A Study of Icosahedral Feature Sound Source Localization Method with Hybrid Dilation Convolution

Zhi-xin Qiu, Jian-wei Niu, Zhan-xu Shen, Yu-xuan Yang, Shu-ju Zhang, Xiao-jun Zhang*, and Zhi Tao*

***Abstract*—Currently available deep learning methods for sound source localization encounter the problems of poor accuracy, high complexity, and susceptibility to environmental interference. To solve these problems, this paper proposes an icosahedral feature-based method of sound source localization that uses hybrid dilated convolutions. The network uses icosahedral convolutional features of the phase-transformed steering response power as input features. A downsampling module consisting of the icosahedral convolution, dilated convolution, and a normalization layer is stacked to form the main structure of the network. In addition, the hybrid dilated convolution is used to enhance the receptive field and acquire multi-scale spatial information for accurate sound source localization. The proposed method was compared with several state-of-the-art models of sound source localization on multiple tasks from the LOCATA database. When tested on audio signals without and with silent segments, the direction of arrival (DOA) obtained by the proposed model had root mean-squared angular errors of 6.34° and 7.15°, respectively, at spherical distances. Both these results were superior to those of advanced models from the literature, and this shows that the proposed method is more accurate and robust than currently used techniques.**

***Index Terms*—Sound source localization (SSL), hybrid dilated convolution (HDC), icosahedral convolution, steered response power with phase transform (SRP-PHAT).**

I. Introduction

SOUND source localization is an important front-end technique for intelligent systems based on acoustic signal processing. Its goal is to estimate the direction of arrival (DOA) of acoustic signals by analyzing them. With the development of intelligent audio processing technology, human–computer interaction is progressing to cover an increasing number and variety of aspects of our lives. Compared with the cumbersome graphical interface, voice interaction is more convenient for the user for controlling different kinds of machines and equipment. This has led to rising demand for sound source localization technology for sound recognition and audio enhancement as well as in smart homes. However, challenges persist in research on sound source localization, and the accuracy of localization and separation of multiple sources of sound, processing of environmental noise, rapid localization of the sources of sound to ensure real-time performance, and capability of generalization of the relevant methods under different environmental conditions need to be improved. Improving the robustness, stability, and real-time performance of systems for sound source localization is thus a popular field of research.

Currently available methods of sound source localization can be roughly divided into two categories: traditional and machine learning-based methods. Traditional methods of sound source localization can be further classified into three kinds. The first kind includes methods based on the time difference of arrival (TDOA), such as those that use cross-correlation and inter-correlation. The delay estimation algorithm, which is based on the Generalized Cross Correlation PHAse Transformation (GCC-PHAT) [1] proposed by Knapp et al. in 1976, is the most widely used technique owing to its simple implementation and small number of arithmetical operations. Moreover, in the context of audio source-based methods of localization, the fusion favoring the correlation (FFC) method fuses the outputs of the Filtered Correlation based method (FCM) and the Energy Differential based method (EDM) to achieve an accuracy of localization of 88% [2]. The second class of traditional methods of sound source localization includes techniques based on beamforming, which was proposed in 1973 by Hahn et al. Beamforming is a method of focusing sound signals by adjusting the directivity of a sensor array. By reasonably designing the directivity of the transducer array, the source of sound can be precisely localized. Dibias et al. used this feature to propose the steered response power with phase transform (SRP-PHAT) [3] in 2000. This method adaptively estimates the weighting coefficients of filters according to the characteristics of the signal and noise to obtain the optimal beamformer. The third kind of traditional method of sound source localization is based on high-resolution spectral estimation. This is a subspace technique based on matrix decomposition that can describe the spatial characteristics of signals of the sound source by using vectors, and is best represented by the multiply signal classification (MUSIC) algorithm proposed by Schmidt in 1979 [4]. Grounded in research by Schmidt et al., the ESPRIT algorithm proposed by Roy et al. [5] can be used to directly locate the source of sound through eigenvalues. The above-mentioned traditional algorithms for

Manuscript received April 11, 2024; revised September 24, 2024.

This work was supported by Undergraduate Training Program for Innovation and Entrepreneurship, Soochow University, under Grant No.202310285031Z.

Zhixin Qiu is an undergraduate student of Soochow University, Suzhou 215000, China(e-mail:2123403022@stu.suda.edu.cn).

Jianwei Niu is an undergraduate student of Soochow University, Suzhou 215000, China(e-mail:2123403016@stu.suda.edu.cn).

Zhanxu Shen is an undergraduate student of Soochow University, Suzhou 215000, China(e-mail:2223403019@stu.suda.edu.cn).

Yuxuan Yang is an undergraduate student of Soochow University, Suzhou 215000, China(e-mail:2123402029@stu.suda.edu.cn).

Shuju Zhang is an undergraduate student of Soochow University, Suzhou 215000, China(e-mail:2123402033@stu.suda.edu.cn).

Xiaojun Zhang is an associate professor of Soochow University, Suzhou 215000, China(corresponding author to provide e-mail: zhangxj@suda.edu.cn).

Zhi Tao is a professor of Soochow University, Suzhou 215000, China(corresponding author to provide e-mail: taoz@suda.edu.cn).

sound source localization are based on ideal mathematical models, are highly reliant on certain assumptions, have limited adaptability, lack robustness under a low signal-to-noise ratio, are sensitive to high reverberations in the environment, and are computationally intensive, often unreliable, and incapable of delivering real-time performance. Traditional methods for DOA estimation have also evolved in recent years. Raghu et al. proposed using intra-block correlations in the SBL framework to improve DOA estimation [6], and to provide new ideas for the relevant research.

