estimation into the localization process, we seek to enhance the overall reliability and robustness of SSL networks.

B. Uncertainty Estimation for Deep Learning

During the inference stage, the deep learning neural networks are usually deployed after sufficient training. Hence, their uncertainty can not be obtained directly. To address this issue, Bayesian neural network [34] (BNN) has been developed to estimate the uncertainty of the deep learning-based models by incorporating probabilistic weight parameters. Another method for uncertainty estimation is Monte Carlo dropout [35], which involves conducting dropout sampling from the weight in the training and testing stages. Recently, evidential deep learning (EDL) [11] is developed to estimate the uncertainty by formulating subjective opinions. EDL can directly model the uncertainty without relying on weight sampling or replacement. By incorporating subjective logic theory (SL) [18] and Dempster-Shafer theory (DST) [17] into the neural network, EDL can provide localization results and uncertainty estimation at the same time. A few studies have demonstrated that EDL has widespread applications, including semantic segmentation [15], action recognition [16],

TABLE I
SUMMARY OF DEEP LEARNING-BASED SSL METHODS

Approach	Input	Output	Network	Uncertainty
SELDNet [19]	phase and magnitude spectrum	DOA	CRNN	No
[20]	sinusoidal IPD	sinusoidal IPD	FC	No
[21]	GCC-PHAT and periodicity degree features	DOA	CNN	No
[22]	GCC-PHAT	Ray Space Transform	CNN	No
[23]	short-time spatial pseudo-sepctrum	DOA	CNN	No
[24]	phase and magnitude spectrograms	DOA	CNN	No
Cross3D [25]	SRP-PHAT	DOA	3D CNN	No
PILOT [9]	STFT coefficients	DOA	Transformer	Yes
[26]	logarithmic power spectra & sinusoidal IPD	time-frequency masks; DOA	BLSTM; LSTM	No
[27]	STFT coefficients	DOA	CNN	No
SRP-DNN [8]	logarithm-magnitude and phase spectrograms	DP-IPD	CRNN	No
SALSA [28]	log-linear spectrograms & normalized eigenvectors	DOA	CRNN	No
[29]	circular harmonic features	DOA	CNN	No
[30]	spherical beamforming map	spherical target map	spherical autoencoder	No
Spherical CRNN [31]	SRP-PHAT	DOA	spherical CRNN	No
[32]	icosahedral SRP-PHAT map	DOA	Icosahedral CNN	No
FN-SSL [10]	STFT coefficients	DP-IPD	(B)LSTM	No
IFAN [33]	SRP-PHAT and SRP-LMS	DOA	Icosahedral CNN	No
TCRNN	STFT coefficients	DOA	CRNN	Yes
TLSTM	STFT coefficients	DOA	LSTM	Yes

multi-view classification [12], etc. However, to the best of our knowledge, the application of EDL theory has never been explored in the context of SSL tasks. This study seeks to address this research gap and comprehensively evaluate the advantages and effectiveness of incorporating EDL with SSL, thereby contributing to a deeper understanding of the capabilities and potentials of this novel approach.

III. METHOD

This section presents an evidence-based neural network for SSL that incorporates extra uncertainty estimation, distinguishing it from previous networks. The theory and fundamentals related to the uncertainty and evidence are introduced first. Then, a novel framework for processing the multi-channel sound recordings and obtaining the DOA of the sound source is elaborated, along with the estimation of uncertainties associated with the predictions of DOA classifications. Finally, the loss function of the training model is established.

A. Uncertainty and Evidence Theory

In this subsection, the fundamentals of EDL [11] are elaborated.DST [17], as one of the generalizations of the Bayesian theory in the perspective of Subjective probabilities, can represent the probabilities of each possible states in a discriminative framework by assigning belief masses. To quantify the belief masses and uncertainty in a discriminative framework, SL [18] is utilized to associate the belief masses and uncertainty with the parameters of Dirichlet distribution. In this way, the overall uncertainty for the current classification and the probabilities of each class can be modeled jointly. Specifically, in a K-class classification problem, the SL assigns belief masses b_k to each class and the overall uncertainty mass u to the whole classification. It should be noted that these $k+1$ values are all non-negative and the sum among them should be one:

$$u + \sum_{k=1}^{K} b_k = 1, \tag{1}$$

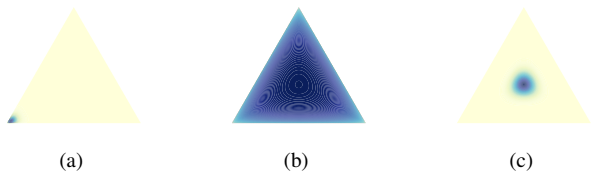
where b_k denotes the belief mass for the k_{th} class and u denotes the overall uncertainty of the whole framework, respectively. The belief mass b_k for a class k can be obtained by the $evidence$, which can be regarded as a kind of support for the specific classification results of a sample. Furthermore, the SL [18] associates the evidence $\boldsymbol{e} = [e_1, ..., e_K]$ with the concentration parameters of the Dirichlet distribution $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_K]$. Specifically, the relationship between the α and e was denoted as $\alpha_k = e_k + 1$. Hence, the belief mass b_k and overall uncertainty u can be acquired by:

$$b_k = \frac{e_k}{S} = \frac{\alpha_k - 1}{S}, u = \frac{K}{S}, \tag{2}$$

where $S = \sum_{i=1}^{K}(e_k + 1) = \sum_{i=1}^{K} \alpha_k$ represents Dirichlet strength. According to Eq. (2), it can be observed that higher evidence obtained for the k_{th} class leads to a higher belief assigned to the k_{th} class. In contrast, the less evidence acquired, the higher overall uncertainty for the classification. This belief assignment can be regarded as a subjective opinion. The probability of the k_{th} class is the mean of the Dirichlet distribution, which can be obtained by $\hat{p}_k = \frac{\alpha_k}{S}$.

To further clarify the significance of the Dirichlet distribution in the classification tasks, we take a triple classification problem as an example. In this case, the Dirichlet distribution can be represented by a triangle, i.e., a standard 2-simplex, where the vertices of the triangle denote the different classes. As shown in Fig. 1(a), when we obtain high evidence for a specific class, e.g., the evidence among 3 classes is $\boldsymbol{e} = \langle 50, 1, 1 \rangle$ and the corresponding Dirichlet distribution concentration parameters is $\boldsymbol{\alpha} = \langle 51, 2, 2 \rangle$, a sharp distribution is described on the top of the triangle. It represents that enough evidence is obtained to ensure accurate classification results.

In contrast, if we acquire little evidence for the classification, e.g., $\boldsymbol{e} = \langle 0.1, 0, 1, 0.1 \rangle$. Correspondingly, the related Dirichlet distribution parameters is $\boldsymbol{\alpha} = \langle 1.1, 1.1, 1.1 \rangle$, the uncertainty mass for this case is around 0.91. As shown in



(a) (b) (c)

Fig. 1. Dirichlet distribution visualization. (a) high confidence. (b) low confidence. (c) out of distribution.

