IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 September 2023, accepted 19 September 2023, date of publication 22 September 2023,
date of current version 29 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3318322

RESEARCH ARTICLE

SELD U-Net: Joint Optimization of Sound Event Localization and Detection With Noise Reduction

YEONGSEO SHIN[1,2], YONG GUK KIM[2], CHANG-HO CHOI[2], DAE-JOONG KIM[2], AND CHANJUN CHUN[1], (Member, IEEE)
[1]Department of Computer Engineering, Chosun University, Gwangju 61452, South Korea
[2]Maritime Research and Development Center, LIG Nex1, Seongnam 13488, South Korea

Corresponding author: Chanjun Chun (cjchun@chosun.ac.kr)

This study was supported by the research fund from Chosun University, 2022. This paper includes Shin's master's thesis.

ABSTRACT Sound event localization and detection (SELD) is a combined task that classifies acoustic events from audio signals, estimates temporal boundaries, and identifies event locations. With the advancement of industries utilizing audio signals, SELD has been applied in various fields, and deep-learning-based research is being conducted for its effective application. However, current deep-learning-based SELD research focuses mainly on performance improvement in noise-free environments, which leads to performance degradation issues in noisy environments. To address this problem, this study proposes a robust SELD U-Net model that performs SELD in noisy environments. The proposed model combines a U-Net to remove noise and a SELDnet to perform SELD. The proposed model was trained and evaluated using noisy environmental data with various sizes. Consequently, it was confirmed that the proposed model has superior performance compared with existing deep learning-based SELD models in environments with high levels of noise.

INDEX TERMS Audio signal, deep learning, noisy environment, sound detection, sound localization.

I. INTRODUCTION

Sound event localization and detection (SELD) is used in various audio applications, such as surveillance systems [1], robotics engineering [2], and voice recognition [3]. SELD comprises two main tasks: sound event detection (SED) and sound source localization (SSL). SED aims to classify acoustic events within an audio signal and identify their onset and offset. In real-life scenarios, acoustic events can occur simultaneously; therefore, overlapping events must be recognized accurately. This task is referred to as polyphonic SED [4], [5]. SED is primarily performed using supervised learning methods that predict the framewise activity of acoustic event classes. Gaussian mixture models [6] and hidden Markov models [7] were initially used. Following the advancement of deep learning algorithms, SED methods utilizing convolutional neural networks (CNNs) and recurrent neural networks (RNNs) were introduced [8], [9], [10], [11].

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

SED performs the functions of the human auditory system in several industries, including audio surveillance [12] and social welfare [13]. SSL aims to estimate the direction or location of a sound source using microphones or sensors by leveraging the configuration of microphone arrays and the time difference of arrival (TDOA) of acoustic signals to determine the location of acoustic events. SSL is performed using various techniques and algorithms, such as TDOA [14], angle of arrival [15], steered response power [16], and multiple signal classification [17]. Recently, various structures (like CNNs) that apply deep learning technologies have been proposed [18], [19]. SSL is used in various fields that employ acoustics, such as source separation [20] and acoustic analysis [21]. SELD combines SED and SSL, classifies acoustic events from audio signals, detects their activity, and estimates the location of the activated events.

One SELD approach involves independently performing SED and SSL and subsequently combining the results of each task. However, this approach encounters a tracking problem when multiple overlapping events occur simultaneously

because it requires the correct mapping of the estimated results from each task to their respective events [22]. This tracking problem can be resolved using a second method that simultaneously performs SED and SSL to execute SELD. Methods developed to perform SED and SSL simultaneously include the combination of the traditional SED and direction of arrival (DOA) execution methods [23] or their implementation using deep learning models [24]. Among these, deep-learning-based methods have shown superior performance compared with traditional methods because they can learn various acoustic features in complex acoustic environments. Furthermore, the advancement of deep learning technologies has improved the performance of SELD, thereby accelerating research on the application of these technologies.

However, most deep-learning-based SELD methods currently being researched are based on the assumption that the studied environments are associated with minimal noise. In other words, because most studies have performed SELD in noise-free or low-level-noise environments, the performance of these deep-learning-based SELD methods can deteriorate in real-world environments when considerable noise exists. In real-life scenarios, various types of noise exist. If a model performing SELD is not robust to noise, its application becomes challenging. Thus, a method that can achieve robust performance (even in noisy environments) is required. Typically, when tasks are performed using audio data from noisy environments, speech enhancement is performed as a preprocessing step for the audio data. Speech enhancement refers to the technology used to restore or enhance a target speech signal from audio data mixed with noise. While past methods were predominantly based on statistical models [25], these statistical model-based methods can undergo performance degradation when the noise characteristics change. Following the advancement of deep learning, research has been conducted on speech enhancement by applying these technologies to compensate for the shortcomings of statistical model-based methods. A representative deep-learning-based speech enhancement technology uses an autoencoder, an unsupervised learning neural network that compresses input data to encode them, and then decodes them to generate an output similar to the original input [26]. Additionally, various speech enhancement methods using U-Net, which is composed of an encoder and a decoder, are being researched. Performing SELD after speech enhancement preprocessing can resolve the performance degradation issues caused by noise. However, the disadvantage of this method is that it requires an additional speech enhancement step before performing SELD; this leads to a decrease in temporal efficiency and makes real-time processing difficult. Therefore, deep learning models that can exhibit robust performance in noisy environments without additional tasks must be studied.

This study aimed to develop a SELD model that exhibits robust performance in noisy environments, thereby overcoming the identified limitations. The proposed model has a structure that connects a U-Net, which performs speech enhancement on the input features extracted from noisy data, and a SELDnet, which performs SELD. The input features of the model utilize the input features of SELDnet extracted from audio data in a noisy environment. In this model, the encoder part of the U-Net serves to remove noisy elements from the input features, and some feature maps generated in the decoder part of the U-Net are used as inputs to SELDnet to perform SELD. Unlike existing methods, the proposed model simultaneously performs noise removal and SELD in an end-to-end manner. Owing to the application of the end-to-end method, the proposed model demonstrated improved temporal efficiency compared with existing methods.

The remainder of this paper is organized as follows: Section II introduces related studies, and Section III describes the proposed SELD U-Net. Section IV presents the execution and results of the experiments. Finally, the conclusions are listed in Section V.