With the rapid development of artificial intelligence technology, deep learning algorithms have been applied in many fields, including sound source localization. Researchers have begun using deep neural networks for the robust estimation of the DOA of the source of sound. Xiao et al. constructed a model by using supervised learning [7] that uses a simple multi-layer perceptron (MLP) network to learn the mapping relationship between features and DOA. In 2016, Vesperini et al. used a DNN model to estimate the 2D coordinates of a speaker in a multi-room scenario [8]. This model can significantly reduce the localization error compared with traditional methods, but its performance needs to be improved in noisy environments. Ma et al. applied the DNN model to a robotic binaural system of sound source localization in 2017 by combining head motion strategies with changes in the input features to the entire cross-correlation function (CCF) to solve the problem of localizing multiple sources of sound [9]. However, DNN networks require many training parameters and a large amount of computation, while CNN networks have the advantage of parameter sharing. In 2015, Hirvonen et al. proposed using the CNN for sound source localization, but provided classifications of the directions of the source in only eight spatial domains [10]. In 2017, Yalta et al. used experiments to reveal that increasing the number of convolutional layers from 11 to 20 can reduce the influence of noise on sound source localization, and that convolutional networks are suitable for dealing with such problems [11]. In 2018, Zhang et al. proposed a CNN-based method of area localization [12] that changes the input features to the mapping of speech signals obtained from a microphone, and used it to localize a single indoor source of sound. They showed that the DOA estimated by the CNN was superior to that obtained by the SVM and MLP. The convolutional recurrent neural network (CRNN) was used as the baseline network for the task of sound source localization in the Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) [13], but its accuracy is unsatisfactory. In 2021, Bohlender et al. made improvements to Chakrabartyd's system [14] by replacing the last dense layer of the architecture with recursive layers of the long short-term memory network (LSTM) and temporal convolutional network (TCN) [15], where this improved the accuracy of sound source localization. Krause et al. found that 3D convolutions can better localize the source of sound in highly reverberant environments than 1D and 2D convolutions [16], but require a large number of computations that are time consuming. Diaz-Guerra et al. proposed the Cross3D model [17], which transforms the input into the SRP-PHAT spatial spectrum, and used a combination of 1D and 3D convolutions to track the source of sound. This fully causal model can localize the source of sound in real time. In 2022, Zhong et al. proposed a spherical CRNN based on SRP-PHAT [18] that reduced the number of requisite parameters by 85.5% and the inference time by 88.6% compared with the Cross3D model. In 2023, Diaz-Guerra et al. made another improvement to the spherical RCNN developed by Zhong et al. by using icosahedral CNNs instead of spherical convolutions [19]. This leads to fewer parameters, and enables the more accurate localization of the sound source. There is still considerable room for improving the accuracy, computational efficiency, and stability of deep learning-based methods of sound source localization. In light of the above challenges in research on sound source localization, this paper proposes an icosahedral features-based network for sound source localization that uses the hybrid dilated convolution. Its main contributions of can be summarized as follows:

1) We introduce the dilated convolution to the task of semantic segmentation and use it to localize the sound source to expand the receptive field of the network. It comprehensively considers the temporal and spatial information of the features, and improves the accuracy of localization.

2) We use the hybrid dilated convolution to extract multi-scale information. This enhances the robustness of the model and mitigates the gridding effect caused by dilated convolutions.

3) We use icosahedral convolutions to propose a network for sound source localization. It is formed by stacking downsampling modules consisting of the icosahedral convolution, dilated convolution, and normalization layer. By appropriately adjusting the number of stacked layers, the network can accurately localize the source of sound while using few computations.

II. SRP-PHAT

Steered response power with phase transform (SRP-PHAT) [3] is a mainstream method of sound source localization that is based on the phase transform-weighted steered response power. The algorithm is robust, and requires a short time for analysis. This, combined with the insensitivity of the phase transform method to the environment around the signal in terms of the estimation of the time delay, renders the system robust to reverberant environments. It can thus localize the sound source in real environments, but its localization performance is poor under a low signal-to-noise ratio.

The basic principle of SRP-PHAT is to compute the sum of functions of the generalized correlation GCC-PHAT that are weighted by the phase transforms of all microphones of the system for the received signal, and to traverse the entire source space to find the point with the largest SRP value as the estimated location of the source of sound.

We use a frame of the signal received by the microphone array to localize the source of sound. Let $X_m(n)$ denote a frame of the data received by the m-th microphone. The SRP-PHAT function can then be expressed as follows:

$$\hat{p}_{(q)} = \sum_{l=1}^{M} \sum_{m=l+1}^{M} \hat{R}_{lm}\left[\tau_{lm}(q)\right] \tag{1}$$

where q is the vector of rectangular coordinates of the hypothetical source, and $\hat{R}_{lm}\left[\tau_{lm}(q)\right]$ is the GCC-PHAT

function of the signals received from the l-th and m-th microphones that can be expressed as follows:

$$\hat{R}_{lm}(\tau) = \frac{1}{K}\sum_{K=0}^{K-1}\frac{X_l(k)\,X_m^*(k)}{|X_l(k)\,X_m^*(k)|}e^{j\omega\tau} \tag{2}$$

In (2), $X_m(k)$ is the FFT of $X_m(n)$, * represents the conjugate, K is the number of FFT points, $omega is the simulated angular frequency, and $\tau_{lm}(q)$ denotes the FFT of the hypothetical sound source with respect to the TDOA of the lth and mth microphones. We use r_l and r_m as the vectors of rectangular coordinates denoting the l-th and m-th primitives, respectively, respectively, and c is as the speed of sound in air (about 342 m/s). We then obtain the following:

$$\tau_{lm}(q) = \frac{\|q - r_m\| - \|q - r_l\|}{c} \tag{3}$$

In (3), $\|\|$ denotes the 2-Norm for finding that vector. The estimated sound source localization can then be expressed as:

$$\hat{q}_x = \underset{q\subseteq Q}{argmax}\hat{P}(q) \tag{4}$$

In (4), Q is a predefined search space. The SRP-PHAT method performs well in terms of sound source localization in reverberant environments. We thus use it as the original feature, and extract its spatial information through the icosahedral convolution to deal with the issue of rotational symmetry.

III. Icosahedral Convolution

Because the microphone array is located on a 2D fluid rather than a regular 2D plane, traditional CNNs struggle to consider the geometrical properties of the data. We use the icosahedral convolution to process these data [19]. Compared with the ordinary CNN, the icosahedral convolutional network can more comprehensively deal with 3D data. It is a gauge-equivariant convolutional network for processing signals on the icosahedron, and considers local canonical symmetry so that the network maintains covariance under local gauge transformations.

For special geometrical objects with a high degree of symmetry, such as icosahedra, it is possible to obtain a mesh of pixels that is almost completely regular and symmetric by subdividing their surfaces into small triangles and placing pixels on each triangle. The icosahedral mesh is constructed from a series of fine triangular meshes. It contains 20 equilateral triangle faces, 30 edges, and 12 vertices. Starting from the vertices, a new point is introduced on the face of each equilateral triangle to subdivide the mesh evenly into four equilateral triangles with smaller faces. By repeating this process R times, a mesh with $5 \times 2^{2r+1} + 2$ points is obtained, as shown in Fig. 1. Five maps are generated in this process. Finally, the icosahedral mesh is mapped into a rectangular mesh of size $5 \times 2^r \times 2^{r+1}$. By gradually subdividing each face of the icosahedron into four equilateral triangular faces and reprojecting each node to a unit distance from the origin, a spherical mesh is obtained. This method of discretization preserves the geometry of the sphere, while making the geodesic distance between any pair of discretized nodes nearly constant. This in turn simplifies the learning of the lifting operator and enables weight sharing.