Fig. 1(b), This indicates that the overall uncertainty is high and we cannot have enough confidence on the classification results. Additionally, when the evidence is equal among different classes, e.g., $\boldsymbol{e} = \langle 16, 16, 16 \rangle$, and the Dirichlet distribution parameters is $\boldsymbol{\alpha} = \langle 17, 17, 17 \rangle$. In this case, the Dirichlet distribution is uniform, as shown in Fig. 1(c). We can also obtain a high uncertainty mass for this case.

B. Framework of Uncertainty Estimation for SSL

In this subsection, the proposed framework for uncertainty estimation of SSL is introduced, as shown in Fig. 2 to Fig. 4.

Overview of the proposed framework: As shown in Fig. 2, the proposed architecture contains four parts: (1) Input: We utilize multi-channel sound signals as the input for the entire framework. (2) Backbone: This component is employed for feature extraction. (3) Uncertainty Estimation: This is calculated based on the output from the backbone. (4) Output: The final output consists of two parts: the predicted DOA for each time frame and the corresponding uncertainty estimation results for the predictions.

To study the feasibility and generality of the Uncertainty Estimation method as much as possible, inspired by SRP-DNN [8] and FN-SSL [10], two of the most commonly used neural networks, the CRNN and LSTM architectures were selected as the backbones to conduct feature extractions. Different from the original models, to further enhance the ability of the model and capture the relationship between the multi channels in terms of frequency, the multi-head self attention module is introduced. The improved CRNN and LSTM-based neural networks are proposed, called Trusted CRNN (**TCRNN**) and Trusted LSTM (**TLSTM**). Next, each part of the proposed framework is introduced in detail.

Input: In this study, we consider binaural sound recordings as the input data. These recordings are first processed by the Short-time Fourier Transform (STFT) before being fed into the backbone. The input for the neural network consists of the concatenated real and imaginary parts of the STFT results.

Backbones: As mentioned previously, two types of backbones are proposed, based on CRNN and LSTM, respectively. For the TCRNN, depicted in the upper part of Fig. 3, four Convolutional Modules (CMs) are utilized to perform initial feature extraction. Each CM consists of two convolutional layers, each followed by a Rectified Linear Unit (ReLU) and a Batch Normalization (BN) layer. After being processed by the CMs, these extracted features are fed into a multi-head self-attention module to further learn the location-related features in the view of frequency. Subsequently, the enhanced features are fed into to the LSTM module to capture temporal features. The learning objective of this study is the DOA at individual time frames. Therefore, transformations of time dimensions are necessary. Max pooling, selected for time dimension compression, is applied after each CM. The final output for the DOA is obtained through a fully-connected layer. Unlike TCRNN, the TLSTM is solely constructed using the LSTM modules. Specifically, two variants of the LSTM, termed Full-band and Narrow-band LSTM, are utilized in this study [10]. These variants can effectively extract the features from different frequencies. The input for the TLSTM is identical to that of the TCRNN. Additionally, to further improve the capability of the TLSTM to learn frequency-related features, a multi-head self-attention module is employed subsequent to the LSTM modules for further feature extraction.

Uncertainty Estimation: After the processing by the neural network backbones, uncertainty estimation is carried out. In this study, we employ an activation function layer, such as ReLU, to process the output of the neural network to yield non-negative values, which are interpreted as evidence. Subsequently, employing the Subjective Logic framework previously mentioned, we construct the belief mass and perform uncertainty calculations. Specifically, the neural network output is first transformed into evidence. Then, the possibility and uncertainty of the predicted DOAs for various DOA candidates are calculated using Eq. (1) and Eq. (2). This approach enables the proposed method to endow the model with the additional capability of calculating uncertainty for different candidate DOAs.

Output: As shown in the latter part of Fig. 3, the proposed framework not only delivers the predicted DOA at each time frame but also provides the uncertainty estimation values for the corresponding predictions.

As shown in Fig. 3, the multi-head self-attention mechanism is employed in both backbones to enhance the ability of the feature extraction for the models. The structure of the multi-head self-attention module is shown in Fig. 4. As described in [36], we can obtain the attention map by conducting matrix multiplication for the DOA feature map itself. Then the Softmax is employed to conduct normalization to the output in order to generate the attention map. Based on the attention map, the critical information related to the DOA can be extracted. Besides that, to prevent information loss during data processing, a residual connection is utilized. This connection supplements the self-attention module's calculation results with the original data. Multi-head refers to performing these operations multiple times in parallel. This allows the neural networks to attend to different parts of the sequence differently. In sum, the data operation in multi-head self-attention module can be represented by Eq. (3), where x denotes the audio data feature map and $\bigotimes$ denotes the matrix multiplication:

$$Output = x + x \otimes \textit{softmax}(x \otimes x). \tag{3}$$

Based on the descriptions of the data flow mentioned above, the shape of input for the multi-head self-attention module is $B \times T \times 2C * F$, where B denotes the batch size, T means the time frames, C means the number of channels of the audio

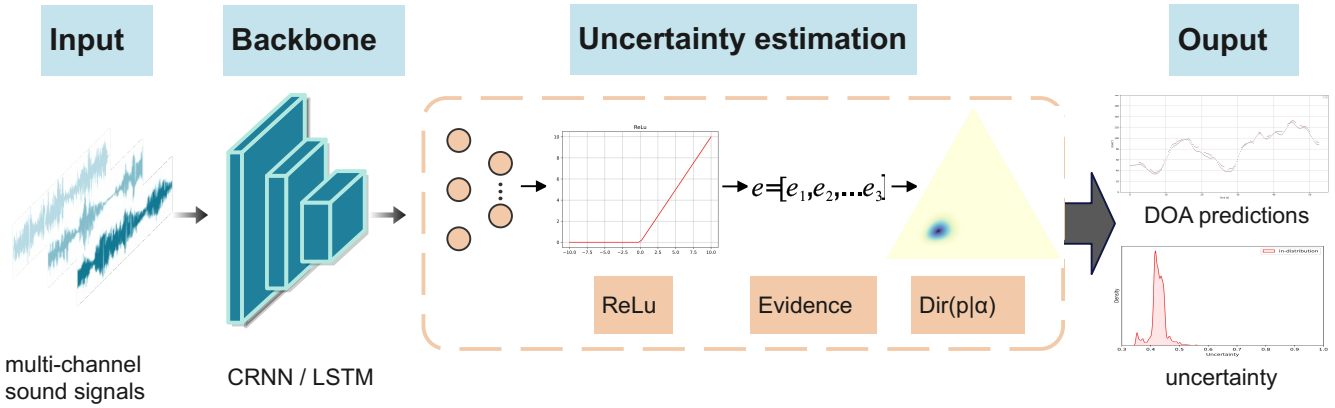


Fig. 2. The paradigm for uncertainty estimation of SSL.

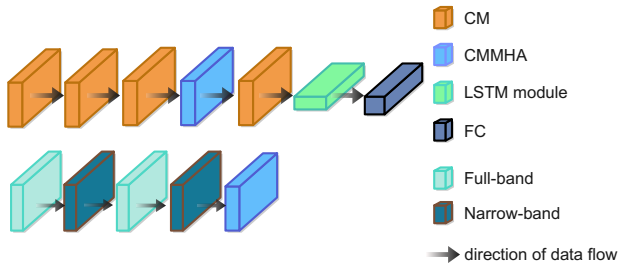


Fig. 3. The architectures of the TCRNN and TLSTM.

