

# FUN-SSL: FULL-BAND LAYER FOLLOWED BY U-NET WITH NARROW-BAND LAYERS FOR MULTIPLE MOVING SOUND SOURCE LOCALIZATION

Yuseon Choi, Hyeonseung Kim, Jewoo Jun, Jong Won Shin

Gwangju Institute of Science and Technology  
{newsun0130, kimhs355, zeusfront}@gm.gist.ac.kr, jwshin@gist.ac.kr

## ABSTRACT

Dual-path processing along the temporal and spectral dimensions has shown to be effective in various speech processing applications. While the sound source localization (SSL) models utilizing dual-path processing such as the FN-SSL and IPDnet demonstrated impressive performances in localizing multiple moving sources, they require significant amount of computation. In this paper, we propose an architecture for SSL which introduces a U-Net to perform narrow-band processing in multiple resolutions to reduce computational complexity. The proposed model replaces the full-narrow network block in the IPDnet consisting of one full-band LSTM layer along the spectral dimension followed by one narrow-band LSTM layer along the temporal dimension with the FUN block composed of one Full-band layer followed by a U-net with Narrow-band layers in multiple scales. On top of the skip connections within each U-Net, we also introduce the skip connections between FUN blocks to enrich information. Experimental results showed that the proposed FUN-SSL outperformed previously proposed approaches with computational complexity much lower than that of the IPDnet.

**Index Terms**— Dual-path processing, Multi-resolution analysis, Sound Source Localization, Multiple Moving Sources

## 1. INTRODUCTION

Sound source localization (SSL) aims to estimate the spatial position of one or multiple sound sources with respect to a given microphone array based on multi-channel audio signals. The output of the SSL is exploited in a variety of downstream applications such as sound source separation, speech enhancement, and automatic speech recognition [1, 2, 3]. Traditional approaches [4, 5, 6, 7] apply statistical analyses on a certain spatial features to localize sound sources, but degrade

in the presence of heavy noises and reverberations. Deep learning-based approaches [8, 9, 10, 11, 12, 13, 14, 15, 16] have shown superior performance in challenging acoustic scenarios when sufficient training data is available. Many of them, however, focus on localization of a single source or static sources, although the localization of multiple moving sources would widen the application of the SSL.

To localize multiple moving sources in adverse environment, the SSL model needs to capture temporal context to track the moving sources and also exploit spectral correlation to deal with challenging scenarios. Commonly adopted network architectures for moving SSL include convolutional neural networks (CNN) [17, 18] and convolutional recurrent neural networks (CRNN) [19, 20, 21]. IPDnet [22] employs dual-path processing along the temporal and spectral dimensions with the full-narrow (FN) network blocks, which achieves remarkable performance in localizing both single and multiple sources under static and dynamic conditions. The full-band layer along the spectral dimension captures inter-frequency correlations, whereas the narrow-band layer models temporal dynamics within individual frequency bands. While this structure was shown to be effective to track multiple moving sources, the computational complexity was rather high.

In this study, we propose a modified network architecture to localize multiple moving sources more efficiently. As the U-Net [23, 24] can effectively capture multi-scale information across different resolutions with relatively low computational cost, we have adopted a U-net structure into the FN network blocks in the original IPDnet. In each repeated block, a Full-band BLSTM layer is followed by a U-Net architecture equipped with Narrow-band LSTM layers in multiple scales, which we refer to as the FUN block. By effectively integrating temporal features across multiple resolutions, FUN-SSL effectively estimates the direct-path relative transfer functions (DP-RTFs) of multiple moving sound sources. To provide richer information to the following FUN blocks, we introduce the skip connections between subsequent FUN blocks at the same resolution like the inter-U-Net skip connections in [25] on top of the skip connections within each U-Net. Experimental results on a simulated dataset demonstrated that the proposed method, FUN-SSL, outperformed previously pro-

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) - ITRC (Information Technology Research Center) grant funded by the Korea government (MSIT) (IITP-2025-RS-2021-II211835, 50%) and IITP grant funded by the Korea government (MSIT) (No.IRIS-2025-25443882, S.A.M.A.N.T.H.A: Sentient Audio Machine for Alive Natural Talking & Human Affection, 50%).

posed methods with a comparable model size and reduced computational cost.