When implementing the icosahedral convolution, the icosahedral structure needs to be constructed and mapped onto the 2D plane by using a set of atlases consisting of five-coordinate cards to describe the icosahedron, and G-padding by adding appropriate padding values around the boundaries of the inputs to extend their shape. This enables efficient convolutional operations. The convolution kernel is then expanded so that it can adapt to the shape of the icosahedral structure by rotating and padding it to this structure. Following this, a 2D convolution operation is applied to the 2D plane to obtain the convolved feature map. Finally, the feature map on the 2D plane is again mapped back to the icosahedral structure to obtain the final icosahedral convolution.

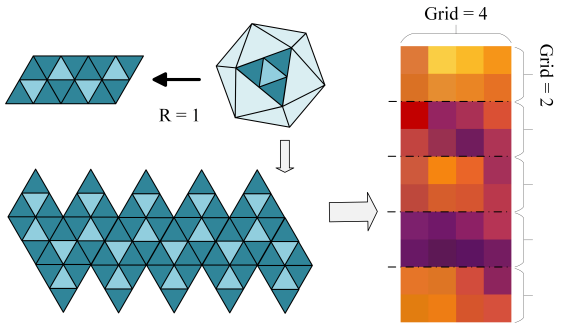


Fig. 1. Icosahedral convolution (R=1).

To work directly with off-the-shelf tools, we use the 3×3 convolution kernel of the 2D convolution to obtain the desired convolution kernel by zeroing the upper-left and lower-right points. This is subsequently canonically equated with a shape that can be convolved for the 2D convolution [20].

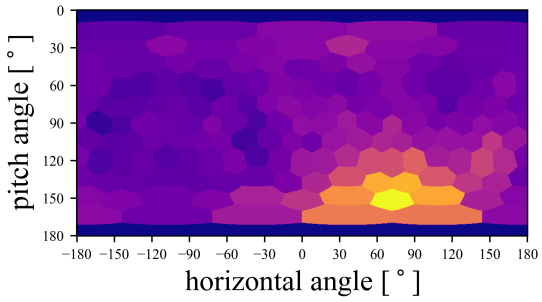


Fig. 2. Icosahedral power spectra of SRP-PHAT.

Mapping information on the location of the sound source onto an icosahedral mesh yields the same information at a lower resolution because no resolution is wasted in oversampling the poles, and the convolution does not need to learn how to deal with the distortion in the projection. The SRP-PHAT is first calculated from the input GCCs and the optimal coefficients of the filters between each pair of microphones. Its icosahedral mapping is then generated, with the results shown in Fig. 2. The canvas appears purple at a low output power, while the maximum output power is shown in yellow. The result shows that the icosahedral SRP-PHAT map can accurately calculate the direction of the sound source.

The specific algorithm is as follows:

$$GConv(f, w) = conv2d(Gpad(f), expand(w)) \tag{5}$$

$$H = 2^r + 2, R = 2^{r+1} + 2 \tag{6}$$

In (5), the weight w is defined as $(C_{out}, C_{in}R_{in}, 7)$, $expand(w)$ is defined as $(C_{out}R_{out}, C_{in}, R_{in}, 3, 3)$, and f and $Gpad(f)$ are defined as $(B, C_{in}, R_{in}, 5H, W)$. The output $GConv$ is defined as $(B, C_{out}R_{out}, 5H, W)$, where H and W are the height and width of each local chart, respectively, C is the number of channels, R is the number of dimensions of the channels, and B is the batch size.

IV. Hybrid dilated convolution

The dilated convolution is a convolution operation that increases the resolution of the given image by inserting zero values into the convolution kernel [21]. This idea of convolution was proposed to solve the problem of a reduced image resolution and the loss of information due to downsampling in problems involving the semantic segmentation of images. We introduce a new parameter, the dilation rate, to the dilated convolution to enable a larger receptive field to be obtained for a convolution kernel of the same size. Accordingly, the number of parameters used in the dilated convolution is smaller than that of a normal convolution, provided that the same receptive field is obtained. We assume that the size of the convolution kernel of the inflated convolution is k and the number of voids is d. The size of the equivalent convolution kernel can then be given as:

$$k' = k + (k-1) \times (d-1) \tag{7}$$

The receptive field is as follows:

$$RF_{i+1} = RF_i + (k'-1) \times S_i \tag{8}$$

where RF_i denotes the receptive field of the previous layer, and S_i denotes the product of the step lengths of all previous layers (excluding this layer). It is computed as follows:

$$S_i = \prod_{i=1}^{i} Stride_i \tag{9}$$

Although the dilated convolution increases the size of the receptive field without reducing the size of the feature map, it introduces a new problem, the gridding effect, that is mainly reflected in the inputs to the convolution. As the convolution kernels are spaced apart, this means that not all inputs are involved in the computation, and a discontinuity in the centroids of the convolutions is reflected in the overall feature map, especially when the superimposed convolutional layers are all used with the same rate of dilation.

A solution called the hybrid dilated convolution (HDC) was proposed in [22]. It helps the network alleviate the gridding effect and increase the size of the receptive field to improve the accuracy of detection of complex objects by concatenating a series of convolutional kernels with different dilation rates, so that the final receptive field of the network covers a square region while avoiding any voids or missing edges.

However, the problem persists if the rate of expansion is exponential [22]. We thus set the rate of expansion of the HDC to [1,2,3], so that it can fully extract information on the icosahedral features to improve the accuracy of sound source localization.

Fig. 3 shows the change in the final receptive field of the network after using the HDC.

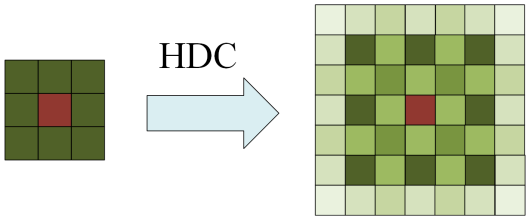


Fig. 3. Receptive fields of the network before and after using the HDC. The shades of color represent the extent to which the network uses the point: The darker the color is, the higher is the use of the point.