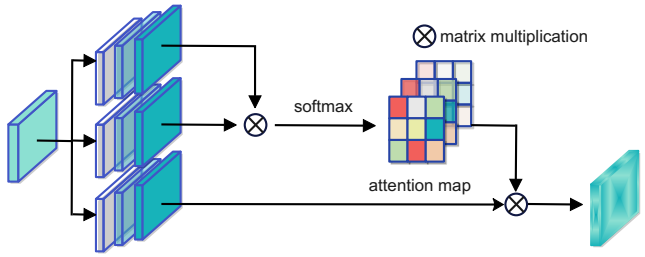


Fig. 4. The structure of the multi-head self attention module.

signals, and F represents the number of frequency bins. The frequency-related features can be further extracted at each time frame. In this way, the obtained attention map can reveal the relationship between the frequency bins and the DOAs. With continued training, the distinctions between these frequency bins become increasingly discernible.

C. Loss Function

In this subsection, the loss function is clarified, which is utilized to train the neural network to learn to form opinions. The loss function in conventional neural network can be denoted as:

$$\mathcal{L}_{ce}(\Theta) = -\sum_{j=1}^{K} y_{ij} \log(p_{ij}), \tag{4}$$

where p_{ij} represents the probability of the sample i for the class j and y_{ij} is a one-hot vector representing the ground truth class of the observation. In this work, the evidence of the i_{th} sample is obtained by the neural network and the concentration parameters of the Dirichlet distribution (i.e., $\alpha_i = e_i + 1$) can be calculated. Then the corresponding multinomial opinions $D(\boldsymbol{p}_i|\boldsymbol{\alpha}_i)$ are formed, where p_i means the class probabilities on a simplex. By conducting the corresponding adjustment in Eq. (4), the modified loss function is denoted as:

$$\begin{aligned}\mathcal{L}_{uce}(\Theta) &= \int \left[\sum_{j=1}^{K} -y_{ij}\log(p_{ij})\right] \frac{1}{B(\boldsymbol{\alpha}_i)} \prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1} d\boldsymbol{p}_i \\ &= \sum_{j=1}^{K} y_{ij}\left(\psi(S_i) - \psi(\alpha_{ij})\right),\end{aligned} \tag{5}$$

where $\psi()$ is the $digamma$ function. It should be noted that the above equation is employed to ensure that the correct predictions contributed to more evidence. However, in this work, the KL divergence is leveraged to ensure that incorrect predictions would generate less evidence:

$$\begin{aligned}\mathcal{L}_{KL}(\Theta) &= KL[D(\boldsymbol{p}_i|\hat{\boldsymbol{\alpha}}_i) \,||\, D(\boldsymbol{p}_i|\mathbf{1})] \\ &= \log\left(\frac{\Gamma(\sum_{k=1}^{K}\hat{\alpha}_{ik})}{\Gamma(K)\prod_{k=1}^{K}\Gamma(\hat{\alpha}_{ik})}\right) \\ &+ \sum_{k=1}^{K}(\hat{\alpha}_{ik}-1)\left[\psi(\hat{\alpha}_{ik}) - \psi\Big(\sum_{j=1}^{K}\hat{\alpha}_{ij}\Big)\right],\end{aligned} \tag{6}$$

where $\Gamma()$ represents $gamma$ function; $\hat{\alpha} = y_i + (1-y_i)\odot\alpha$ is utilized to avoid penalizing the evidence of the ground truth of the class to 0. It can be viewed as the adjusted parameter of the Dirichlet distribution. Hence, the overall loss function can be defined as:

$$\mathcal{L}_i(\Theta) = \mathcal{L}_{ue} + \lambda_t \mathcal{L}_{KL}, \tag{7}$$

where λ_t represents the balance factor, which is gradually increased to prevent the neural network from overemphasizing the KL loss in the early training stage [12].

IV. Experimental Results and Discussions

A. Datasets and Implementation Details

In this work, two datasets are employed to train and evaluate the performance of the proposed method. Similar to [25]

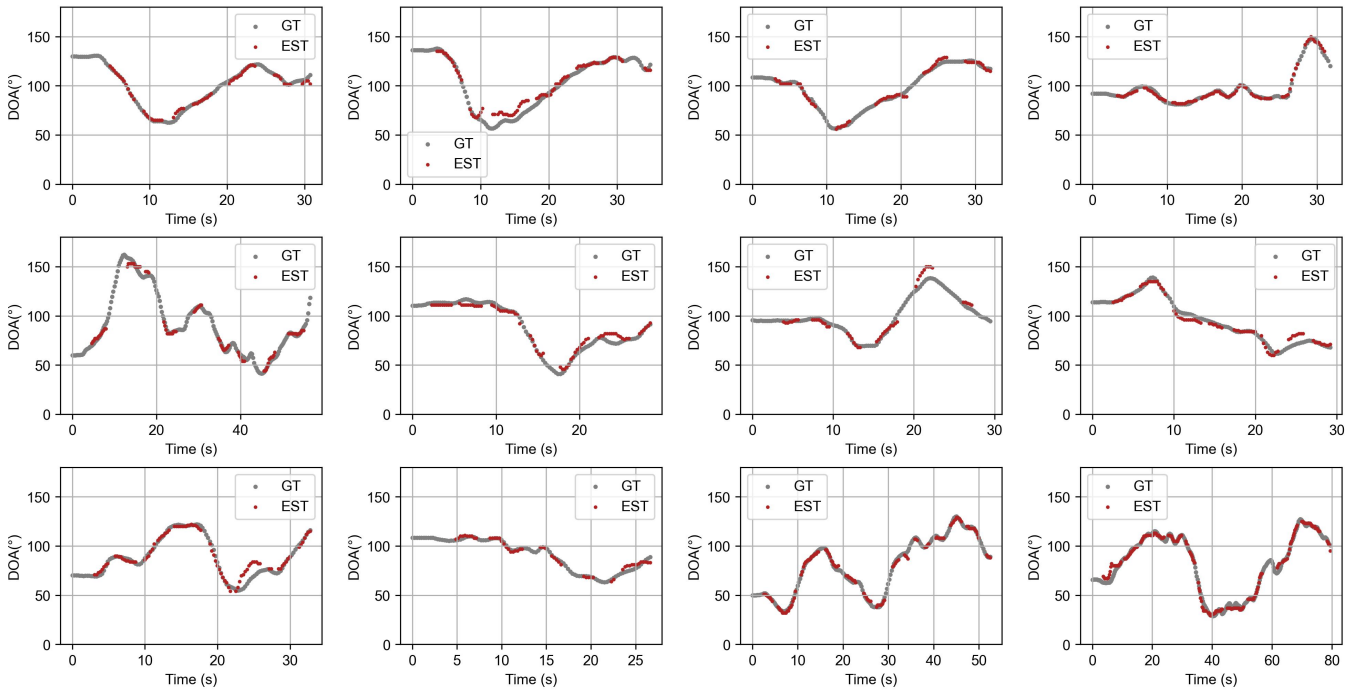


Fig. 5. DOA(trajectory) estimation for the LOCATA datasets.

