PLDISET: PROBABILISTIC LOCALIZATION AND DETECTION OF INDEPENDENT SOUND EVENTS WITH TRANSFORMERS

Peipei Wu[1], *Jinzheng Zhao*[1], *Yaru Chen*[1], *Berghi Davide*[1], *Yi Yuan*[1], *Chenfei Zhu*[2],
Yin Cao[3], *Yang Liu*[4], *Philip J.B. Jackson*[1], *Mark D. Plumbley*[1], *Wenwu Wang*[1]

[1] University of Surrey, Guildford, UK,
[2] Daqian Information, Wuhan, China,
[3] Xi'an Jiaotong-Liverpool University, Suzhou, China,
[4] Meta, Seattle, USA,

ABSTRACT

Sound Event Localization and Detection (SELD) is a task that involves detecting different types of sound events along with their temporal and spatial information, specifically, detecting the classes of events and estimating their corresponding direction of arrivals at each frame. In practice, real-world sound scenes might be complex as they may contain multiple overlapping events. For instance, in DCASE challenges task 3, each clip may involve simultaneous occurrences of up to five events. To handle multiple overlapping sound events, current methods prefer multiple output branches to estimate each event, which increases the size of the models. Therefore, current methods are often difficult to be deployed on the edge of sensor networks. In this paper, we propose a method called Probabilistic Localization and Detection of Independent Sound Events with Transformers (PLDISET), which estimates numerous events by using one output branch. The method has three stages. First, we introduce the track generation module to obtain various tracks from extracted features. Then, these tracks are fed into two transformers for sound event detection (SED) and localization, respectively. Finally, one output system, including a linear Gaussian system and regression network, is used to estimate each track. We give the evaluation results of our model on DCASE 2023 Task 3 development dataset.

***Index Terms*—** Sound Event Localization and Detection, Transformer, Linear Gaussian System

1. INTRODUCTION

Currently, applications in various fields, such as robotics and surveillance, rely on Sound Event Localization and Detection (SELD) technology. Therefore, conducting in-depth research on this topic is crucial. Since 2019, DCASE has been hosting relevant challenges that have significantly improved SELD systems [1, 2].

The first notable method in SELD is SELDNet [3]. However, it is limited in dealing with multiple overlapping events from the same class with different locations. To address this issue, EINv2 introduced a new track-wise output format [4]. Since then, Permutation-Invariant Training (PIT) has been utilized in SELD [5], which forms part of the baseline system used in DCASE 2023 Task 3. However, EINv2 still requires multiple output branches to estimate the corresponding track, which increases the model's size. Especially if the number of overlapping events is higher than the number of output branches, EINv2 cannot predict all events simultaneously. In other words, some events might be ignored.

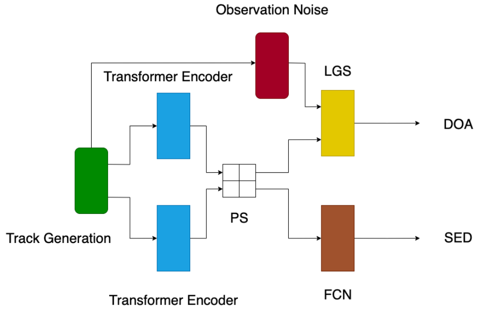

Figure 1: The new output branch for SELD. PS, LGS, and FCN denote parameter-sharing, linear Gaussian systems, and fully-connected networks, respectively.

Building upon previous work, this paper presents a novel three-stage solution for SELD. First, in contrast to EINv2, we generate different tracks from the extracted features prior to the attention module. Second, we employ a transformer instead of a simple convolutional recurrent neural network (CRNN) in SELD. Third, we introduce a linear Gaussian system to predict the Direction of Arrival (DOA) from each track rather than relying on regression networks. It is worth noting that in EINv2, the number of output branches is double the number of tracks, as each track requires separate output networks for DOA and SED predictions, respectively. If the number of tracks is large, this can pose challenges for EINv2, whereas our proposed model handles this efficiently.

In the following Section 2, we review the related work which we used in our proposed method. Section 3 introduces the proposed method in detail. Section 4 showcases the experimental results along with their corresponding analysis. The last section concludes our contribution and future work.