V. Proposed method

The structure of the icosahedral feature network based on the expanded convolutions proposed in this paper is shown in Fig. 4. The features of SRP-PHAT are used as inputs to the model after the icosahedral convolution to obtain the icosahedral features and better extract the 3D spatial information of the signal of the source of sound.

The ReLU activation function compares only the size of the inputs and zeros, has a very large gradient over positive intervals, and helps alleviate the vanishing gradient problem. Applying the ReLU activation function behind the icosahedral convolution provides a better fit to the data and facilitates the learning of complex nonlinear mapping relationships by the network to improve its performance.

Although the effect of depth downsampling on the resolution of the image was mitigated to some extent in [19] by combining two convolutions with icosahedral pooling to construct a unit, there is still considerable room for improvement. To solve this problem, we use the following strategy: The dilated convolution can be applied to the icosahedral features to enable the 1D expansion convolution to comprehensively consider the background temporal information and ensure the gauge-equivariant properties of the model. This increases the size of the receptive field through the insertion of zeros to the original convolution kernel. This model has five convolution kernels and one step. An HDC framework is used, and the rate of dilation is set to [1,2,3] to avoid the problem of meshing.

The normalization layer normalizes the inputs along 32 channels and six directions of the kernel, where this improves the capability of generalization of the model and enables it to converge more quickly. We also stack the downsampling units in the model, but the number of stacked units is appropriately reduced to avoid model overfitting and reduce the computational cost. After stacking the downsampling units, maximum pooling is applied to the six directions of the kernel of the output, and the resulting icosahedral mapping is sent to the soft-argmax function to obtain the direction of the sound source.

Because the traditional fully connected layer incurs a high computational cost, even using the convolutional and pooling layers to reduce the number of parameters reaching the fully connected layer will significantly increase the number of parameters of the model that need to be trained. We thus use the soft-argmax function instead of the fully connected layer

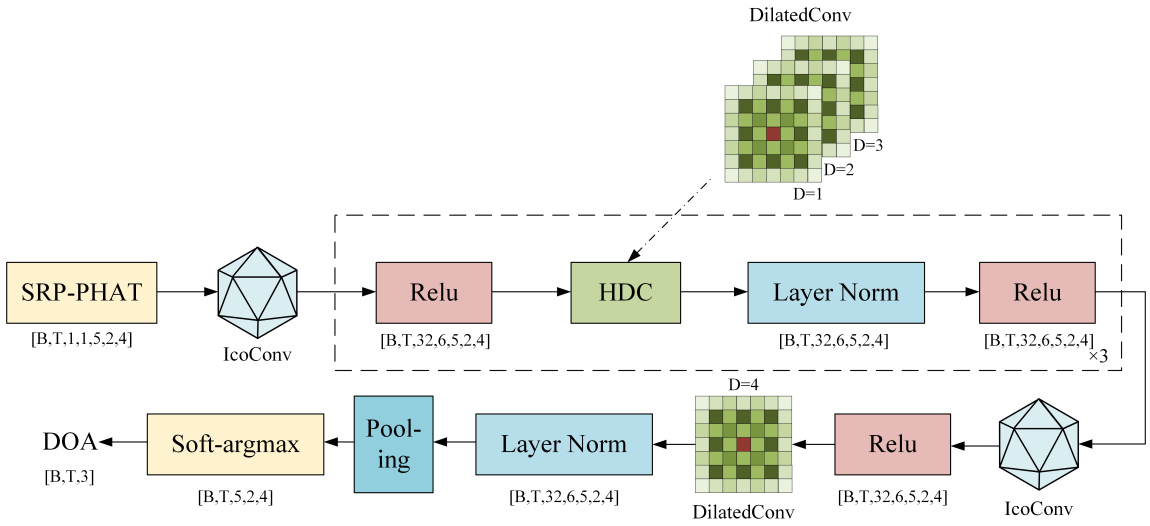


Fig. 4. Structure of the proposed model. D is the dilation rate, B is the batch size, and T is the number of frames.

to significantly reduce the amount of computation required. This ensures that the entire activation mapping sums to one by applying a softmax layer on the activation map. The softmax operation can be viewed as the process of normalizing positional information:

$$soft\text{-}max\left(P\left(x\right)\right) = \frac{e^{P(x)\text{-}max(P(x))}}{\sum_{x \in X} e^{P(x)\text{-}max(P(x))}} \tag{10}$$

where $P\left(x\right)$ denotes the output of the final convolutional layer of the model, and $x \epsilon X$ denotes the coordinates of the points sampled from the icosahedral grid.

The probability distribution of the output can be expressed as the relative weights of multiple localizations to ensure that the output is a probability distribution. Soft-argmax can be obtained by summing the product of each pixel coordinate with its probability:

$$soft\text{-}argmax\left(P\left(x\right)\right) = \sum_{x \in X} x \cdot soft\text{-}max\left(P\left(x\right)\right) \tag{11}$$

The output of the soft-argmax function results in three time series of length T with elements in the range (-1, 1), which means that the coordinates of the vectors point in the direction of the sound source in each time frame. The soft-argmax function is regarded as a differentiable version of the argmax function that can interpret the outputs of the convolutional layers as probability distributions, thus allowing us to treat DOA estimation as a regression problem. This reduces the costs of computation and memory while avoiding the introduction of non-isotropic layers.