and [10], the LibriSpeech corpus dataset [37] is randomly selected to represent the sound source signals. The NOISEX-92 dataset [38], containing white, babble, and factory noises, etc, is employed as the noise source signals. To generate a diffuse sound field, the method described in [39] is used. The simulated dataset is created following the approaches outlined in [25], [32]. Specifically, the room size is randomly set within the range of 6×6×1.5 m to 10×8×6 m. The corresponding reverberation time (RT60) is assigned randomly between 0.2 s and 1.3 s. To synthesize sound signals by mixing the noise and clean source signals together, the Signal-to-Noise Ratio (SNR) selection is crucial. The SNR is defined as the ratio of the power of the signal to the power of the noise, which is expressed as,

$$\text{SNR} = 10 \times \log_{10}\left(\frac{P_{\text{signal}}}{P_{\text{noise}}}\right), \tag{8}$$

where P_{signal} and P_{noise} are the power of the signal and noise, respectively. It should be noted that the unit of SNR is dB. In this paper, the SNR is randomly chosen from -5 dB and 15 dB. To receive the sound signals, two microphones are defined in the acoustic field with a distance of 8 cm between them. The pair of microphones is then randomly positioned in the acoustic field in different samples. The dataset consists of 10,000 samples for training, 998 samples for validation, and 5,000 samples for testing.

As for the real-world dataset, the LOCATA dataset is used [40]. Similar to [10], tasks 3 and 5 are selected to evaluate the effectiveness of our proposed method. In this work, only the sound recordings with azimuth angles in the range of 0° and 180° are used. Note that the simulated dataset was utilized to train the model, while the real-world dataset was solely used to evaluate the neural network performance.

B. Baseline Methods

In this paper, three relevant methods were also implemented to conduct a cross-comparison, which include: (1) SELDNet: this neural network serves as the baseline model for the DCASE22 challenge [19]. It is designed to output both the classification and the DOA of sound events at the same time. In this work, we only use the DOA branch of the SELDNet to generate predictions. (2) SRP-DNN: it is a casual convolutional recurrent neural network aiming to solve the multiple moving SSL problems [8]. (3) FN-SSL: this neural network processes the direct-path inter-phase difference (DP-IPD) of multi-channel sound recordings using narrow-band extraction and full-band correlation techniques [10].

C. Performance Metrics

Consistent with previous studies [8], [10], we employ the same indicators to evaluate model performance. Localization Accuracy (ACC) quantifies the percentage of time frames where the localization error is below the predefined thresholds of 5°, 10°, and 15°. For example, a sample is considered accurate if the error between the predicted and actual DOA is less than the 5° threshold, and ACC(5°) represents the percentage of such accurate predictions within the test dataset. Mean Absolute Error (MAE) measures the absolute error between the predicted and the corresponding ground truth DOA, directly highlighting the deviation from the ground truth.

D. Experimental Results

Results on simulated data are presented in Table II. The results indicate that the proposed method, TCRNN, achieves

TABLE II
RESULTS ON SIMULATED DATASET

Baseline	ACC(5°)[%]	ACC(10°)[%]	MAE[°]
SELDNet [19]	20.72	33.89	31.7
SRP-DNN [8]	71.82	90.85	4.4
FN-SSL [10]	86.33	96.98	2.8
TCRNN	**91.23**	**98.49**	**2.2**
TLSTM	85.26	97.09	2.8

TABLE III
RESULTS ON LOCATA DATASET

Baseline	ACC(15°)[%]	ACC(10°)[%]	MAE[°]
SELDNet [19]	37.19	25.80	28.9
SRP-DNN [8]	94.13	84.08	6.6
FN-SSL [10]	95.51	91.68	5.1
TCRNN	**97.15**	**92.85**	**4.7**
TLSTM	94.30	91.33	5.1

the highest performance among various SSL methods in terms of both ACC and MAE. The proposed TLSTM model performs equally well as the FN-SSL method, even though TLSTM contains less blocks. When compared to CRNN-based methods like SRP-DNN and LSTM-based methods like FN-SSL, the proposed method demonstrates superior ability in revealing the relationship between channels and frequency due to the introduction of the self-attention module, resulting in more accurate DOA estimation. Overall, the inclusion of uncertainty computation and the self-attention module enhances the model's performance without any detrimental effects.

Results on real-world datasets are shown in Table III. As mentioned in the Section *Datasets and Implementation Details*, these models were initially trained using the simulated dataset and subsequently evaluated using the LOCATA dataset. The acoustic conditions in the LOCATA dataset are nearly noise-free, with a reverberation time of around 0.55 s. Hence, all the aforementioned methods can yield satisfactory localization results, as shown in Table III. Notably, the proposed TCRNN approach outperforms the other compared methods. The TLSTM model performs similarly to the FN-SSL model, but with less parameters. These results indicate that the proposed methods exhibit superior performance when applied to real-world datasets, which is consistent with the outcomes observed in the simulated dataset. Note that different from the [10], in this paper, the 4° is not reduced for all comparison methods. To qualitatively assess the performance of the proposed method, the predicted DOA(trajectory) is shown in Fig. 3, which is obtained using TCRNN. The Fig. 5 clearly demonstrates that the proposed method achieves accurate localization across various scenarios.

Uncertainty estimation by using the proposed methods is illustrated in Fig. 6. In this study, the distribution of in-distribution and out-of-distribution samples is visualized based on their uncertainty. The in-distribution data is directly from the original simulated test dataset and LOCATA dataset, while the out-of-distribution data is obtained by adding Gaussian noise with specific SNRs. To comprehensively evaluate the effects of different noise levels on the uncertainty estimation, the SNR values of -5 dB, -10 dB, and -15 dB are selected to generate the out-of-distribution datasets.

It should be noted that, we use the value between 0 and 1 to describe the uncertainty. To evaluate the ability of the uncertainty estimation for the proposed methods, we visualize the distribution of in-/out-of distribution samples in terms of their uncertainty values, similar to previous studies [12], [13].

The corresponding uncertainty estimation results are presented in Fig. 6 and Fig. 7. Based on these results, the following observations can be made: (1) In comparison to the in-distribution samples, higher uncertainties are estimated for the out-of-distribution samples with lower SNRs. This implies that the proposed method can effectively capture the changes in the data through uncertainty estimation when it is contaminated by noise. (2) As the SNR decreases, which means the level of noise increases, the uncertainty of the localization results also increases. This trend is evident in Figs. 6(b)-6(d), where the peak value of the uncertainty density becomes higher as the SNR decreases. The uncertainty estimation is also more concentrated in the high uncertainty area, as depicted in Fig. 6, where the entire uncertainty estimation area shifts to the right. The two proposed models, both TCRNN and TLSTM show the same trend. (3) Regarding the performance indicators shown in Fig. 7, as the SNR decreases, the ACC of the proposed method decreases, while the MAE increases. Note that as the noise levels increase, the TLSTM fails to detect significant changes in the data through the performance indicators. Both the ACC and MAE exhibit slight changes. This is because there exists a threshold for the performance indicators to capture the change in the data. Beyond that threshold, the performance indicators can hardly capture the change in the data. However, uncertainty estimation can effectively detect the changes in the data, as shown in Fig. 6. In summary, these findings demonstrate that the effectiveness of our proposed model in estimating uncertainty, as it properly captures increased uncertainty in the presence of lower SNR levels. Furthermore, both the proposed CRNN-based and LSTM-based methods, TCRNN and TLSTM, exhibit consistent trends in uncertainty estimation, providing strong evidence for the feasibility and generalization ability of the proposed uncertainty estimation method.