## 2. BACKGROUND

Assuming a free- and far-field scenario, the signal captured at the  $m$ -th microphone in the short-time Fourier transform (STFT) domain for the  $n$ -th frame and the  $k$ -th frequency bin,  $X_m(n, k)$ , can be modeled as

$$X_m(n, k) = \sum_{p=1}^P A_m(k, \theta_p(n)) S_p(n, k) + V_m(n, k), \quad (1)$$

where  $S_p(n, k)$  is the  $p$ -th source signal out of  $P$  sources,  $\theta_p(n)$  represents the direction of arrival (DoA) for  $p$ -th source at the  $n$ -th time frame,  $A_m(k, \theta_p(n))$  is the direct-path acoustic transfer function (DP-ATF), and  $V_m(n, k)$  is noise including all non-direct-path contributions. The DP-RTF is defined as the ratio of the DP-ATFs for the  $m$ -th and the first microphones, which is given by

$$\begin{aligned} D_{m,1}(k, \theta_p(n)) &= \frac{A_m(k, \theta_p(n))}{A_1(k, \theta_p(n))} = \frac{e^{-j2\pi\nu_k\tau_m(\theta_p(n))}}{e^{-j2\pi\nu_k\tau_1(\theta_p(n))}} \quad (2) \\ &= e^{-j2\pi\nu_k\Delta\tau_{m,1}(\theta_p(n))} \\ &= \cos(\text{IPD}_{m,1}(k, \theta_p(n))) \\ &\quad + j \sin(\text{IPD}_{m,1}(k, \theta_p(n))) \end{aligned}$$

in which  $\nu_k$  is the frequency for the  $k$ -th bin,  $\tau_m(\theta)$  is the time delay in the  $m$ -th microphone signal for the source coming from the direction  $\theta$ ,  $\Delta\tau_{m,1}(\theta) = \tau_m(\theta) - \tau_1(\theta) = \frac{d\cos(\theta)}{c}$  is the inter-channel time difference between the  $m$ -th and the first microphones where  $d_m$  is the distance from the  $m$ -th microphone to the first one and  $c$  is the speed of sound, and  $\text{IPD}_{m,1}(k, \theta) = -j2\pi\nu_k\Delta\tau_{m,1}(\theta)$  is the inter-channel phase difference (IPD) for the DoA  $\theta_p(n)$ .

IPDnet [22] estimates DP-RTFs for multiple microphones using the full-band and narrow-band fusion network to localize multiple moving sources. In this paper, we propose a modified architecture which is computationally more efficient than the full-narrow network block of the IPDnet, while the input, output, training target and loss function stay the same.

## 3. METHOD

### 3.1. Overall architecture

The overall architecture of the proposed FUN-SSL is illustrated in Fig. 1. The network receives the STFT representation of multi-channel audio signals  $\mathbf{x} \in \mathbb{R}^{N \times K \times 2M}$  as input, where  $N$ ,  $K$ , and  $M$  are the numbers of frames, frequency bins, and microphones, and 2 is for the real and imaginary parts. This input  $\mathbf{x}$  is normalized by the Laplace normalization as in [22], and then processed by a series of FUN

blocks. Each FUN block is composed of an embedding module, a full-band BLSTM layer, and a U-Net with narrow-band LSTM layers, which is composed of a sequence of down-sampling layers and a series of narrow-band LSTM layers followed by up-sampling layers. The output from the last FUN block is finally processed by a causal convolutional block to produce the estimates for the real and imaginary parts of the DP-RTF vectors for each of the  $M - 1$  microphone pairs for all  $Q$  sources [22], in which the depth-wise separable convolutions with channel dimension  $C_2$  are applied instead of normal convolutions used in [22] to reduce the number of parameters and computational complexity. As in [22], the estimates for DP-RTFs are obtained once for 12 frames with the time pooling layers. The DP-RTF estimates for each of the  $Q$  candidate sources are compared with the theoretical DP-RTFs for possible directions and then the source activity and the DoA are determined.

### 3.2. FUN block

The FN block in IPDnet [22] consists of a full-band BLSTM layer and a narrow-band LSTM layer with skip connections that concatenate the input of the block into the output of each layer to enrich the information. To reduce the computational complexity and exploit temporal correlations in multiple scales, we propose to employ FUN blocks which is a dual-path processing block built upon the FN block. FUN block consists of an embedding module, a full-band BLSTM layer, down-sampling and up-sampling modules, and narrow-band LSTM layers in multiple resolutions. FUN block does not have any skip connection from the block input which contributes to a significant amount of computational complexity, but has skip connections inside the U-Net and between the neighboring FUN blocks in multiple scales.