VI. Experimental results

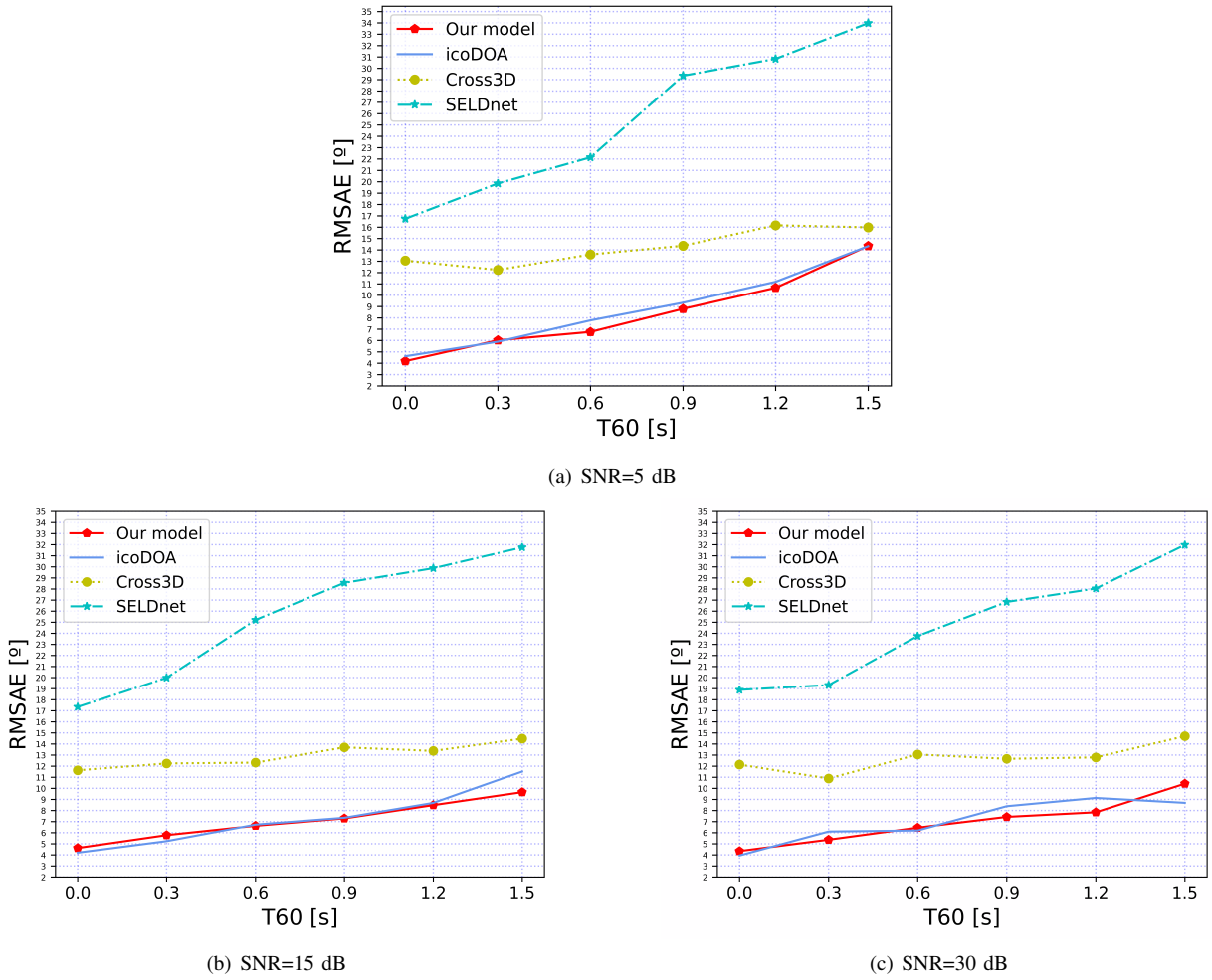
A. Training set and simulation experiments

We used simulated signals for model training. To increase the diversity of the acoustic conditions of training, each training sample was generated in real time from a random combination of parameters, such as the trajectory of the sound source, position of the microphone, source and noise signals, reverberation time, and SNR. We chose speech from the train-clean-100 subset of the LibriSpeech corpus [23] as the data source. This subset of the corpus contained 100 h of speech extracted from LibriVox audiobooks, for a total of 6.2 GB of data. The sampling frequency was 16 kHz. One randomly selected audio from the train-clean-100 subset was used as the sample of the source signal. Twenty seconds of speech were then intercepted from that audio to obtain the audio data. Although the audiobooks contained clearer speech signals, some of the audio contained strong background noise. To prevent the network from learning inaccurate information, we applied Voice Active Detection (VAD) to identify the silent parts and remove their signals to clean the data. The SNR and reverberation time (T60) were also randomly selected, and ranged from 5 dB to 30 dB, and from 0.2 s to 1.3 s, respectively.

To analyze the performance of the proposed model under different acoustic conditions, we generated the simulated signals by using the same method that was used to generate the training dataset. However, unlike the training data, the test set used the test-clean subset of the LibriSpeech corpus as the acoustic source for the simulation. In addition, some of the test samples contained a short silent section at the beginning that affected the results of estimation by the model. Localization errors in the first five frames of each trajectory were thus not considered.

We performed simulations of static source localization to evaluate the model. Simulation scenarios corresponding to several specific reverberation times T60 and SNRs were generated and tested to analyze the robustness of the proposed model. Values of the reverberation time T60 were set for six indoor scenarios, and ranged from zero to 1.5 s at internals of 0.3 s, while SNR values of 5, 15, and 30 dB were used. During the generation of the simulation data, the position of the sound source was fixed, and did not change over time.

Fig. 5 shows the results of comparison of the simulation data involving static sources of sound in different acoustic scenarios. The vertical coordinate is the root mean-squared angular error (RMSAE) of the spherical distance. SELDnet used a 2D CNN to learn the amplitude and phase spectra of the sound sources, had the lowest accuracy, was significantly



(a) SNR=5 dB

(b) SNR=15 dB

(c) SNR=30 dB

Fig. 5. Comparative simulations in different acoustic scenarios

affected by the ambient reverberation time, T60, and thus had the poorest robustness. Cross3D used a 3D CNN, could better retain the spatial and temporal relationships, and was superior to SELDnet in terms of both accuracy and robustness. The developers of icoDOA [19] concluded that when R is set to two, the input feature map contains all the information useful for sound source localization, and continuing to further reduce R does not improve the accuracy of the model. Therefore, we set R to two in the proposed model. Its accuracy of localization and robustness were higher than those of all other models considered in indoor acoustic scenarios with SNRs of 5 dB and 15 dB. When the SNR was 30 dB, its accuracy was slightly lower than that of icoDOA as the reverberation time increased. This is because the dilated convolution extracted more redundant information under high reverberations, which reduced the accuracy of the model.

B. Model testing on the LOCATA dataset

We also tested the proposed model on the LOCATA dataset, which is a database of sound signals recorded in real environments, and was created by the Computational Laboratory of the Humboldt University of Berlin. The data were recorded in a laboratory with dimensions of $7.1 \times 9.8 \times 3\ m^3$, and a duration of reverberations T60 of 0.55 s. We also tested the state-of-the-art models of sound source localization SELDnet [13], Cross3D [17], and icoDOA [19] on the LOCATA dataset, and compared them with the proposed model. We used Tasks 1 to 6 from the dataset as well as Tasks 2, 4, and 6, which involved multiple sound sources. Because the other models were tested only on tasks involving a single source of sound, we tested our model on Tasks 1, 3, and 5 as well so that the tests covered most real-world scenarios of application.