Although uncertainty estimation is more easily achieved when significant differences exist between data samples, such as in the case of in-distribution and out-of-distribution samples mentioned earlier, it is important to further investigate the sensitivity of uncertainty estimation for the proposed method. To this end, the uncertainties under five different SNR levels (-1 dB to -5 dB) are estimated, and the corresponding results are shown in Fig. 8. Upon closer inspection, it can be observed that while there exists a slight variation in the uncertainty estimation for different SNR levels, the proposed method remains effective in capturing on both simulated and real-world datasets. Take Fig. 8(c) as an example, we can observe that the peak value of uncertainty estimation is lower when the SNR is higher. The highest peak value is attained at a SNR of -5 dB, while the lowest peak value is obtained at a SNR of -1 dB. Additionally, with lower SNR levels, the uncertainty estimation area tends to shift to the right. Similar trends are

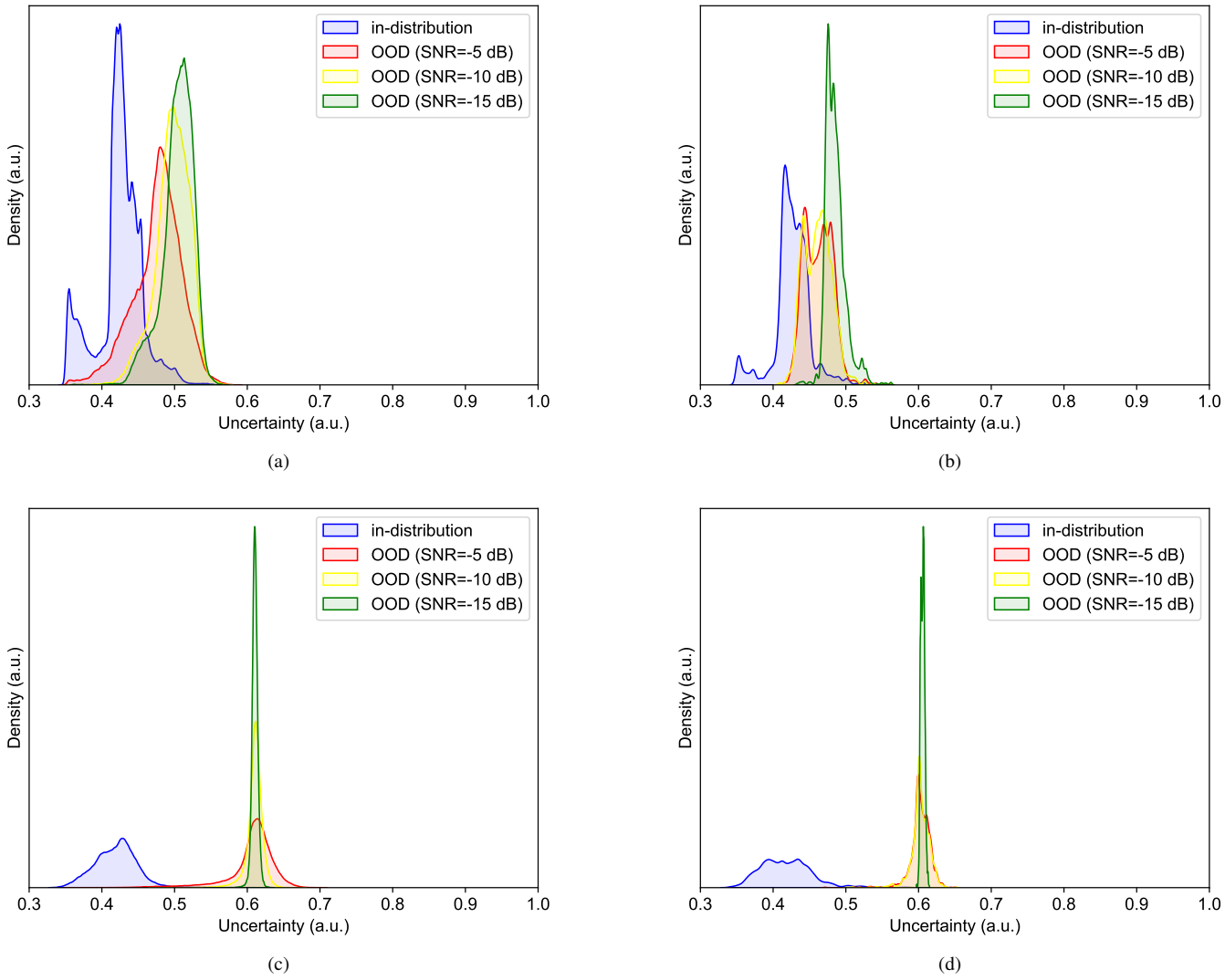


Fig. 6. Density of uncertainty estimation for the test and LOCATA datasets. (a) TCRNN test dataset. (b) TCRNN LOCATA dataset. (c) TLSTM test dataset. (d) TLSTM LOCATA dataset.

observed in the uncertainty estimation results obtained by TCRNN and TLSTM, respectively.

E. Computation Complexity Analysis

Although model performance is crucial, model complexity also significantly impacts the efficiency of SSL, particularly in computation-limited scenarios. In this study, three indicators are utilized to analyze the computational complexity of the proposed methods and other SSL models: (1) Parameters, which represent the total number of weights and biases involved in optimizing model performance; (2) FLOPs, or Floating Point Operations Per Second; and (3) Inference time, defined here as the total time required to infer 5,000 samples. The corresponding results are summarized in Table IV. From this table, it is evident that SELDNet [19] has the lowest computational complexity, while FN-SSL [10] exhibits the highest complexity in terms of Parameters and FLOPs. Regarding the proposed TCRNN, although it demonstrates higher computational complexity compared to SRP-DNN [8] and SELDNet [19], these two models significantly underperform in model performance relative to TCRNN, as indicated in Table II and III. Additionally, the inference time for TCRNN is comparable to that of SELDNet [19] and SPR-DNN [8] but is substantially faster than that of FN-SSL [10]. Overall, the proposed method achieves a well-balanced trade-off between complexity and model performance. The complexity analysis further underscores the superiority of the proposed architecture.

It should be noted that the results mentioned above are based on the two-channel sound signals. To investigate the potential impact of different microphone arrays on the experimental results, we conducted a series of experiments using various microphone arrays. The experimental results are discussed in the Appendix section.