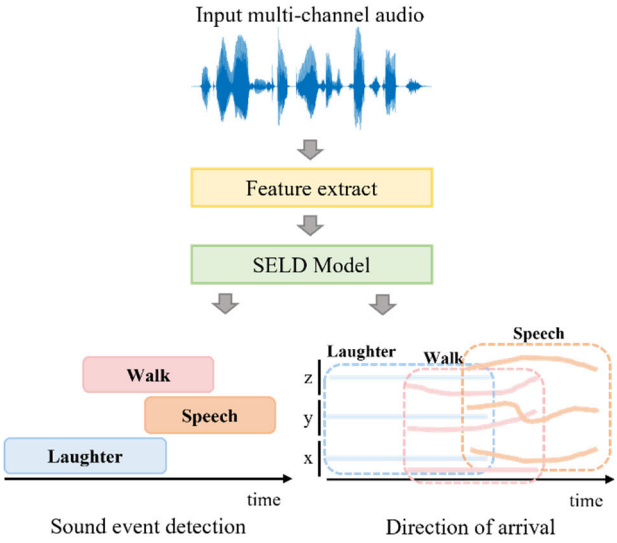

FIGURE 1. Schematic of deep-learning-based sound event localization and detection (SELD) method.

II. RELATED WORKS

A. SELD

SELD uses multichannel audio to classify acoustic events, determine their activities, and identify their locations. A deep-learning-based SELD is depicted in Fig. 1. Given the advancements of deep learning, various research efforts are underway to apply these technologies to solve the SELD problem. Research on SELD has been accelerated considerably by the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge for artificial-intelligence-based acoustic events and scene recognition technologies.

The deep-learning-based SELD methods proposed in the first DCASE Challenge held in 2019 included two-stage [27] and two-branch [28] methods. The two-stage method uses

separate networks to perform SED and SSL, whereas the two-branch method divides tasks using fully connected (FC) layers within a single model. However, the two-stage method has a higher model size owing to the use of separate networks, whereas it is difficult for the two-branch method to adjust weights appropriately, as it performs two objectives simultaneously in one model. To address the issues of these two methods, Kazuki et al. proposed in 2020 the activity-coupled Cartesian direction of arrival (ACCDOA) output format [29]. The ACCDOA is an output format that assigns the location of acoustic events based on their activity, thus providing simultaneously both the location information and active status of the event, thus enabling the performance of SELD without any weight-adjustment difficulties using a single model. However, the ACCDOA output format has the disadvantage of not detecting the occurrence of acoustic events of the same class simultaneously because it assigns one location to one class. To address this disadvantage, Kazuki et al. proposed a multi-ACCDOA output format [30], which extended the ACCDOA to a track format and enabled the detection of multiple acoustic events of the same class. Based on the proposed output format, a SELD model that detected acoustic events of the same class (that appeared simultaneously) could be constructed using a single model.

Furthermore, to enhance the performance of SELD, research has been conducted on the input features during the DCASE Challenge. In the first challenge, held in 2019, phase and magnitude spectrograms were primarily used as input features, regardless of the type of audio data and whether first-order ambisonic (FOA) or MIC were used [28], [31]. Additional features, such as log-mel spectrograms and intensity vectors (IVs), were also used [32]. Performance differences occur even when using the same input features depending on the audio type. Therefore, a method was proposed during the 2020 competition to use appropriate input features according to the audio format. Regarding the MIC format, mel spectrograms and generalized cross-correlation (GCC) were used as input features; regarding the FOA format, mel spectrograms and IVs were used as input features [33]. However, even when different input features were used according to the audio format, the SELD performance was relatively low when the MIC format data were used compared with the FOA format. As an improvement measure, Nguyen et al. proposed input features that could achieve high performance for both the FOA and MIC formats. They proposed a spatial cue-augmented log-spectrogram (SALSA) input feature using a log-linear spectrogram and principal eigenvector [34]. When using a SALSA as an input feature, a performance improvement was confirmed compared with the results obtained using the log-mel spectrogram and GCC in the MIC format. Nguyen et al. proposed a lightweight SALSA-Lite, which changes the SALSA using eigenvector-based spatial features for polyphonic SELD and uses the normalized interchannel phase difference as a spatial feature [35].

B. SPEECH ENHANCEMENT

Speech enhancement improves the quality of voice signals by amplifying the speaker's voice or eliminating background noise, thereby enabling the extraction of the desired signal from the audio. In other words, when a signal y[n] is composed of the sum of the noise-free audio signal x[n] and noise signal n[n], the goal is to restore y[n] to a form similar to x[n]. Methods such as the Wiener [36], matched [37], and Kalman filters [38] have been applied to achieve speech enhancement. Recently, with the advancement of deep learning technologies, various deep-learning-based speech enhancement techniques have received attention. Deep learning technologies, which demonstrate superior performance, have facilitated continuous research efforts into various speech enhancement applications.

One of the most common approaches for deep-learning-based speech enhancement is noise removal. This method identifies and removes the noise part from an audio signal while preserving and restoring only the desired signal. A denoising autoencoder (DAE) is typically used for this purpose [39]. The DAE takes a noisy audio signal as input, encodes it, and then decodes the encoded feature map to restore it to a noise-free signal. The model is trained to minimize the difference between the noise-free signal and restored output. Through this process, the model automatically learns the noise filtering process.

A representative model structure used to perform this process is U-Net [40]. U-Net is a model with encoder and decoder structures, wherein skip connections are applied to combine the intermediate feature maps of the encoder and decoder. Originally developed for segmentation tasks, the U-Net structure, composed of an encoder and decoder, can also be applied to speech enhancement tasks. Various models based on U-Net have been studied, including the Nested U-Net [41] and Wave-U-Net [42], and studies on speech enhancement using these modified models have been conducted. Xiang et al. applied self-attention to the Nested U-Net structure, which contains skip connections at each stage of the encoder and decoder, to perform speech enhancement using contextual information at various scales [43]. Macartney and Weyde used Wave-U-Net, which applies U-Net to a one-dimensional time domain, to separate vocals and accompaniments in music for speech enhancement [44]. Furthermore, various deep-learning-based speech enhancement studies continue to be conducted using various methods.