2. RELATED WORKS

2.1. Trackwise output format

This format type is first introduced in [4]. It can be defined as:

$$\boldsymbol{Y}_{\text{Trackwise}} = \{(y_{\text{SED}}, y_{\text{DOA}} | (y_{\text{SED}} \in \mathbb{1}_{\mathbf{S}}^{M\times K}, y_{\text{DOA}} \in \mathbb{R}^{M\times 3})\} \quad (1)$$

where y_{SED} and y_{DOA} are predictions for SED and DOA, respectively, $\mathbb{1}$ denotes one-hot encoding, M is the number of tracks, K

is the number of classes, $\mathbf{S}$ is the set of sound event classes, and $\mathbb{R}^{M\times 3}$ represents spatial information by Cartesian coordinates.

However, this format type can lead to a track permutation problem. In most cases, $M \ll K$ indicates that not all classes of sound events happen in each frame. In other words, events are not consistently predicted in fixed tracks. As a result, in the training process, tracks do not know which ground truths are corresponded to themselves correctly. To address this issue, permutation-invariant training is employed as a solution.

2.2. Permutation-Invariant training

Permutation-invariant Training was first introduced for speaker separation in [5]. Let t denote the frame index. Given a frame-level permutation set $\mathbf{P}(t)$, which consists of all possible prediction-label pairs, ground truth labels are assigned based on the possible combinations within this set of pairs. The lowest loss is then used for backpropagation. The PIT loss can be defined as follows:

$$\mathcal{L}^{\text{PIT}} = \min_{\alpha\in\mathbf{P}(t)} \sum_{M} \{\ell_{\alpha}^{\text{SED}}(t) + \ell_{\alpha}^{\text{DOA}}(t)\} \tag{2}$$

where α is one of the possible prediction-label pair, $\ell_{\alpha}^{\text{SED}}(t)$ and $\ell_{\alpha}^{\text{DOA}}(t)$ are SED and DOA loss, respectively.

2.3. Linear-Gaussian system

The linear Gaussian system represents a linear relationship between variables, where the observed variables are corrupted by Gaussian noise. This modeling approach has been widely utilized in various tasks, including detection or tracking tasks. A simple linear Gaussian system can be described by the following equation:

$$y = \mathbf{H}x + \omega \tag{3}$$

where y represents the observed state, x represents the latent state (which is hidden), $\mathbf{H}$ is the observation matrix, and ω represents the observation noise. A more complex version of the linear Gaussian system can refer to the Bayesian filters, involving parameter optimization, such as Kalman Filter [6].

3. THE PROPOSED METHOD

In this section, we will discuss the proposed method in detail. Firstly, we introduce parameter-sharing (PS) technology to enable multi-task learning. Then, we discuss the network in three stages: Feature Extraction, Transformer, and Tracks Estimation. At last, we will give a summary of the proposed method's structure.

3.1. Parameter-Sharing

Due to SELD involving both sound event detection and corresponding localization, this task is considered a complex multi-task rather than a single task. Therefore, joint SELD learning can benefit from multi-task learning (MTL) [7]. Considering that SED and DOA predictions have different noise patterns, a good representation F can average the noise patterns from both sides. Additionally, certain features R in F may be easily obtained from one side (SED or DOA) but difficult from the other side. MTL can aid in obtaining a good representation F.

Parameter-sharing (PS) is a classical MTL method, including soft PS and hard PS [8]. The comparison between soft PS, hard PS, and no PS can be seen in [4]. Thanks to their work, in this paper, we select soft PS directly. The cross-stitch is used for soft PS. Let D_c, D_t, and D_f denote the dimensions of feature maps, time steps, and frequency, respectively. The learnable parameters are denoted as $\delta_{i,j} \in \mathbb{R}^{D_c}$. From the original feature maps $(x^{\text{SED}}, x^{\text{DOA}})$, the new feature map updated by cross-stitch is given as:

$$[\hat{x}^{\text{SED}}, \hat{x}^{\text{DOA}}]^{\text{T}} = \mathbf{\Delta}[(x^{\text{SED}}, x^{\text{DOA}})]^{\text{T}} \tag{4}$$

where $\hat{x}^{\text{SED}}, \hat{x}^{\text{DOA}} \in \mathbb{R}^{D_c\times D_t\times D_f}$ is the new feature map, $\mathbf{\Delta}$ is a matrix with the dimension of 2×2 consisting the learnable parameters, and T means transpose operation.

3.2. Feature Extraction

The first stage, Feature Extraction, includes a CNN-based feature extractor, the track, and the observation noise generation module. The primary objective of this stage is to obtain feature embedding and observation noise.

SELDnet introduces a three-layer CNN-based feature extractor, but its simple structure is considered less sensitive to small-sized features. Moreover, SELDnet didn't provide extractors for SED and DOA branches separately. As a result, it might ignore some specific features R in F, as discussed earlier. Therefore, this simple extractor is not ideal for joint SELD learning. We adopted the extractor from EINv2 [4] directly. Same we provide different inputs for SED and DOA extractors. Only the DOA extractor will be applied observation noise generation module.