Unlike the FN block in which the input is directly fed into a full-band BLSTM layer, FUN block introduces the embedding module composed of a fully-connected (FC) layer, a PReLU activation function, and cumulative layer normalization (cLN) [26]. Since the subsequent down-sampling and up-sampling modules perform depth-wise convolutions only, a fully connected layer is employed to mix information across channels and compensate for this limitation. For the  $i$ -th FUN block, the processing within the embedding module is represented as

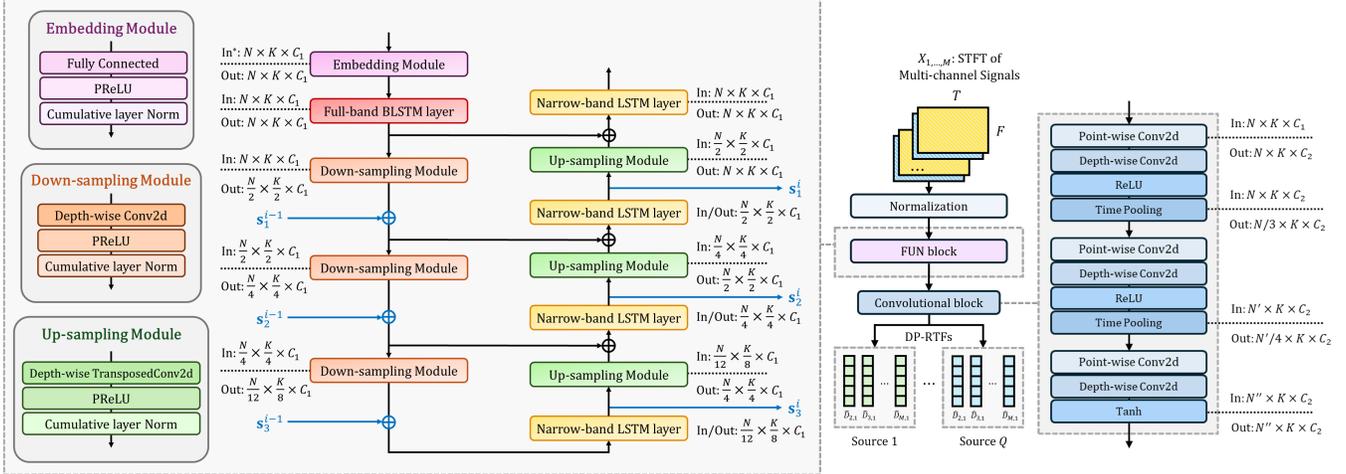
$$\mathbf{x}_{\text{out}}^i = \text{cLN}(\text{PReLU}(\text{FC}(\mathbf{x}_{\text{in}}^i))) \in \mathbb{R}^{N \times K \times C_1}, \quad (3)$$

where  $\mathbf{x}_{\text{in}}^i \in \mathbb{R}^{N \times K \times C_1}$ ,  $i \geq 2$ , and  $\mathbf{x}_{\text{in}}^0 = \mathbf{x} \in \mathbb{R}^{N \times K \times 2M}$ .

The output of the embedding module is then processed by a single full-band BLSTM layer as in the FN block, i.e.,

$$\mathbf{d}_0^i = \text{BLSTM}_{\text{full}}(\mathbf{x}_{\text{out}}^i) \in \mathbb{R}^{N \times K \times C_1}, \quad (4)$$

where  $\mathbf{d}_0^i$  denotes the features after full-band processing, and the layer  $\text{BLSTM}_{\text{full}}$  operates along the frequency axis.