III. PROPOSED METHOD

Conventional approaches to perform SELD involve separate noise removal and SELD. However, this leads to the repetition of the feature extraction and model training steps for each task, thus resulting in a decrease in temporal efficiency. To address this issue, this study proposes a model structure that combines speech enhancement and SELD models to perform noise removal and SELD simultaneously. In this study,

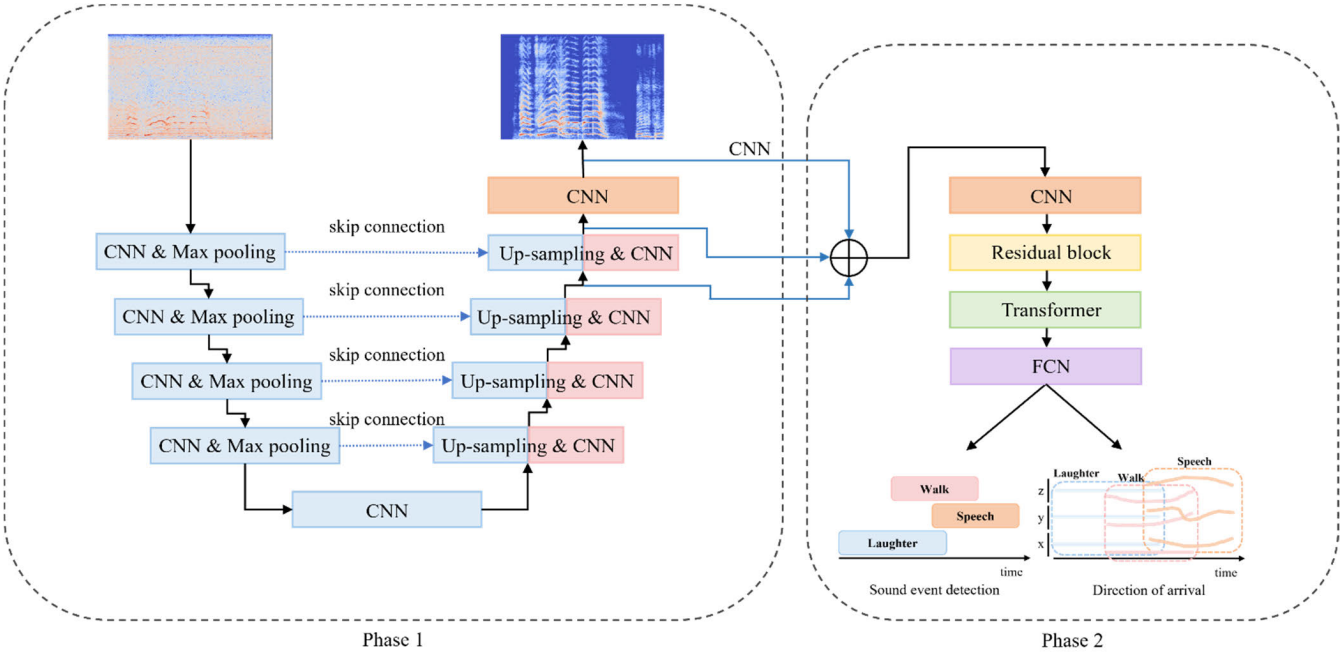

FIGURE 2. Proposed SELD U-Net architecture.

we combine U-Net for speech enhancement and SELDnet for SELD, and we refer to this model as SELD U-Net. The structure of SELD U-Net is shown in Figure 2. The process can be divided into two phases. Phase 1 corresponds to the speech enhancement process, and Phase 2 corresponds to the SELD process. The proposed model was trained separately during each phase. Training for Phase 1 was completed before proceeding to Phase 2. During the training process of Phase 2, the weights learned in Phase 1 were fixed, and only the model specific to that phase was trained.

In SELD U-Net, some of the feature maps were extracted from the U-Net decoder and used as input features. The U-Net encoder removes noise from the extracted input features of the audio signal in a noisy environment and extracts essential audio information. During the decoding process, the input features extracted from the denoised signal are reconstructed and the denoised input features are inferred. In this process, the feature maps in the decoder contain crucial information when the noise is removed, and they are used as input features for SELDnet. In this case, the feature maps from the decoder (used as input features for SELDnet) have different data shapes. Therefore, convolutional operations are applied to adjust the shapes of the feature maps, which are then combined into the input features.

A. U-Net

The proposed model incorporates three types of U-Net to construct the SELD U-Net, and the performances of all the models are evaluated and compared. The U-Net models used in this study include U-Net [40], Nested U-Net [41], and ResU-Net [45], as depicted in Figures 3, 4, and 5, respectively. All three U-Nets use a 7-channel input feature (obtained by combining 4-channel mel spectrograms and 3-channel IVs) extracted from noisy audio data. This input feature is the same as that used in the existing SELDnet. In U-Net, training is conducted to infer the denoised input features from the input features. The feature maps from the decoder part of the trained U-Net are combined and used as input features for SELDnet. The feature maps from the U-Net decoder contain essential information without noise, which is expected to improve the SELD performance.

As shown in Figure 3, U-Net consists of five layers in depth. In the encoder, the input feature passes through blocks composed of convolution layers (kernel size $= 3 \times 3$), batch normalization, and a rectified linear unit (ReLU). During this process, the number of channels of the input feature increases to 64, 128, 256, 512, and 1024. Subsequently, in the decoder, restoration is performed using skip connections by using feature maps from the encoder.

The structure of the Nested U-Net, which consists of five layers, is shown in Figure 4. In the encoder, downsampling is performed using convolution layers (kernel size $= 3 \times 3$), batch normalization, and ReLU. In the decoder, deconvolution layers are applied to the encoder features to upsample them back to their original input size. The difference between Nested U-Net and U-Net is the addition of an additional encoder–decoder block between the encoder and decoder to implement nesting in the former. Based on this nesting structure, the model achieves higher levels of abstraction and better restoration of detailed information.

The structure of ResU-Net consists of four layers, as shown in Figure 5. Before going through the convolution layer