Afterward, we generate M tracks from feature embeddings, where M is a fixed input value. Therefore, we design a fully-connected network (FCN) to implement. First, two embeddings are flattened. Then, a linear layer is designed to increase the dimension M times. Last, we reshape the embedding and obtain M tracks. Also, the cross-stitch method is applied to the FCN.

Considering that the linear Gaussian system (LGS) is only applied to the direction of arrival (DOA) branch, we solely adopt the observation noise module for the DOA's feature map. The observation noise module consists of a linear layer to convert the feature map into the observation state noise dimension (2-D or 3-D, depending on the requirements).

3.3. Transformer

The Transformer was first proposed in [9], and we adopted it for handling temporal information. We design separate Transformers for SED and DOA, similar to the previous stage. Considering Transformer requires input with positional information. Thus, we apply a fixed absolute positional encoding on each track as follows:

$$P_{t,2i} = 0.1\sin\left(t/10^{8i/D_c}\right), \quad P_{t,2i+1} = 0.1\sin\left(t/10^{8i/D_c}\right), \tag{5}$$

where t represents the time step and i denotes the feature map index. Then, the positional encoded features will be fed into the Transformer's encoder. Each encoder layer contains 8 multi-head self-attention structures, and the input embedding dim is 512. Between each encoder layer, soft PS is applied to balance the gap between SED and DOA's representations. The entire Transformer consists of two encoder layers.

3.4. Tracks Estimation

The last stage, Tracks Estimation, aims to estimate SED and DOA in each track. In EINv2, each track has two FCNs to estimate SED and DOA. If there is more than one track, EINv2 needs to add more FCNs to cover the additional tracks. For instance, if there are three tracks, EINv2 will need 6 FCNs to cover all estimations. Different from it, we design the re-useable estimation block to cover inputs from different tracks to estimate the SED and the DOA of each track.

For SED estimation, we employ a regression method to obtain. The transformer's output is fed into FCN and activated by the sigmoid function. As for the DOA estimation, we adopt the linear Gaussian system (LGS) to calculate the posterior estimation. The calculation process is as follows:

$$\mathbf{I} = \mathbf{H}\mathbf{E}\mathbf{H}^{\mathrm{T}} + \mathbf{N}_{\mathrm{o}} \tag{6}$$

Here, $\mathbf{I}$ represents the innovation covariance matrix, $\mathbf{H}$ is the observation matrix as defined in Equation 3, $\mathbf{E}$ denotes the identity matrix, and $\mathbf{N}_{\mathrm{o}}$ is the output from the observation noise module. The observation noise is obtained by passing the observation embeddings (with a dimension of 512) through a linear layer. This projection maps the observation embeddings to the state embedding, which has a dimension of 3. The posterior covariance matrix $\mathbf{C}_{\mathrm{p}}$ is then obtained as:

$$\mathbf{C}_{\mathrm{p}} = (\mathbf{E}^{-1} + \mathbf{I})^{-1} \tag{7}$$

where $[\cdot]^{-1}$ denotes the inverse operation. Next, the residual matrix $\mathbf{R}$ is calculated as:

$$\mathbf{R} = \mathbf{H}(\mathbf{x} - \mathbf{B}_{\mathrm{o}}) \tag{8}$$

where $\mathbf{x}$ represents the state embedding transferred from observation embedding, and $\mathbf{B}_{\mathrm{o}}$ is the bias in the observation model. Finally, the DOA estimation, also known as the posterior mean matrix, is obtained as follows:

$$\hat{\mathbf{x}}^{\mathrm{DOA}} = \mathbf{C}_{\mathrm{p}}\mathbf{E}^{-1} + \mathbf{R}. \tag{9}$$

3.5. PLDISET and loss function

In the previous section, we discussed Permutation Invariant Training (PIT) but did not provide detailed information about the loss functions for sound event detection (SED) and direction of arrival (DOA). In this subsection, we will explain the loss functions and provide an overview of the PLDISET method.

We select Binary Cross Entropy (BCE) as the loss function for the SED task, which is a classification task. It measures the cross-entropy between the predictions and the labels for SED. For the DOA task, the evaluation is based on the distance between the estimations and the ground truths. Since Cartesian coordinates are introduced, we can use the mean squared error between two points in Cartesian coordinates as the loss function for DOA.

To train the SELD model and optimize its performance in both SED and DOA tasks, these loss functions are used. The overall loss is computed by summing the individual losses for SED and DOA with appropriate weights.