V. Conclusion

In this work, we introduce a novel framework for SSL that incorporates SL and DST to address the gap in un-

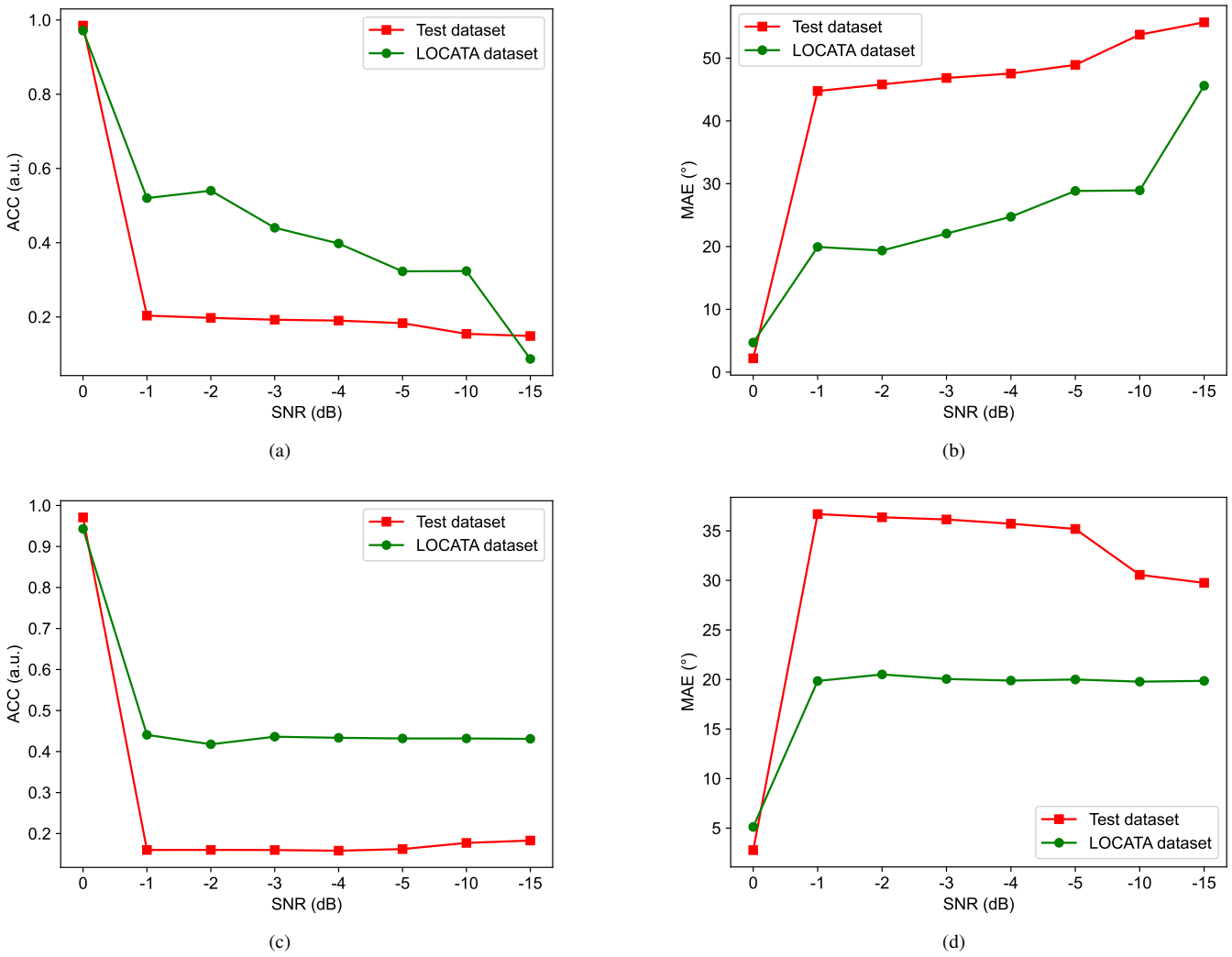


Fig. 7. Performance comparison on different levels of noise. (a) The performance of TCRNN on ACC. (b) The performance of TCRNN on MAE. (c) The performance of TLSTM on ACC. (d) The performance of TLSTM on MAE.

TABLE IV
COMPLEXITY ANALYSIS OF SOUND SOURCE LOCALIZATION MODELS

Baseline	Parameters(M)	Flops(G)	Inference times(s)
SELDNet [19]	0.015	0.196	149
SRP-DNN [8]	0.77	5.43	136
FN-SSL [10]	2.51	806.77	769
TCRNN	1.45	22.83	161
TLSTM	1.85	517.77	690

certainty estimation. Additionally, we enhance SSL performance by proposing new backbones that integrate multi-head self-attention mechanisms. The performance of the proposed method is extensively evaluated on both simulated and real-world datasets. The experimental results demonstrate that the proposed method, TCRNN, outperforms other existing methods in terms of SSL accuracy. The effectiveness of the proposed method in uncertainty estimation is investigated through experiments on both the simulated test and LOCATA dataset with their noise-added version of these two datasets, respectively. The results indicate that the proposed method captures the uncertainty well, with a higher uncertainty associated with lower Signal-to-Noise Ratios (SNRs). Future studies will focus on audio-related downstream tasks, integrating uncertainties into relevant applications, and analyzing their benefits. These include the fusion of various sources, such as visual and audio, or the decision-making processes of auditory robots.

ACKNOWLEDGMENTS

The authors acknowledge the funding support from the Research Institute for Intelligent Wearable Systems (RI-IWEAR) and the Research Institute for Artificial Intelligence of Things (RIAIoT) of the Hong Kong Polytechnic University.

APPENDIX

To investigate the potential impact of different microphone arrays on the experimental results, we conducted a series of experiments using three distinct types of microphone arrays to collect multi-channel sound signals. The specific types of microphone arrays employed are shown in Fig. A.1. The