The horizontal and pitch angles are commonly used experimental metrics to assess performance in terms of sound source localization. The spherical distance between the predicted and actual positions of the source of sound is usually expressed based on its angle. Error in the computed spherical distance reflects the accuracy of sound source localization. The smaller this error is, the more accurate is the sound source localization by the corresponding model. We used the RMSAEs of the horizontal angle, pitch angle, and spherical distance angle as indices to evaluate the results. They can be expressed as:

$$RMSAE_{\vartheta} = \frac{180}{\pi}\sqrt{\frac{\sum_{i=1}^{n}(\Delta\vartheta_i)^2}{n}} \tag{12}$$

$$RMSAE_{\phi} = \frac{180}{\pi}\sqrt{\frac{\sum_{i=1}^{n}(\Delta\phi_i)^2}{n}} \tag{13}$$

$$RMSAE_{\delta} = \frac{180}{\pi}\sqrt{\frac{\sum_{i=1}^{n}(\Delta\delta_i)^2}{n}} \tag{14}$$

where n is the total number of frames of the tested speech samples, ϑ_i is the angular difference between the estimated horizontal angle of the i-th frame and the actual horizontal angle, ϕ_i is the angular difference between the estimated pitch angle of the ith frame and the actual pitch angle, and δ_i is the difference between the spherical distances of the i-th frame.

The audio samples in the LOCATA dataset included clips with silent segments and those with no silent segments. We separately tested both kinds of samples, and their results are shown in Tables I and II, respectively.

TABLE I
RMSAEs (°) OF THE DOA OF THE PROPOSED MODEL AND STATE-OF-THE-ART METHODS OF SOUND SOURCE LOCALIZATION ON THE LOCATA DATASET (CLIPS CONTAINING SILENT SEGMENTS)

Model		Task 1	Task 3	Task 5	Mean
SELDnet [13]	Horizontal angle	8.22	12.78	14.67	10.61
	Pitch angle	19.02	22.70	20.34	20.11
	Spherical distance	20.84	26.51	25.25	20.84
Cross3D [17]	Horizontal angle	4.66	8.80	10.96	6.93
	Pitch angle	3.31	2.52	6.00	3.73
	Spherical distance	5.98	9.25	12.06	8.02
icoDOA [19]	Horizontal angle	4.52	8.37	10.25	6.60
	Pitch angle	3.41	2.60	5.76	3.75
	Spherical distance	5.93	8.87	11.28	7.73
Our model	Horizontal angle	3.91	7.61	9.00	5.78
	Pitch angle	2.92	2.27	6.66	3.21
	Spherical distance	5.50	8.02	11.37	7.15

TABLE II
RMSAEs (°) OF THE DOA OF THE PROPOSED MODEL AND STATE-OF-THE-ART METHODS OF SOUND SOURCE LOCALIZATION ON THE LOCATA DATASET (NO SILENT SEGMENTS)

Model		Task 1	Task 3	Task 5	Mean
SELDnet [13]	Horizontal angle	7.52	12.08	13.58	9.83
	Pitch angle	16.56	20.32	17.61	17.61
	Spherical distance	18.35	24.13	22.39	20.49
Cross3D [17]	Horizontal angle	4.15	6.80	8.13	5.59
	Pitch angle	3.90	5.25	5.56	4.55
	Spherical distance	5.74	8.90	10.25	7.41
icoDOA [19]	Horizontal angle	4.72	6.60	7.43	5.72
	Pitch angle	3.41	2.77	5.37	3.70
	Spherical distance	6.09	7.20	9.07	6.98
Our model	Horizontal angle	3.61	6.90	3.98	5.06
	Pitch angle	2.90	2.12	4.44	3.06
	Spherical distance	5.31	7.30	8.04	6.34

The average RMSAEs of the horizontal angle, pitch angle, and spherical distance determined by the proposed model were 5.78, 3.21, and 7.15 degrees, respectively, when clips containing silent segments were considered in the tests. When such clips were not considered, the average RMSAEs of our model were 5.06, 3.06, and 6.34 degrees respectively. It outperformed all the other models in the latter case.

The experimental results show that the RCNN used by SELDnet was ineffective in mapping the spectral features of the sound signals back to the location of the source. Cross3D used a 2D SRP-PHAT map as the input, which had a low resolution. We considered only the highest accuracy obtained by this method, the network structure of which was relatively complex and computationally intensive. In addition, its RMSAEs of the horizontal angle, pitch angle, and spherical distance were smaller than those of icoDOA, and it was more robust.

Fig. 7 visualizes the results of our model of sound source localization in case of a single source on the LOCATA dataset. The solid line represents the true DOA while the dashed line denotes the estimated DOA. The gray part of the figure represents the silent segments detected by the VAD. The model needed to track the source of sound of vocalized words in between the short silent segments, but its results might have deviated from the correct DOA values because it could not receive valid information during the silent period in each clip. However, its results of estimation were always close to the actual DOA of the sound source in the audio containing silent segments, which verifies its robustness.