The overview of the PLDISET is depicted in Figure 2. For the sound event detection (SED) task, we use log mel spectrogram as the input feature. In the case of the direction of arrival (DOA) task, both log mel spectrogram and intensity vector map are selected as the input features.

4. EXPERIMENT AND EVALUATION

4.1. Dataset and data augmentation

The DCASE 2023 development dataset consists of multichannel recordings of sound scenes captured in different rooms and environments. The dataset includes temporal and spatial annotations for prominent events belonging to a set of target classes. The total duration of the dataset is 7.5 hours. However, due to the limited size of the dataset, it is insufficient to train a competitive deep-learning-based model. To overcome this limitation, we utilized the simulation generator script provided by the DCASE 2022 challenge to generate an additional 30 hours of recordings. The generated dataset includes two versions: a noiseless version and a noisy version.

4.2. Metrics

We use the DCASE challenge's metrics to evaluate our method. The evaluation metrics used in this challenge are based on true positives (TP) and false positives (FP), taking into account not only correct or wrong detections but also the proximity to a distance threshold T° (angular threshold in our case) from the reference. For this challenge, the threshold is set to $T = 20^{\circ}$. The details can be seen in [10, 11, 12].

4.3. Hyper-Parameters

We apply the Fast Fourier Transform (FFT) on the recordings using a 1024-point Hann window with a hop size of 600 points. To extract the log-mel spectrogram from the FFT result, we select 256 mel bands. Next, we segment the audio clips into chunks of a fixed length of 4 seconds without overlapping. The intensity vector map is obtained as well.

For model training, we utilize the AdamW optimizer for optimization. The initial learning rate is set to 0.0005 for the first 80 epochs and is then reduced to 0.00005 for the subsequent 10 epochs. During the finetuning of the model, the scheduler strategy changes to use a learning rate of 0.0005 for the first 10 warm-up epochs. Afterward, the learning rate is multiplied by 0.1 every 10 epochs. The weighted term for the Permutation Invariant Training (PIT) loss is selected as 0.5 for both the sound event detection (SED) and direction of arrival (DOA) losses.

4.4. Baseline system

We evaluate our proposed method by comparing it to the baseline system (SELDNet) provided by the DCASE challenge, which has been widely used as a benchmark [3, 13, 14, 15]. The baseline system extends the original SELDNet [3] by introducing multi-head self-attention blocks, using the Multi-ACCDOA output format, and employing SALSA-lite features to handle multiple overlapping sound events. Furthermore, we add EINv2 for comparison as well.

4.5. Evaluation

We compare the proposed method with the baseline and EINv2 in three steps. First, we trained all three algorithms on the noiseless dataset using respective default settings. Table 1 shows their performances. On the SED task, PLDISET and EINv2 achieved similar performance and much better than the baseline. As for the DOA

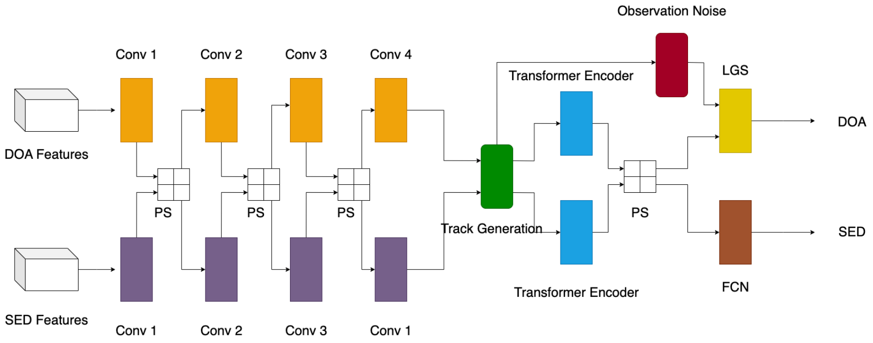

Figure 2: Network Architecture of the PLDISET.

task, PLDISET is slightly lower than the baseline, with a minor gap.

Table 1: Official metrics on the noiseless generated dataset

Methods	ER_{20}	F_{20}	LE_{CD}	LR_{CD}
baseline	0.52	49.2	18.8	58.9
EINv2	0.36	55.7	**11.3**	**79.8**
PLDISET	**0.35**	**56.1**	19.1	58.1

Afterward, we finetuned the models from the first step on the noisy datasets. The evaluation results on the test dataset are provided in Table 2. EINv2 performed best on both tasks. The proposed method achieved similar results on the SED task and was not far from the baseline on the DOA task.