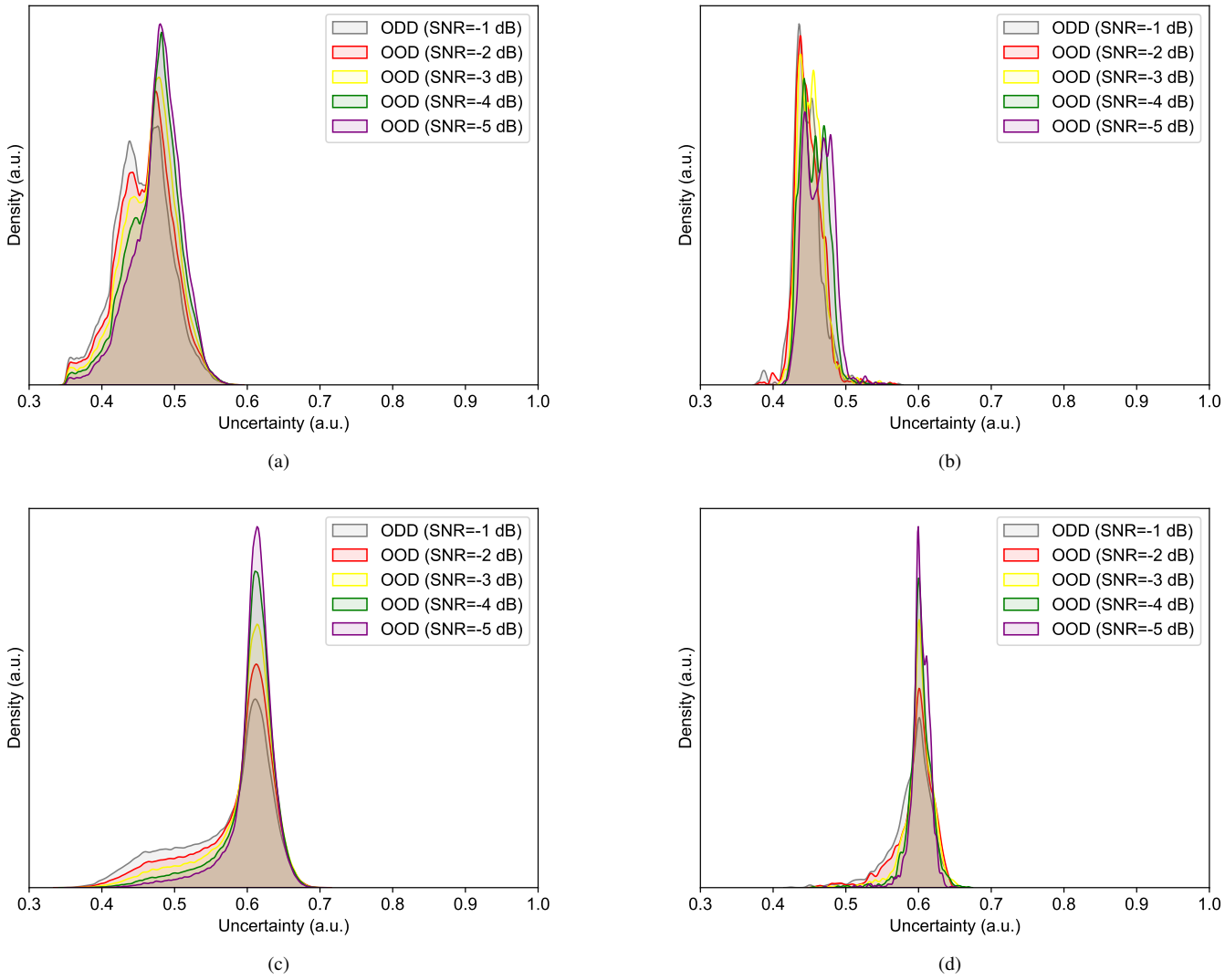
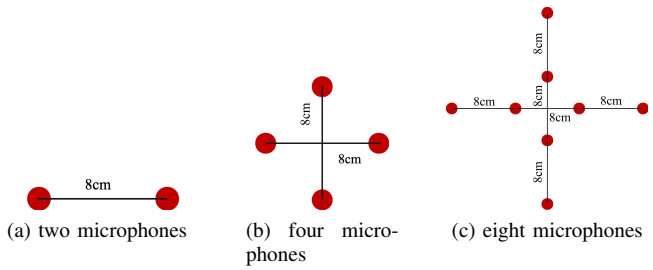


Fig. 8. Sensitive of uncertainty estimation for the test and LOCATA datasets. (a) TCRNN on test dataset. (b) TCRNN on LOCATA dataset. (c) TLSTM on test dataset. (d) TLSTM on LOCATA dataset.

impact of these arrays was evaluated in two key aspects: (1) the performance of the SSL and (2) the uncertainty estimation for the SSL. It is important to note that these experimental results were obtained using the simulated test dataset.



(a) two microphones (b) four microphones (c) eight microphones

Fig. A.1. The types of the microphone arrays.

(1) The performance of the SSL

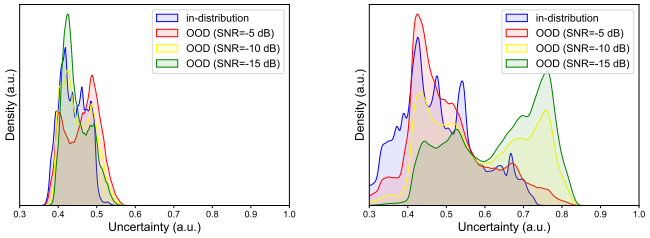
Regarding the localization accuracy of the sound source, we conducted experiments using various microphone arrays. The results, presented in Table A.1, indicate that the performance of the proposed method is not significantly influenced by the choice of microphone arrays. The proposed methods consistently achieve high accuracy in SSL tasks across different microphone arrays.

TABLE A.1
THE PERFORMANCE OF THE PROPOSED METHOD WITH DIFFERENT MICROPHONE ARRAYS

Microphone array type	ACC(5°)[%]	ACC(10°)[%]	MAE[°]
2MIC	91.23	98.49	2.2
4MIC	89.92	96.38	3.35
8MIC	90.46	97.96	2.72

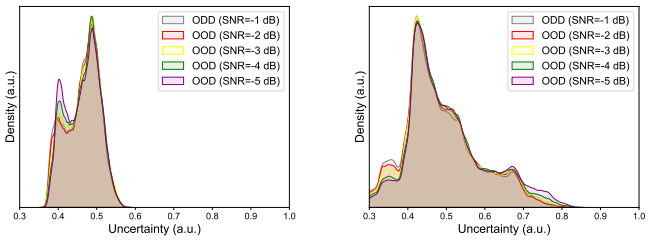
(2) The uncertianty estimation for the SSL

The distribution of uncertainty estimation values for in-distribution and out-of-distribution samples is visualized in Fig. A.2. The results demonstrate a consistent pattern where the peak value of uncertainty estimation increases as the SNR



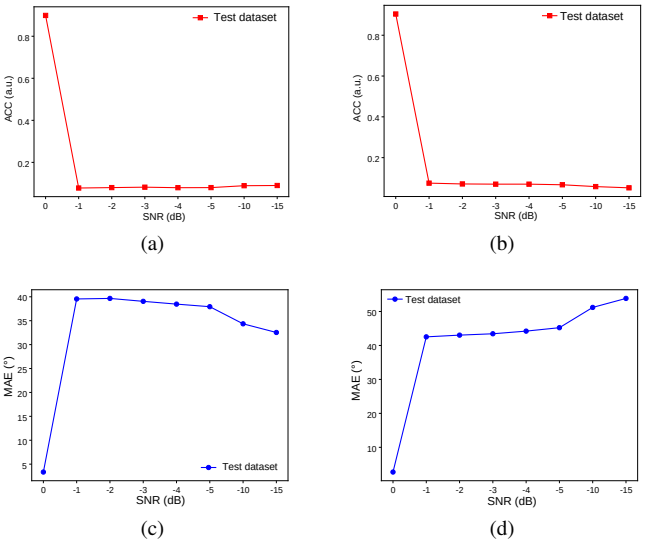
(a) (b)

Fig. A.2. Density of uncertainty estimation for the test dataset. (a) uncertainty of four microphones. (b) uncertainty of eight microphones.



(a) (b)

Fig. A.3. Sensitivity of uncertainty estimation for the test dataset. (a) sensitivity of four microphones. (b) sensitivity of eight microphones.



(a) (b) (c) (d)

Fig. A.4. The performance of the proposed method with different microphone arrays. (a) ACC of four microphones. (b) ACC of eight microphones. (c) MAE of four microphones. (d) MAE of eight microphones.

decreases, compared to the in-distribution data. Additionally, the uncertainty density becomes more concentrated in areas of high uncertainty as the SNR decreases.