\*  $N \times K \times 2M$  for the first block,  $N \times K \times C_1$  for the rest of the blocks

**Fig. 1.** Network Architecture of the proposed FUN-SSL.

The output of the full-band BLSTM layer is processed by a U-Net with narrow-band LSTM layers detailed in Fig. 1. Each of three down-sampling modules comprises a depth-wise 2D convolutional layer, PReLU, and cLN. As in [27], we employ a depth-wise convolution instead of a standard convolution for more efficient processing. The output from the  $j$ -th down-sampling module becomes

$$\mathbf{d}_j^i = \text{cLN} \left( \text{PReLU} \left( \text{DConv}_{(5, 2h_j)}^{(2, h_j)} \left( \mathbf{d}_{j-1}^i \right) \right) \right) + \mathbf{s}_j^{i-1}, \quad (5)$$

in which  $h_j$  is the down-sampling factor set to be  $h_1 = 2$ ,  $h_2 = 2$ , and  $h_3 = 3$ , and the kernel size and stride for the depth-wise convolution is  $(5, 2h_j)$  and of  $(2, h_j)$ .  $\mathbf{s}_j^{i-1}$  is the output of the narrow-band LSTM layer operating at the same scale with  $\mathbf{d}_j^i$  in the  $(i-1)$ -th FUN block connected with a inter-U-Net skip connection. For the first block,  $\mathbf{s}_j^0$  is initialized as a zero vector.

For each scale, the signal is processed by a narrow-band LSTM layer. The output of the  $j$ -th narrow-band LSTM layer is denoted as  $\mathbf{s}_{4-j}^i$  as it has the same scale with the  $(4-j)$ -th down-sampling module output, and is given by

$$\mathbf{s}_{4-j}^i = \text{LSTM}_{\text{narrow}}(\mathbf{u}_{j-1}^i + \mathbf{d}_{4-j}^i), \quad j = 1, 2, 3, 4, \quad (6)$$

in which  $\mathbf{u}_{j-1}^i$  is the output of the  $(j-1)$ -th up-sampling module with  $\mathbf{u}_0^i = \mathbf{0}$ . Here,  $\text{LSTM}_{\text{narrow}}$  operates along the time axis. Each of three up-sampling modules takes  $\mathbf{s}_{4-j}^i$  as input and processes it with a depth-wise transposed 2D convolutional layer, PReLU, and cLN, i.e.,

$$\mathbf{u}_j^i = \text{cLN} \left( \text{PReLU} \left( \text{DConvTransposed}_{(5, h_{4-j})}^{(2, h_{4-j})} \left( \mathbf{s}_{4-j}^i \right) \right) \right). \quad (7)$$

The output of the last narrow-band LSTM layer,  $\mathbf{s}_0^i$ , becomes the output of the  $i$ -th FUN block.

It is noted that in the causal convolutional block which processes the output of the last FUN block with a depth-wise separable convolutions, the point-wise convolution is applied prior to the depth-wise convolution unlike the conventional depth-wise separable convolution, as the up-sampling modules in the FUN block consist of depth-wise convolutions.

#### 4. EXPERIMENTAL SETUP

The proposed model was trained and evaluated on a simulated dataset with a sampling rate of 16 kHz similar to the one used in [21] and [22]. The microphone signals were synthesized by convolving clean speech signals from the LibriSpeech [28] corpus with room impulse responses (RIRs) generated by gpuRIR [29]. Reverberation time (RT60) was sampled between 0.2 s and 1.3 s, and room dimensions were randomly selected within the range of  $6 \times 6 \times 2.5$  m to  $10 \times 8 \times 6$  m. A maximum of two static or moving sound sources were considered. The proportions for static and moving sources as well as those for a single source and two sources cases were 50% and 50%. As in [10, 21, 15, 22], the moving trajectory of each sound source was constructed by selecting random start and end points within the room, connecting them with a straight line, and adding sinusoidal perturbations along each axis (x, y, z) to generate curved paths. Two microphones were placed at random positions on the same horizontal plane with an inter-microphone distance of 8 cm. Diffuse noises generated by the ANF generator [30] from the white, babble, and factory noise from the NOISEX-92 database [31] were added at a random signal-to-noise ratio (SNR) between -5 dB and 15 dB. 300,000, 4,000, and 4,000 samples with a duration of 4.5 s were generated to form the training, validation, and test sets, respectively. SRP-DNN [21] and IPDnet [22] were compared.