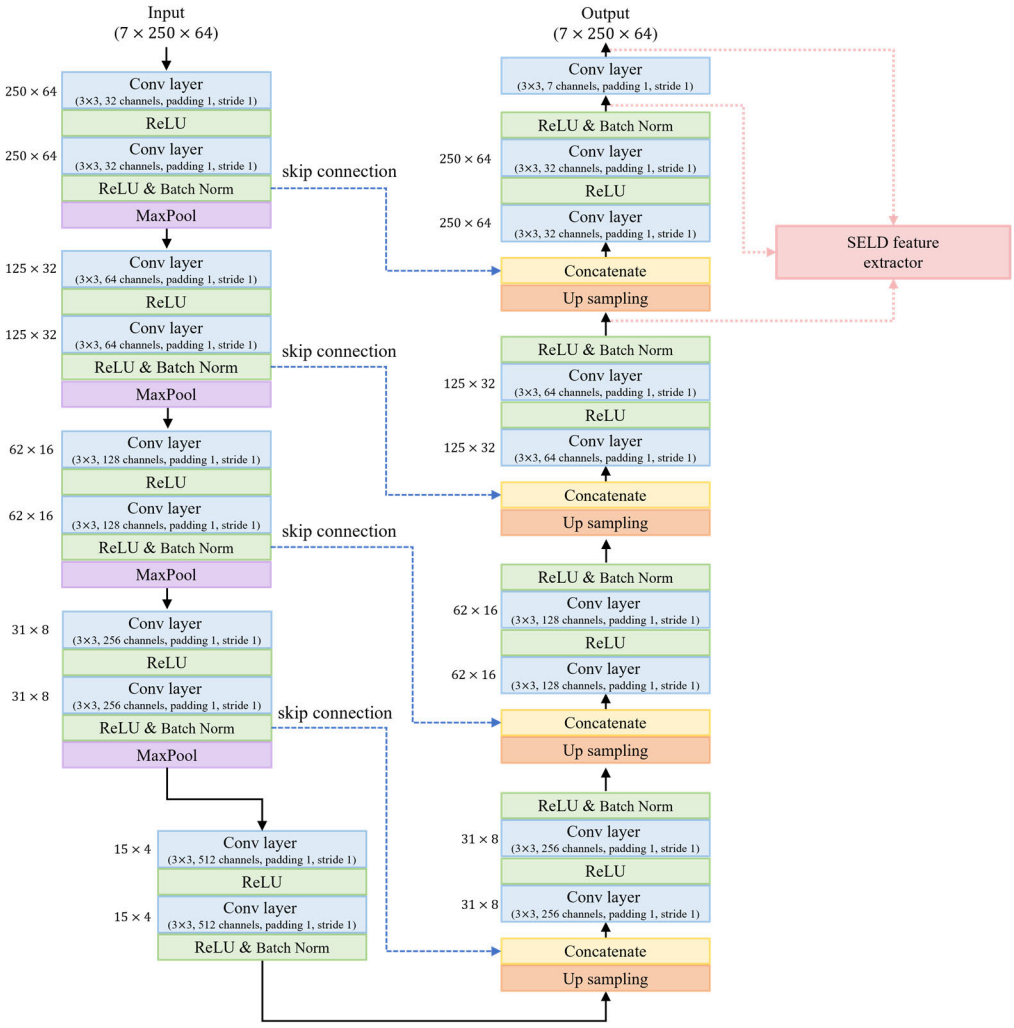

FIGURE 3. U-Net model architecture used in SELD U-Net.

for downsampling, the input features pass through a convolution layer that changes their shapes for residual connection. Subsequently, residual connections are performed in both the encoder process (for downsampling) and decoder process (for upsampling). These residual connections enable the construction of a deep network while addressing the problem of gradient vanishing, thus resulting in more accurate inference results.

To compare the performance based on the number of parameters, we evaluated the parameter counts of the three models. The model utilizing U-Net has 8.836M parameters, the Nested U-Net has 10.061M parameters, and the Res U-Net has 9.121M parameters.

The U-Net model employs the L1 loss function, and the losses for the mel spectrograms and IVs are calculated separately and combined for training. Based on the training results of U-Net, the restoration of IVs was less successful than that of the mel spectrograms. To address this issue, a weighting factor was assigned to IV loss during the training process. The loss function used for U-Net training is given by (1),

$$Loss = (\sum_{i=1}^{S} |Spec_i - \widehat{Spec}_i|) + (\sum_{i=1}^{S} |IV_i - \widehat{IV}_i|) * \lambda, \tag{1}$$

where $Spec_i$ and IV_i represent the mel spectrogram and IV of the noise-free audio, respectively, whereas $\widehat{Spec}_i$ and $\widehat{IV}_i$ refer to the mel spectrogram and IV obtained after performing speech enhancement on the noisy audio. S denotes the number of samples, and λ represents the weighting factor for the loss associated with the IV. In this study, a value of 10 was applied to λ during the training process. In this study, the speech enhancement was conducted using a learning approach commonly applied in 2D image enhancement. Specifically, the loss was calculated by measuring the differences between two images. Since the noise in the audio samples used in the experiments was evenly distributed without any specific outliers, we opted for the L1 loss function, which is less sensitive to outliers, for our calculations.