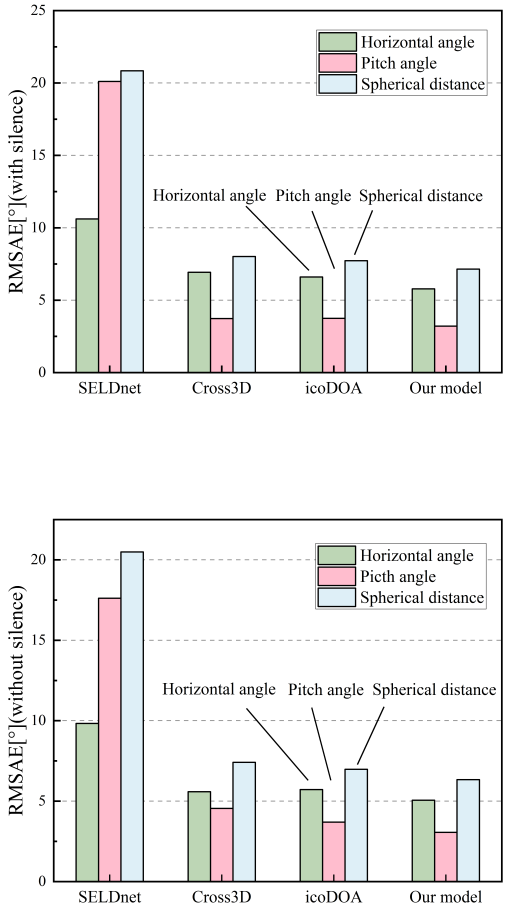
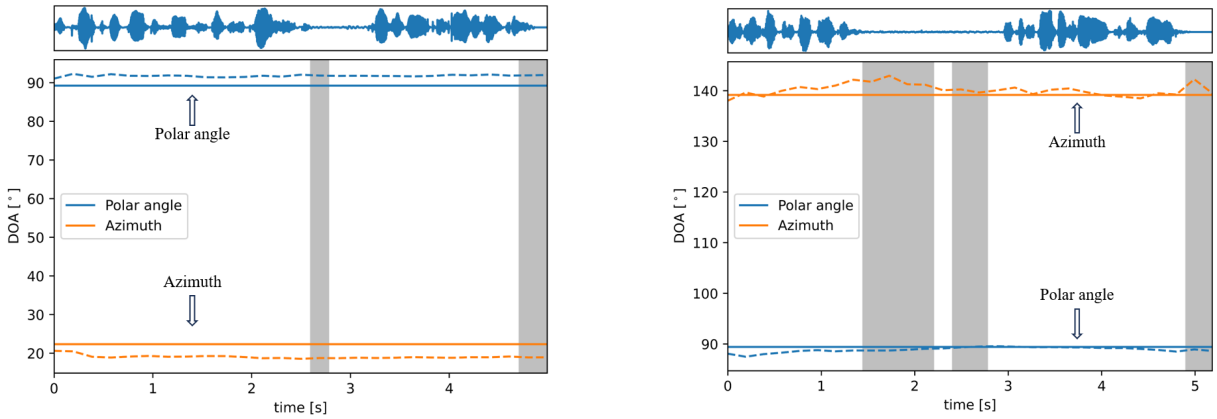


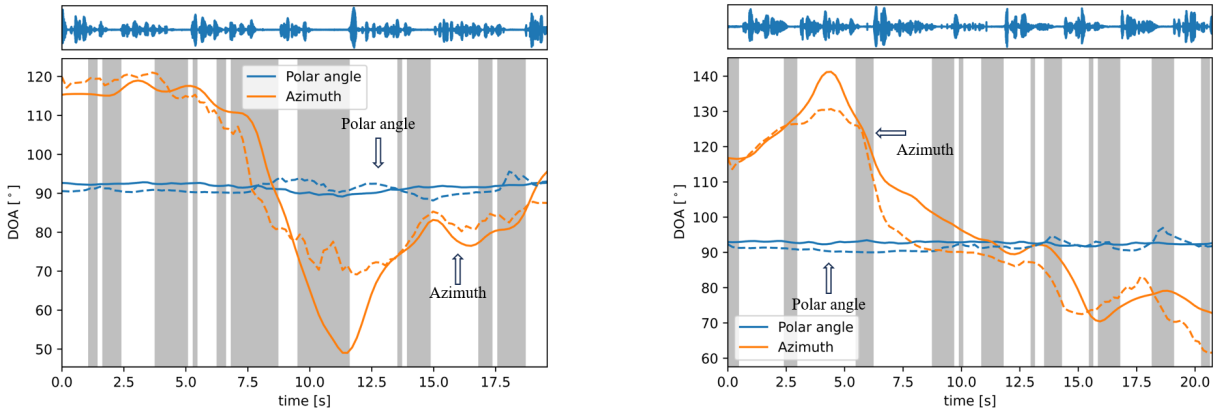
Fig. 6. Comparison of errors in the DOA between the proposed model the state-of-the-art methods of sound source localization. The top graph represents the results for audio clips containing silent segments while the bottom graph shows the results for clips that did not contain such segments.

C. Ablation experiment

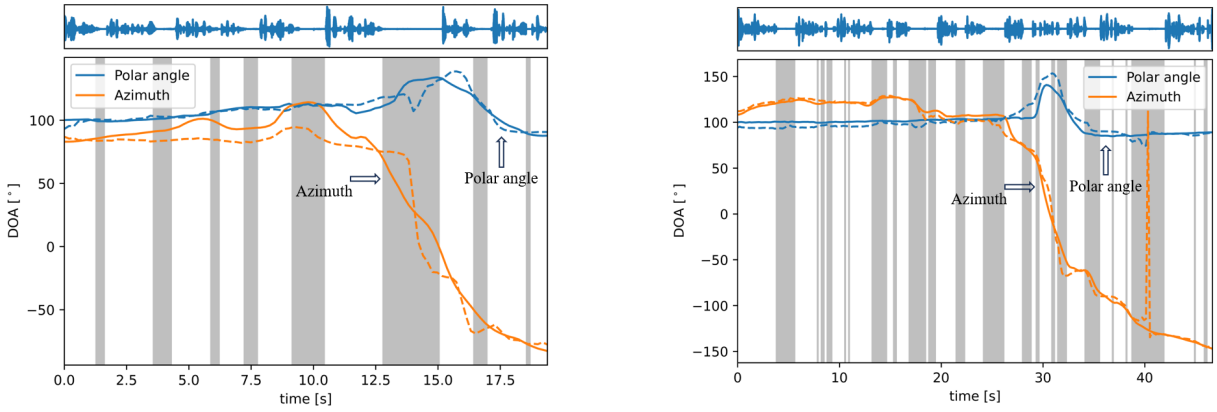
Our model uses the HDC framework to solve the gridding effect brought about by the dilated convolution, which increases the size of the receptive field and helps extract multi-scale features to improve the robustness of the model. To test the effectiveness of the HDC framework, we used the LOCATA dataset to conduct an ablation experiment. The parameters of the HDC-free model were set as follows: The rate of dilation D of the dilated convolution was set to one in all cases, while the rest of the model was kept consistent with the proposed model that contains the HDC. The ablation experiment also considered the models Cross3D, SELDnet,



(a) Task 1



(b) Task 3



(c) Task 5

Fig. 7. Example of DOA estimation on data from the LOCATA dataset

icoDOA, and the proposed model for comparison. All models were compared only in terms of the RMSAE of the spherical distance, because this index could comprehensively reflect their accuracy of sound source localization.

Fig. 8 shows that the RMSAE of the HDC-free model was close to that of icoDOA and larger than that of the HDC-containing model. The results show that when the rates of dilation D of the dilated convolution in the model were all set to one, the dilated convolution could not obtain comprehensive information owing to the gridding effect. Because the rate of dilation D was always one, multi-scale features could not be extracted, and this further affected model performance. The HDC framework improved the performance of the proposed model from various aspects, and helped it localize the sound source highly precisely in complex scenarios of application.