**Table 1.** Complexity and performance comparison with previously proposed methods. †: experiments with the code from the authors.

| Model        | # Params. | FLOPs    | Gross Acc     | Fine Error  | FAR          |
|--------------|-----------|----------|---------------|-------------|--------------|
| SRP-DNN [21] | 0.8 M     | 2.3 G/s  | 80.1 %        | 2.9°        | 13.1 %       |
| IPDnet [22]  | 0.7 M     | 19.4 G/s | 91.7 %        | 2.1°        | 7.7 %        |
| IPDnet†      | 0.7 M     | 19.4 G/s | 93.0 %        | 2.0°        | 7.1 %        |
| FUN-SSL      | 0.8 M     | 10.8 G/s | <b>94.2 %</b> | <b>1.9°</b> | <b>5.8 %</b> |

The length of a window was 512 samples with a 50% overlap, and 512 point STFT was applied. The maximum number of sources  $Q$  was set to 2. The threshold for sound source activity detection was set so that the miss detection rate (MDR) and the false alarm rate (FAR) became the same. The number of FUN blocks was set to 2. The channel dimensions for (B)LSTM layer and convolutional layer were configured as  $C_1 = 96$  and  $C_2 = 128$ , respectively. As in the IPDnet [22], a permutation-invariant training (PIT) was employed with a mean squared error loss between the target and estimated DP-RTF. The Adam optimizer was used and the batch size was 16. The model was optimized for 40 epochs using a learning rate initialized to 0.001 with an exponential decay factor of 0.95.

The performance of source localization was evaluated only for speech-active frames. The azimuth candidates were discretized at a resolution of  $1^\circ$ , and the angular estimation error was defined as the absolute difference between the estimated and the target azimuth. Gross accuracy measures the ratio of the speech active frames detected and the error is less than  $ET = 10^\circ$ , FAR is the number of estimated sources which are not active or the error are more than  $ET$  divided by the number of active sources, and Fine Error is the mean absolute error for the frames with the errors less than  $ET$  [22, 32].

## 5. RESULTS

The performance and computational complexity of each localization model is presented in Table 1. It is noted that the performances for SRP-DNN [21] and IPDnet [22] are from the corresponding papers, for which the test data were constructed in a similar manner but not identical to the test data used to evaluate FUN-SSL. The experimental results using the official code from the authors<sup>1</sup> are denoted as IPDnet†. The performance of the SRP-DNN [21] was inferior to other models, although it was a much lighter model. The IPDnet† showed slightly better performance to that reported in the paper, and FUN-SSL outperformed IPDnet† in all metrics with slightly more parameters and almost half of the computations.

Table 2 presents the ablation study for the number of FUN

**Table 2.** Complexity and performance according to the number of FUN blocks.

| # Blocks | # Params. | FLOPs [G/s] | Gross Acc     | Fine Error  | FAR          |
|----------|-----------|-------------|---------------|-------------|--------------|
| 1        | 0.4 M     | 5.6 G/s     | 91.9 %        | 2.2°        | 8.1 %        |
| 2        | 0.8 M     | 10.8 G/s    | 94.2 %        | <b>1.9°</b> | 5.8 %        |
| 3        | 1.2 M     | 16.0 G/s    | <b>94.5 %</b> | <b>1.9°</b> | <b>5.9 %</b> |

**Table 3.** Performance comparison with the SSL using FN blocks with the same number of LSTM layers and FLOPs.

| Block        | # Params. | FLOPs    | Gross ACC     | Fine Error  | FAR          |
|--------------|-----------|----------|---------------|-------------|--------------|
| 2 FUN Blocks | 0.8 M     | 10.8 G/s | <b>94.2 %</b> | <b>1.9°</b> | <b>5.8 %</b> |
| 5 FN Blocks  | 0.4 M     | 10.8 G/s | 93.0 %        | 2.2°        | 7.0 %        |

blocks in the FUN-SSL. The results implied that the second block was necessary for good performance, but the third block had a marginal impact on the performance.

Due to the multi-scale structure in the FUN block, it contains five (B)LSTM layers in total, while the FN block in the IPDnet has two (B)LSTM layers. To examine whether the performance improvement was from the increased number of LSTM layers or the proposed structure of the FUN block, we performed an additional experiment comparing the FUN-SSL using two FUN blocks with  $C_1 = 96$  with the SSL replacing two FUN blocks with five FN blocks having 10 (B)LSTM layers in total, for which  $C_1$  was adjusted to 56 to have similar computational cost with the FUN-SSL. The results are shown in Table 3. FUN-SSL achieved superior localization performance than the SSL with five FN blocks with the same computational complexity. This confirmed that the performance improvement was not merely due to the increased number of LSTM layers, but stemmed from the effectiveness of the proposed FUN block architecture including U-Net architecture and inter-block skip connections.