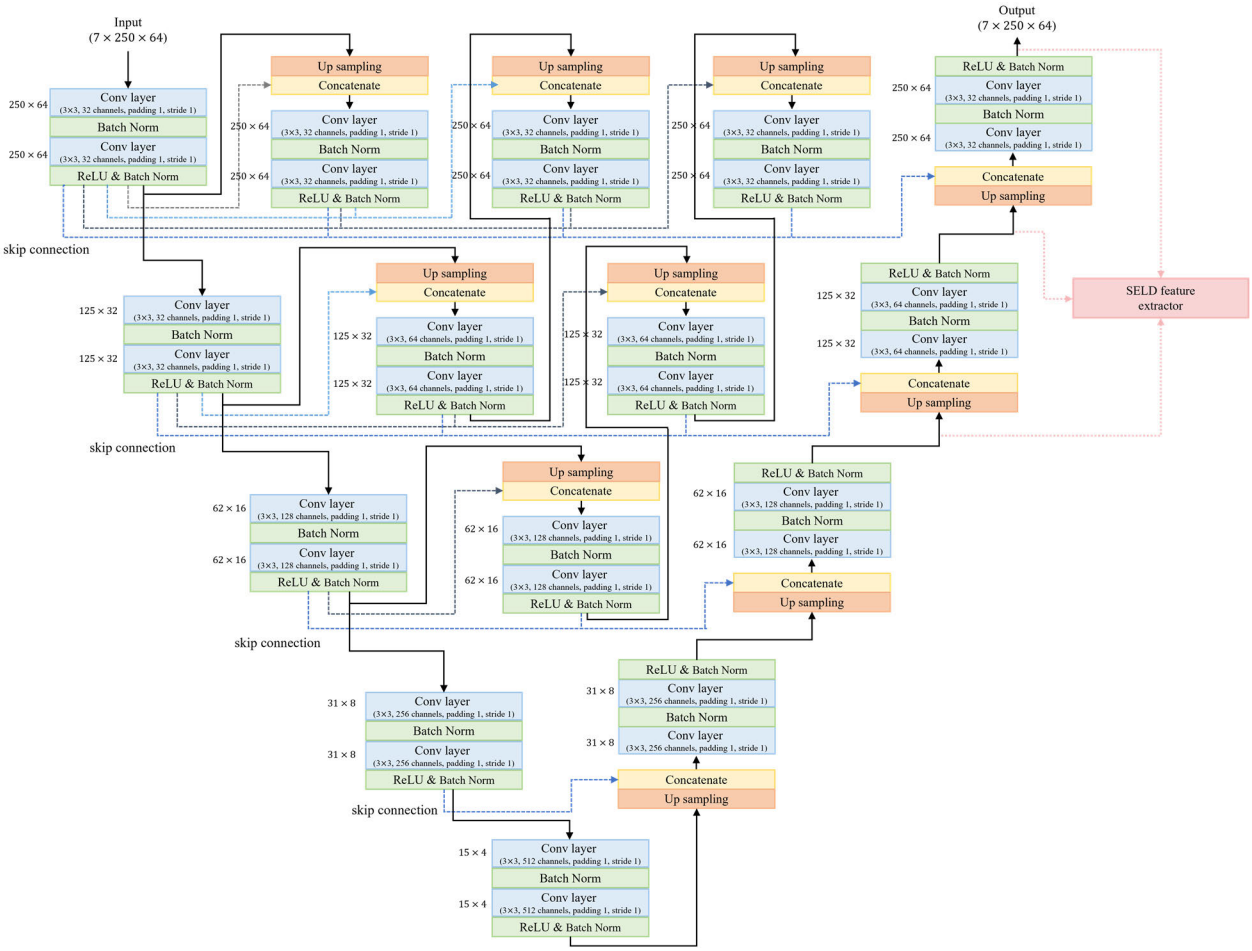

FIGURE 4. Nested U-Net model architecture used in SELD U-Net.

B. SELDnet

SELDnet, which performs SELD in the proposed SELD U-Net, is shown in Figure 6. SELDnet uses a part of the feature maps from the U-Net decoder as input features. To prepare feature maps for use as input features for SELDnet, a convolution operation is applied to each feature map to ensure that they have the same shape after being extracted and combined. SELDnet is designed based on a convolutional recurrent neural network (CRNN) structure [28]. The CRNN-based SELDnet has CNN and RNN layers, where the CNN extracts audio signal features and the RNN models the temporal information flow. In the proposed model, the CNN layers of SELDnet are replaced with residual convolutional blocks (RCB), and the RNN layers are replaced with transformer encoders.

The RCB is a structure derived from the residual network (ResNet) [46] that uses residual blocks that add input values to the output values. In the proposed model, the RCB consists of two convolution layers followed by batch normalization and an activation function. The output values from these layers are added to the input values of the block. The final output is then obtained by passing the result through the average pooling and dropout layers. This structure can address the vanishing gradient problem, thereby enabling the construction of deep neural network models. Furthermore, because the input values are directly carried over and only the remaining information is learned, convergence becomes easier and faster during training compared with the training of the entire input. Three RCB are used in the proposed model.

The transformer encoder is a sequence model that preserves the positional information of the input data and simultaneously learns the relationships among all the positions in the input sequence [47]. The transformer encoder consists of three main components: positional encoding, multihead self-attention, and feedforward blocks. In the positional encoding block, each element of the input sequence is assigned positional information, which is typically implemented using periodic functions (e.g., sine or cosine functions). The multihead self-attention block then performs multiple rounds of attention operations on the input data to capture various representations of the input. Each attention head has an independent attention mechanism that enables it to perform attention operations on each input using different sets of weights. The outputs generated by each head are combined

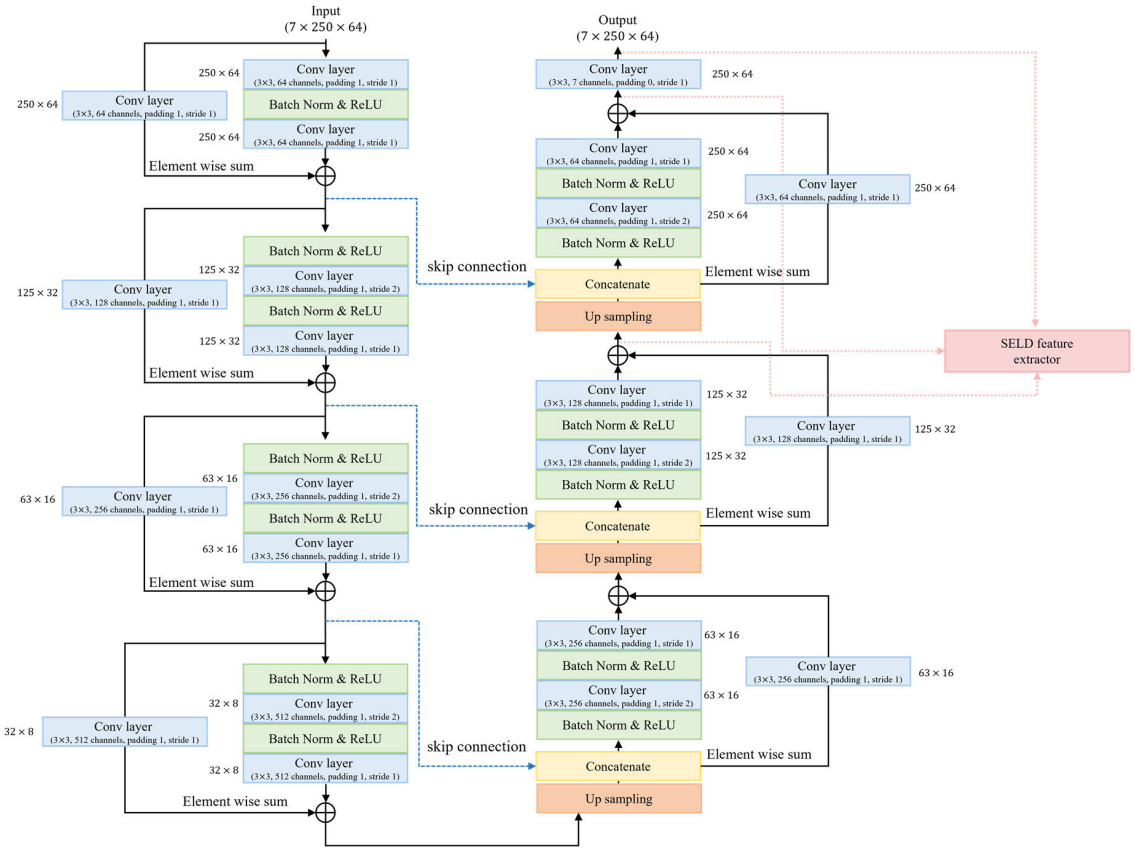

FIGURE 5. ResU-Net model architecture used in SELD U-Net.

to produce the final output. Through these operations, the transformer encoder can focus on the same data from different perspectives, thereby capturing complex and diverse meanings. Because the output of the multihead self-attention block is a combination of outputs from each head, an information fusion process is required. For this purpose, the feedforward block receives the output of the multihead self-attention block and performs a linear transformation. By performing linear transformations, the model can integrate individual output values and learn the patterns in the data. Compared with the RNN, the transformer encoder has the advantage of being able to process input sequences in parallel as it handles them simultaneously. In addition, by using the self-attention mechanism to learn the relationships among all elements of the input sequence, the transformer encoder can address the issue of long-term dependences. In the proposed model, a single layer of the transformer encoder was used with four attention heads and a feed-forward layer (dimension = 512).