D. Evaluation of model computation

The number of floating-point operations per second (FLOPS) represents the volume of computation (time complexity of computation), and can be used to measure the complexity of a given algorithm. FLOPS are often used as an indirect measure of the speed of neural network models. The number of FLOPS of the convolutional layer is calculated as follows:

$$FLOPS = 2HW(C_{in}K^2 + 1)C_{out} \tag{15}$$

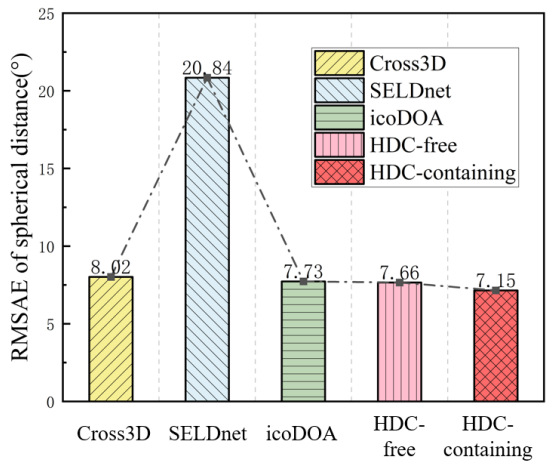


Fig. 8. Results of ablation experiment.

where C_{in} is the number of channels of the input tensor of the convolutional layer and C_{out} is that of its output tensor, while k is the size of the convolutional kernel. Then, the constant term can be removed to simplify the calculation as follows:

$$FLOPS = HW(C_{in}K^2)C_{out} \tag{16}$$

From the above equation, it is clear that the number of FLOPS of the model during computations is proportional to the square of the size of the convolution kernel in the convolutional layer. H and W are determined in the proposed model based on the resolution of the map R, which was set to two in this study. H and W were thus constant. The number of FLOPS of our model was thus determined only by the number of convolutional layers and the size of the convolutional kernel of each. Because our model uses dilated convolutions, the increase in the receptive field implies an increase in the equivalent convolution kernel, and its number of FLOPS become uncontrollable if the number of convolutional layers is not limited. $Params$ is the total number of model parameters to be trained, and represents indicates the spatial complexity of the computations. It is widely used to measure the size of the model. We use the numbers of FLOPS and $Params$ to measure the requisite numbers of computations and parameters, respectively, of the proposed model and state-of-the-art models to further evaluate their performance:

TABLE III
COMPARISON OF THE NUMBER OF COMPUTATIONS AND PARAMETERS OF DIFFERENT MODELS.

Model	FLOPS	Params[M]
SELDnet [13]	0.55G	5.34
Cross3D [17]	7.77G	21.35
icoDOA [19]	74.47M	0.29
Our model	0.88G	0.02

Table III shows that when there were four convolutional layers, our model had 0.88G FLOPS, between those of icoDOA and Cross3D, and similar to that of SELDnet. Further experiments showed that the dilated convolutional layer in our model had 0.88G FLOPS, which accounted for 99.38% of its total computations. This is because the dilated convolutional layer had a larger receptive field than the general convolutional layer, and collected and used information more comprehensively. This inevitably increased the number of computations of the model while improving its accuracy. The increase in its computational volume affected its real-time performance in terms of sound source localization, but it still recorded better results than other models (shown in Fig. 6). This shows that the number of computations of our model, 0.88G, was within an acceptable range for the sound source localization task. The value of $Params$ of our model was 0.02M, significantly lower than those of other models, and led to lower requirements of video memory for training it.

VII. CONCLUSION

This paper proposed a network for sound source localization that takes the icosahedral features of mapping of the SRP-PHAT as the input. It was formed by stacking downsampling modules consisting of an icosahedral convolution, a dilated convolution, and a normalization layer. The gridding effect caused by the dilated convolution was alleviated by introducing the HDC framework, while multi-scale information was extracted from the data to enhance the robustness of the model.

The results of experiments showed that the proposed model delivered better performance and had a higher accuracy of source localization than state-of-the-art techniques. This shows that the expanded convolution can be used in sound source localization networks to optimize the performance of the model. Our model can be applied to real-time tasks of sound source localization in scenarios involving mobile sources of sound, and can maintain a small average RMSAE even on challenging datasets of real recordings.

However, the proposed model cannot currently localize multiple sources of sound, and we plan to address this shortcoming in our future work. In addition, we will seek to enhance the capability of the network for multi-feature fusion, and to introduce an attention mechanism to it.

REFERENCES

[1] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[2] H. Sayoud, S. Ouamour, and S. Khennouf, "Virtual system of speaker tracking by camera using an audio-based source localization," *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2012, WCE 2012*, pp. 819–822, 4-6 July, 2012, London, U.K.

[3] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone arrays: signal processing techniques and applications*. Springer, 2001, pp. 157–180.

[4] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[5] R. Roy, A. Paulraj, and T. Kailath, "Esprit–a subspace rotation approach to estimation of parameters of cisoids in noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 5, pp. 1340–1342, 1986.

[6] K. Raghu and K. Prameela, "Direction of arrival estimation by employing intra-block correlations in sparse bayesian learning through covariance model," *Engineering Letters*, vol. 31, no. 1, pp. 82–92, 2023.

[7] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2814–2818.

[8] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2016, pp. 1–6.

[9] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2444–2453, 2017.

[10] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[11] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 37–48, 2017.

[12] X. Zhang, H. Sun, S. Wang, and J. Xu, "A new regional localization method for indoor sound source based on convolutional neural networks," *IEEE Access*, vol. 6, pp. 72 073–72 082, 2018.

[13] 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, 2018.

[14] S. Chakrabarty and E. A. Habets, "Multi-speaker doa estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, 2019.

[15] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Exploiting temporal context in cnn based multisource doa estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1594–1608, 2021.

[16] D. Krause, A. Politis, and K. Kowalczyk, "Comparison of convolution types in cnn-based feature extraction for sound source localization," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 820–824.

[17] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Robust sound source tracking using srp-phat and 3d convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 300–311, 2020.

[18] 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 *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 5063–5067.

[19] D. Diaz-Guerra Aparicio, A. Miguel, and J. R. Beltran, "Direction of arrival estimation of sound sources using icosahedral cnns," 2023.

[20] T. Cohen, M. Weiler, B. Kicanaoglu, and M. Welling, "Gauge equivariant convolutional networks and the icosahedral cnn," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1321–1330.

[21] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[22] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Ieee, 2018, pp. 1451–1460.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.