Table 2: Official metrics on the noisy generated datasets

Methods	ER_{20}	F_{20}	LE_{CD}	LR_{CD}
baseline	0.55	48.9	20.0	49.9
EINv2	**0.38**	**52.5**	**13.1**	**75.2**
PLDISET	**0.38**	52.1	21.5	47.1

In the last step, we evaluated those methods on the development dataset of the DCASE Challenge 2023. We finetuned models from previous steps on the training part. Table 3 demonstrates the results on the evaluation set. The proposed method and EINv2 outperform well on the SED task with an error rate of around 0.39. The performance of PLDISET on the DOA task is close to the baseline.

Table 3: Official metrics on the DCASE development dataset

Methods	ER_{20}	F_{20}	LE_{CD}	LR_{CD}
baseline	0.57	48.7	22.0	47.7
EINv2	**0.38**	**53.3**	**14.5**	**72.4**
PLDISET	0.39	52.6	23.6	47.4

The proposed method shows its advantages on the SED task in the three comparisons, with the lowest error rate of 0.35 and the highest of 0.39. Considering that some datasets consist of real-world recordings that are more challenging than the simulated data, the proposed method shows its excellent capability in handling the SED tasks under different complex scenarios. As for the DOA task, unlike other works, we adopt a probabilistic method for localization instead of a regression-based approach. However, the PLDISET method shows a gap in the DOA task compared to EINv2 and the baseline. The possible reason for the disadvantage is the LGS may result in lower accuracy in the DOA estimations due to inaccurate prior information or an inappropriate model.

Compared to other works, one of the distinguishing features of PLDISET is its ability to estimate all tracks using a single output branch. For most methods, they require assigning output modules for each track. But PLDISET can reuse the output module for each track. The experimental results demonstrate that PLDISET performs well in SED tasks, showing its strong ability to accurately detect and classify sound events without multiple regression networks. Although the localization ability may not be as refined as in some other works, it still achieves satisfactory results. Overall, PLDISET balances sound event detection and localization tasks well. Considering that the parameters of the LGS can be updated and constrained by certain rules, there are potential research prospects in further exploring and refining this aspect. By improving the prior information and refining the model, it may be possible to enhance the accuracy of DOA estimations in the PLDISET method. Besides that, PLDISET shows its prospects of extending into a tracking version. In tracking problems, different numbers of targets appear in each frame which is quite common. Currently, PLDISET reuses the single output branch to cover all tracks, which can be improved to handle different tracks input. In addition, temporal information can be considered in the tracking problem. Therefore, some historical information, such as the Kalman Filter decreasing the error by regression in the transaction, can be used to adjust the LGS to improve tracking accuracy.

5. CONCLUSION AND FUTURE WORK

In this study, we introduced a novel network called PLDISET for SELD. We design the new output branch to estimate all tracks rather than create several branches for each track. The proposed method is evaluated on three datasets by comparing the baseline and EINv2 to show its advantages and potential. The source code and improving work based on the proposed method for sound event tracking will be released in the future.

6. REFERENCES

[1] W.-G. Choi and J.-H. Chang, "Confidence regularized entropy for polyphonic sound event detection," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

[2] J. Hu, Y. Cao, M. Wu, Q. Kong, F. Yang, M. D. Plumbley, and J. Yang, "Sound event localization and detection for real spatial sound scenes: Event-independent network and data augmentation chains," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

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

[4] Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang, and M. D. Plumbley, "An Improved Event-Independent Network for Polyphonic Sound Event Localization and Detection," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 00, pp. 885–889, 2021.

[5] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245.

[6] G. Bishop, G. Welch, *et al.*, "An introduction to the kalman filter," *Proc of SIGGRAPH, Course*, vol. 8, no. 27599-23175, p. 41, 2001.

[7] P. Vafaeikia, K. Namdar, and F. Khalvati, "A brief review of deep multi-task learning and auxiliary task learning," *arXiv*, 2020.

[8] D. S. Sachan and G. Neubig, "Parameter sharing methods for multilingual self-attentional translation models," *arXiv*, 2018.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv*, 2017.

[10] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in dcase 2019," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9306885

[11] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 333–337.

[12] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in dcase 2019," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2021.

[13] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji, "Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," *arXiv*, 2022.

[14] T. N. T. Nguyen, D. L. Jones, K. N. Watcharasupat, H. Phan, and W.-S. Gan, "SALSA-lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2022. [Online]. Available: https://doi.org/10.1109%2Ficassp43922.2022.9746132

[15] P. Sudarsanam, A. Politis, and K. Drossos, "Assessment of self-attention on learned features for sound event localization and detection," 2021.