Finally, an FC layer was used to generate the final prediction values. The proposed model performed both SED and SSL tasks using a single model and adopted a multi-ACCDOA output format to detect multiple events in the same class [30]. The multi-ACCDOA is an extended output format of the ACCDOA, which assigns the location of acoustic events based on their activity status. Each track corresponding to a class has a target vector representing the activity and location of an acoustic event. The empty tracks of each class are assigned to one of the duplicated target vectors from the other tracks; in this way, all possible permutations are considered to determine the best permutation. Subsequently, all possible combinations of permutations are generated, and the loss is computed for each permutation to determine the optimal permutation. This output format enabled the model to consider various permutations and compute the loss for each permutation. Because the dataset used in this study allowed a maximum of three overlapping sound sources, the number of tracks was set to three. Matrix P, which represents this configuration, is defined by (2),

$$P \in \mathbb{R}^{3\times N\times C\times T}, \tag{2}$$

where N, C, and T represent the number of tracks, number of classes, and frame index, respectively, and the number 3 denotes the x-, y-, and z-axes.

In the multi-ACCDOA format, the proposed auxiliary duplicating permutation invariant training (ADPIT) method is applied, and the mean-squared error (MSE) loss is computed according to (3) and (4).

$$\mathcal{L}^{PIT} = \frac{1}{CT}\sum_{c}^{C}\sum_{t}^{T} \min_{\alpha\in Perm(ct)} l_{\alpha,ct}^{ACCDOA}, \tag{3}$$

$$l_{\alpha,ct}^{ACCDOA} = \frac{1}{N}\sum_{n}^{N} MSE(P_{\alpha,nct}^{*}, \widehat{P}_{nct}), \tag{4}$$

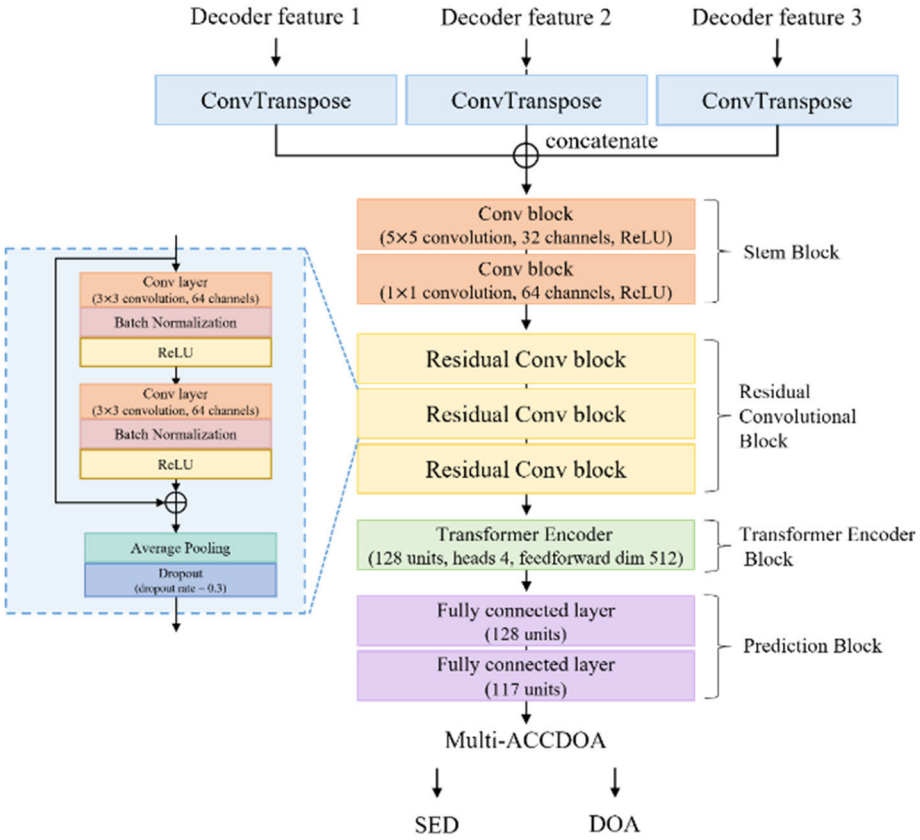

FIGURE 6. SELDnet model architecture used in SELD U-Net.

where C represents the class, and T represents the frame. *Perm* denotes the set of all possible permutations owing to the permutations of classes, and $\alpha \in Perm(ct)$ refers to one of the permutations for class c at frame t. The class-wise ADPIT is performed according to (3) to calculate the loss for each permutation. The permutation with the lowest loss is used as the output of the model. The formula for calculating the loss between the predicted permutation from the model and the target permutation is given in (4). $P^*_{\alpha,nct}$ represents the ACCDOA target for permutation α, and $\widehat{P}_{nct}$ represents the ACCDOA prediction for track n, class c, and frame t. The model training results in the output of the shape (batch, sequence length, class $\times$ 3 $\times$ 3), thus indicating that the information for the x, y, and z coordinates of each class is the output for the three tracks.

C. FEATURE EXTRACTION AND EXPERIMENTAL SETUP

To use the data effectively, extracting the appropriate features is important. In this study, the input features were extracted from a 4-channel FOA B-format. A total of 7-channel input features were used (which consisted of 4-channel mel spectrograms and 3-channel IVs stacked together). The hop and window lengths were set to 512 and 1024, respectively, and a Hanning window was used. The short-time Fourier transform (STFT) size was set to 1024 to extract the spectrograms, and a mel filter bank was applied to obtain 4-channel mel spectrograms. The IV was calculated from the 4-channel spectrogram obtained by applying the STFT, as described in (5) [48],

$$I_{t,f,ch} = \begin{bmatrix} \Re(W^*_{f,t} \circ X_{f,t}) \\ \Re(W^*_{f,t} \circ Y_{f,t}) \\ \Re(W^*_{f,t} \circ Z_{f,t}) \end{bmatrix}, \tag{5}$$

where $W, X, Y,$ and Z represent the ambisonic channels, W denotes the omnidirectional channel, and $X, Y,$ and Z represent the channels corresponding to the x-, y-, and z-axes, respectively. f denotes the frequency bin index, ch represents the channel index, * denotes complex conjugation, R denotes the real part, and $\circ$ denotes element-wise multiplication.