## 6. CONCLUSION

In this paper, we propose a novel architecture for sound source localization improving the one for the IPDnet, FUN-SSL, which is composed of a full-band layer followed by a U-Net with multi-scale narrow-band layers. By introducing a U-Net structure with depth-wise convolutions, the temporal dependencies were exploited at multiple scales computationally efficiently. In addition, inter-U-Net skip connections are introduced to further enhance the model capacity by preserving important spatial information from previous blocks and propagating it to a subsequent processing stage. FUN-SSL outperformed the IPDnet with a comparable number of parameters and significantly lower computational cost, demonstrating the effectiveness of the proposed architecture.

<sup>1</sup><https://github.com/Audio-WestlakeU/FN-SSL>

## 7. REFERENCES

- [1] A. S. Subramanian, C. Weng, S. Watanabe, M. Yu, and D. Yu, "Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition," *Comput. Speech & Lang.*, vol. 75, pp. 101360, 2022.
- [2] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *J. Acoustical Soc. America*, vol. 152, no. 1, pp. 107, July 2022.
- [3] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, 2017.
- [4] 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, 2003.
- [5] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas and Propag.*, vol. 34, no. 3, pp. 276–280, 1986.
- [6] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ild and itd," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 68–77, 2010.
- [7] H. Song and J. W. Shin, "Multiple sound source localization based on interchannel phase differences in all frequencies with spectral masks," in *Interspeech*, 2021, pp. 671–675.
- [8] J. Pak and J. W. Shin, "Sound localization based on phase difference enhancement using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1335–1345, 2019.
- [9] 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, Lang. Process.*, vol. 29, pp. 1303–1317, 2021.
- [10] 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, Lang. Process.*, vol. 29, pp. 300–311, 2020.
- [11] J. H. Cho and J. H. Chang, "Sr-srp: Super-resolution based srp-phat for sound source localization and tracking," in *Interspeech*, 2023, pp. 3769–3773.
- [12] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Direction of arrival estimation of sound sources using icosahedral cnns," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 313–321, 2022.
- [13] B. Yang, H. Liu, and X. Li, "Learning deep direct-path relative transfer function for binaural sound source localization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3491–3503, 2021.
- [14] R. Pi and X. Yu, "Uncertainty estimation for sound source localization with deep learning," *IEEE Trans. Instrumentation and Measurement*, 2024.
- [15] Y. Wang, B. Yang, and X. Li, "Fn-ssl: Full-band and narrow-band fusion for sound source localization," in *Interspeech*, 2023, p. 3779–3783.
- [16] Y. Xiao and R. K. Das, "Tf-mamba: A time-frequency network for sound source localization," in *Interspeech*, 2025, pp. 948–952.
- [17] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Exploiting temporal context in cnn based multisource doa estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1594–1608, 2021.
- [18] L. Wang, Z. Jiao, Q. Zhao, J. Zhu, and Y. Fu, "Framewise multiple sound source localization and counting using binaural spatial audio signals," in *ICASSP*, 2023, pp. 1–5.
- [19] 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, pp. 716–720.
- [20] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal Select. Topics Sig. Process.*, vol. 13, no. 1, pp. 34–48, 2018.
- [21] B. Yang, H. Liu, and X. Li, "Srp-dnn: Learning direct-path phase difference for multiple moving sound source localization," in *ICASSP*, 2022, pp. 721–725.
- [22] Y. Wang, B. Yang, and X. Li, "Ipdnet: A universal direct-path ipd estimation network for sound source localization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2024.
- [23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MIC-CAI*, 2015, pp. 234–241.
- [24] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv:1806.03185*, 2018.
- [25] A. A. Albishri, S. J. H. Shah, and Y. Lee, "Cu-net: Cascaded u-net model for automated liver and lesion segmentation and summarization," in *BIBM*, 2019, pp. 1416–1423.
- [26] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [27] E. Tzinis, Z. Wang, and P. Smaragdīs, "Sudo rm-rf: Efficient networks for universal audio source separation," in *MLSP*, 2020, pp. 1–6.
- [28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.
- [29] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpurir: A python library for room impulse response simulation with gpu acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021.
- [30] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating non-stationary multisensor signals under a spatial coherence constraint," *J. Acoustical Soc. America*, vol. 124, no. 5, pp. 2911–2917, 2008.
- [31] A. Varga and H. J. M. 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.
- [32] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1503–1512, 2012.