As for the sensitivity of the uncertainty estimation, we obtained similar conclusions to those derived from experiments using two-microphone arrays. The proposed method is capable of distinguishing subtle differences between out-of-distribution data collected with different microphone arrays.

Additionally, Fig. A.4 illustrates the changes in the performance of the proposed method across different SNRs. The results indicate that the performance of the model deteriorates as the SNR decreases, regardless of the microphone arrays used.

In summary, the experimental results discussed above indicate consistent conclusions regarding SSL accuracy and uncertainty estimation across the three types of microphone arrays tested. It can be concluded that the type of microphone array does not significantly affect the performance of the proposed method.

REFERENCES

[1] I. An, Y. Kwon, and S.-e. Yoon, "Diffraction- and reflection-aware multiple sound source localization," *IEEE Trans. Robot.*, vol. 38, no. 3, pp. 1925–1944, 2022.

[2] H. W. Löllmann, A. Moore, P. A. Naylor, B. Rafaely, R. Horaud, A. Mazel, and W. Kellermann, "Microphone array signal processing for robot audition," in *Hands-free Speech Commun. and Microphone Arrays*, 2017, pp. 51–55.

[3] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2351–2364, 2023.

[4] 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. and Signal Process.*, vol. 189, p. 110067, 2023.

[5] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[7] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ild and itd," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 18, no. 1, pp. 68–77, 2010.

[8] B. Yang, H. Liu, and X. Li, "Srp-dnn: Learning direct-path phase difference for multiple moving sound source localization," in *IEEE Int. Conf. Acouts., Speech, Signal Process.*, 2022, pp. 721–725.

[9] 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*, 2021, pp. 2117–2120.

[10] Y. Wang, B. Yang, and X. Li, "Fn-ssl: Full-band and narrow-band fusion for sound source localization," in *Proc. INTERSPEECH*, 2023, pp. 3779–3783.

[11] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Advances in Neural Inf. Process. Syst.*, 2018, pp. 3183–3193.

[12] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2551–2566, 2023.

[13] M. Wang, T. Lin, L. Wang, A. Lin, K. Zou, X. Xu, Y. Zhou, Y. Peng, Q. Meng, Y. Qian *et al.*, "Uncertainty-inspired open set learning for retinal anomaly identification," *Nature Commun.*, vol. 14, no. 1, p. 6757, 2023.

[14] K. Zou, X. Yuan, X. Shen, M. Wang, and H. Fu, "Tbrats: Trusted brain tumor segmentation," in *Int. Conf. on Med. Image Comput. and Comput.-Assisted Intervention*, 2022, pp. 503–513.

[15] Q. Wang, C. Yin, H. Song, T. Shen, and Y. Gu, "Utfnet: Uncertainty-guided trustworthy fusion network for rgb-thermal semantic segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[16] W. Bao, Q. Yu, and Y. Kong, "Evidential deep learning for open set action recognition," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2021, pp. 13 329–13 338.

[17] A. P. Dempster, "A generalization of bayesian inference," *J. of the Roy. Statistical Soc.: Ser. B (Methodological).*, vol. 30, no. 2, pp. 205–232, 1968.

[18] A. Jsang, *Subjective Logic: A formalism for reasoning under uncertainty*. Springer, 2018.

[19] 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, 2019.

[20] J. Pak and J. W. Shin, "Sound localization based on phase difference enhancement using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1335–1345, 2019.
[21] R. Varzandeh, K. Adiloğlu, S. Doclo, and V. Hohmann, "Exploiting periodicity features for joint detection and doa estimation of speech sources using convolutional neural networks," in *IEEE Int. Conf. Acouts., Speech, Signal Process.*, 2020, pp. 566–570.
[22] L. Comanducci, F. Borra, P. Bestagini, F. Antonacci, S. Tubaro, and A. Sarti, "Source localization using distributed microphones in reverberant environments based on deep learning and ray space transform," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2238–2251, 2020.
[23] T. N. T. Nguyen, W.-S. Gan, R. Ranjan, and D. L. Jones, "Robust source counting and doa estimation using spatial pseudo-spectrum and convolutional neural network," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2626–2637, 2020.
[24] D. Krause, A. Politis, and K. Kowalczyk, "Comparison of convolution types in cnn-based feature extraction for sound source localization," in *Eur. Signal Process. Conf.*, 2021, pp. 820–824.
[25] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Robust sound source tracking using srp-phat and 3d convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 300–311, 2021.
[26] L. Cheng, X. Sun, D. Yao, J. Li, and Y. Yan, "Estimation reliability function assisted sound source localization with enhanced steering vector phase difference," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 421–435, 2021.
[27] W. He, P. Motlicek, and J.-M. Odobez, "Neural network adaptation and data augmentation for multi-speaker direction-of-arrival estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1303–1317, 2021.
[28] T. N. T. Nguyen, K. N. Watcharasupat, N. K. 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, Language Process.*, vol. 30, pp. 1749–1762, 2022.
[29] K. SongGong, W. Wang, and H. Chen, "Acoustic source localization in the circular harmonic domain using deep learning architecture," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 2475–2491, 2022.
[30] S. Y. Lee, J. Chang, and S. Lee, "Deep learning-enabled high-resolution and fast sound source localization in spherical microphone array system," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.
[31] T. Zhong, I. M. Velázquez, Y. Ren, H. M. P. Meana, and Y. Haneda, "Spherical convolutional recurrent neural network for real-time sound source tracking," in *IEEE Int. Conf. Acouts., Speech, Signal Process.*, 2022, pp. 5063–5067.
[32] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Direction of arrival estimation of sound sources using icosahedral cnns," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 313–321, 2023.
[33] X.-C. Zhu, H. Zhang, H.-T. Feng, D.-H. Zhao, X.-J. Zhang, and Z. Tao, "Ifan: An icosahedral feature attention network for sound source localization," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–13, 2024.
[34] R. M. Neal, *Bayesian Learning for Neural Networks*. Springer, 1996.
[35] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.
[36] 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 *Advances in Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[37] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *IEEE Int. Conf. Acouts., Speech, Signal Process.*, 2015, pp. 5206–5210.
[38] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.
[39] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *The J. of the Acoustical Soc. of America*, vol. 124, no. 5, pp. 2911–2917, Nov 2008.
[40] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, "The locata challenge data corpus for acoustic source localization and tracking," in *Sensor Array and Multichannel Signal Process. Workshop*, 2018, pp. 410–414.

Rendong Pi is now pursing the Ph.D. degree in the Department of Mechanical Engineering, the Hong Kong Polytechnic University. He has graduated from the School of Transportation Science and Technology, Harbin Institute of Technology with a bachelor's degree and from the School of Qilu Transportation, Shandong University with a master's degree. His research interests include audio-visual localization, physics-informed neural network, object detection and tracking, and semantic segmentation.

Xiang Yu is an assistant professor at the Department of Mechanical Engineering, the Hong Kong Polytechnic University. He obtained his BEng with first class honors and PhD from the same department in 2011 and 2015, respectively. His research primarily focuses on the development of theories, numerical methods, advanced materials, and metamaterials related to the field of acoustics and vibrations. He has published over 50 papers in SCI journals and serves as an Assistant Editor for the Journal of Sound and Vibration.