Overfitting during model training can lead to the degradation of model performance. A data-frequency masking augmentation technique was applied to prevent overfitting. As shown in Figure 7, frequency masking involves masking certain frequency channels of the spectrogram by completely removing the frequency information within random regions. During model training, frequency masking was applied to the entire sequence, and 8 of the 64 mel bins were masked. Additionally, in Phase 1 of training the U-Net, the L1 loss function was used, while in Phase 2, where the SLEDnet was trained, the MSE loss function was employed for training.

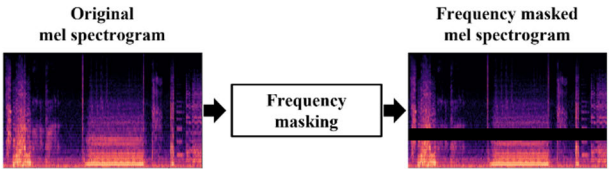

FIGURE 7. Example of frequency masking application.

All other training parameters were set identically. The model was trained using the Nesterov momentum Adam optimizer [49] with a learning rate of 0.001, and the batch size was set to 128. Both models were trained for a total of 1000 epochs.

In addition, the model was trained using the MSE loss function and the Nesterov momentum Adam optimizer [49] with a learning rate of 0.001. Training was conducted for 1000 epochs.

IV. EXPERIMENTS AND RESULTS

Objective evaluations were conducted to validate the performance of the proposed model. For this purpose, in the same manner as in the DCASE 2022 Challenge, the data used for model training were generated using synthetic data. The synthetic data comprised pairs of noisy data at various signal-to-noise ratios (SNRs) and noise-free data. The Freesound dataset 50 k (FSD50K) was used for the synthetic process [50]. In this study, approximately 300 sound events were used for each class to generate the training data. In addition, 100 separate sound events were used to construct the evaluation dataset. The audio data in the generated dataset were in the FOA format with four channels, each with a length of 1 min and a sampling rate of 24 kHz. To evaluate performance at various noise levels, noise data were generated using different SNR values. To verify robust performance in noisy environments, training and evaluation are conducted using datasets that include audio with lower SNR values than those typically used for standard SELD performance evaluation. The SNR values were set to +30, +20, +10, −10, −20, and −30. Each dataset, categorized by SNR, consisted of 1200 samples for training and 300 samples for evaluation.

To evaluate the performance of the proposed model, an objective performance evaluation was conducted using evaluation metrics for speech enhancement and SELD. The speech-enhancement module of the proposed SELD U-Net yields two-dimensional images representing the mel spectrograms and IVs. Thus, evaluation metrics commonly used for image quality assessment, such as MSE, structural similarity index (SSIM) [51], and peak SNR (PSNR), were applied to assess the performance of the model.

The MSE is an evaluation metric used in regression analysis, image processing, and signal processing. It measures the average squared difference between the predicted and actual values. Because the MSE squares the differences, it is sensitive to outliers, and larger errors result in higher MSE values. A lower MSE value indicates that the predicted values are closer to the original values. The MSE was calculated using (6),

$$MSE = \frac{1}{S}\sum_{i=1}^{S}(Y_i - \widehat{Y}_i)^2, \tag{6}$$

where S represents the number of elements in Y and $\widehat{Y}$, Y_i represents the actual values, and $\widehat{Y}_i$ represents the predicted values. SSIM is a method used to quantify digital image quality by quantitatively evaluating the structural similarity between two images. It considers three components—luminance, contrast, and structure—to measure the structural changes between the two images. The SSIM was calculated using (7),

$$SSIM\left(Y, \widehat{Y}\right) = \frac{(2\mu_Y\mu_{\widehat{Y}} + c_1)(2\sigma_{Y\widehat{Y}} + c_2)}{(2\mu_Y^2 + \mu_{\widehat{Y}}^2 + c_1)(\sigma_Y^2 + \sigma_{\widehat{Y}}^2 + c_2)}, \tag{7}$$

where Y represents the mel spectrogram and IV of the noise-free audio data, whereas $\widehat{Y}$ represents the mel spectrogram and IV after speech enhancement. μ_Y and σ_Y denote the mean and variance of Y, respectively, whereas $\mu_{\widehat{Y}}$ and $\sigma_{\widehat{Y}}$ denote the mean and variance of $\widehat{Y}$, respectively. $\sigma_{Y\widehat{Y}}$ represents the covariance between Y and $\widehat{Y}$, $c_1 = (K_1L)^2$ and $c_2 = (K_2L)^2$ are constants used for stability, where L represents the maximum range of energy values. Typically, $K_1 = 0.01$ and $K_2 = 0.03$.

PSNR is an evaluation metric used to quantify the reconstruction error in two-dimensional signals, such as images or videos. It is commonly used to assess the quality of the original signal and its noisy counterpart; higher PSNR values indicate higher quality. The PSNR was calculated using (8),

$$PSNR = 10log_{10}\left(\frac{R^2}{MSE}\right), \tag{8}$$

where MSE represents the mean squared error between Y and $\widehat{Y}$, and R represents the maximum value of the magnitude of the mel spectrogram and IV. The PSNR was measured in decibels (dB), and the pixel values of the mel spectrogram and IV represent the energy values of the audio signal.

The proposed SELD U-Net SELDnet performs the SED and DOA estimations. To evaluate the performance of the SELD, the SELD evaluation metrics proposed by DCASE were used [52]. To assess the SED performance, the error rate and F1-score were employed as evaluation metrics. The localization error and localization recall were used as evaluation metrics to quantify the SSL performance. Furthermore, the SELD score, which considers all four evaluation metrics, was used for comprehensive evaluations.

The error rate was computed by measuring the total number of substitutions (*Sub*), insertions (*Ins*), and deletions (*Del*) relative to the number of activated events (*EN*) in the reference. *Sub* represents the cases in which the system output incorrectly labels an event as active, *Ins* represents false positives (excluding *Sub*), and *Del* represents false negatives

(excluding *Sub*). *Sub*, *Ins*, and *Del* are defined as follows:

$$Sub\,(t) = min\,(FN(t), FP(t))\,, \tag{9}$$
$$Del\,(t) = max\,(0, FN(t) - FP(t))\,, \tag{10}$$
$$Ins\,(t) = max\,(0, FP(t) - FN(t))\,, \tag{11}$$

where t denotes a frame. The error rate was calculated by summing all segments and dividing the sum by the total number of frames T. This was computed as follows:

$$Error\ rate = \frac{\sum_{t=1}^{T} Sub\,(t) + \sum_{t=1}^{T} Del\,(t) + \sum_{t=1}^{T} Ins\,(t)}{\sum_{t=1}^{T} EN\,(t)}. \tag{12}$$

The F-score was calculated using precision (P) and recall (R), and the F1-score was calculated as follows:

$$P = \frac{\sum_{t=1}^{T} TP\,(t)}{\sum_{t=1}^{T} TP\,(t) + \sum_{t=1}^{T} FP\,(t)}, \tag{13}$$
$$R = \frac{\sum_{t=1}^{T} TP(t)}{\sum_{t=1}^{T} TP(t) + \sum_{t=1}^{T} FP(t)}, \tag{14}$$
$$F = \frac{2P \cdot R}{P + R}, \tag{15}$$

where $TP(t)$, $FP(t)$, and $FN(t)$ represent true positives, false positives, and false negatives, respectively, in the t-th frame. For the location-dependent SED evaluation, if the angle difference between the predicted location of the system and the ground truth location was within 20°, it was considered a *TP*. Otherwise, it was considered as a *FN*.

The localization error and localization recall are class-dependent localization metrics. Class-aware localization error (LE_c) and class-aware localization recall (LR_c) were calculated for all predicted and reference events. LE_c and LR_c were calculated as follows:

$$LE_c = \frac{||A_c \odot D_c||_1}{||A_c||_1}, \tag{16}$$
$$LR_c = \frac{\sum_l ||A_c^{(l)}||_1}{\sum_l N_c^{(l)}}, \tag{17}$$

where $c \in [1, \cdots, C]$ represents the class index, and $t = 1, \cdots, T$ refers to each temporal partition of the data if data are divided over time. D_c represents the $M_c \times N_c$ distance matrix, where M_c denotes the number of predictions for class c, and N_c denotes the number of reference events. A_c is the association matrix for matrix D_c, denoted as $\mathcal{H}(D_c)$, where $\mathcal{H}(\cdot)$ represents the Hungarian algorithm. Based on the consideration of all predicted events, LE_c and LR_c, we can calculate LE_{CD} and LR_{CD}, which consider only events belonging to the same class. LE_{CD} and LR_{CD} are expressed by (18) and (19), respectively.

$$LE_{CD} = \frac{1}{C \cdot T} \sum_c \sum_l LE_c^{(t)}, \tag{18}$$
$$LR_{CD} = \frac{1}{C} \sum_c LR_c. \tag{19}$$

Finally, to represent the overall performance of the neural network model, we used the SELD score, which combines the evaluation metrics for SED (error rate, F1-score) and SSL (localization error, localization recall). The SELD score was calculated using (20),

$$SELD\ score = \frac{ER + (1 - F) + (LE/180) + (1 - LR)}{4}, \tag{20}$$

where ER is the error rate, F is the F1-score, LE is the localization error, and LR is the localization recall.

Two comparative models were used to evaluate the performance of the proposed SELD U-Net. The first is the baseline model of the DCASE 2022 Challenge, which is based on the CRNN architecture. The second model corresponds to the SELDnet of the proposed SELD U-Net. Both models use the multi-ACCDOA output format, similar to the proposed model, and use 7-channel input data (consisting of 4-channel mel spectrograms and 3-channel IVs) extracted from noisy audio data. By comparing these two models, we could assess the performance of the proposed model, compare it with a general SELD model, and examine the impact of the noise-reduction process on the SELD performance.

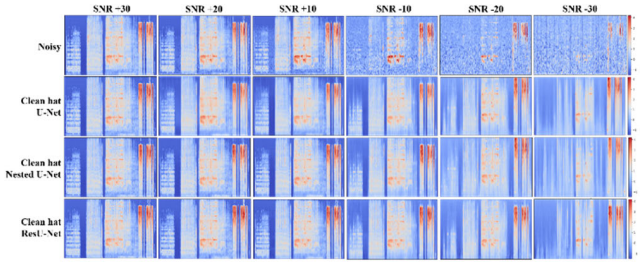

FIGURE 8. Speech enhancement results (mel spectrograms).

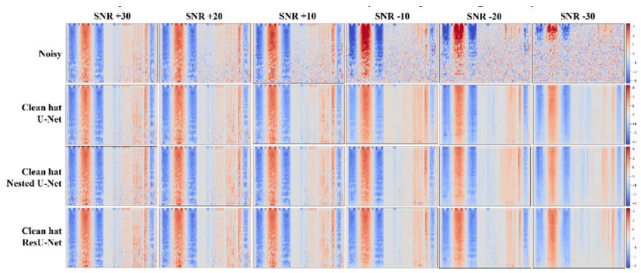

FIGURE 9. Speech enhancement results (intensity vectors).

Figure 8 shows the results of speech enhancement on mel spectrograms using U-Net, Nested U-Net, and ResU-Net, whereas Figure 9 shows the results of speech enhancement on the IV. A visual comparison of the inference results of the models indicates that the restoration results of U-Net, Nested U-Net, and ResU-Net were similar. A comparison of the restored mel spectrograms and IVs with the noise-free state reveals that restoration was performed effectively for

SNR values equal to +30, +20, and +10. However, for SNRs of −20 and −30, the mel spectrograms successfully detected features in high-magnitude regions but could not capture details in low-magnitude regions. These findings indicate that speech enhancement was achieved to some extent in noisy environments but the restoration of fine-grained details was not effectively achieved. Similarly, the IV performs an approximate restoration in noisy environments but lacks fine-grained detail.

TABLE 1. Evaluation results for speech enhancement (SNR = +30, +20, +10).

		Spectrogram			Intensity vector		
		Mean-squared error (MSE)	Structural similarity index (SSIM)	Peak signal-to-noise-ratio (PSNR)	MSE	SSIM	PSNR
SNR +30	Noise	0.0	1.0	Infinity	0.0	1.0	Infinity
	U-Net	0.041	0.639	62.066	0.043	0.861	61.901
	Nested U-Net	0.001	0.979	77.563	0.001	0.999	77.170
	ResU-Net	1e-07	0.999	114.02	6e-07	0.999	113.41
SNR +20	Noise	0.047	0.782	61.885	0.040	0.543	62.501
	U-Net	0.052	0.642	61.312	0.048	0.631	61.641
	Nested U-Net	0.009	0.895	69.111	0.008	0.722	69.483
	ResU-Net	0.008	0.907	69.609	0.007	0.712	70.032
SNR +10	Noise	0.142	0.553	56.983	0.124	0.380	57.549
	U-Net	0.066	0.596	60.100	0.065	0.492	60.146
	Nested U-Net	0.023	0.801	64.867	0.022	0.570	65.112
	ResU-Net	0.021	0.798	65.201	0.020	0.563	65.465

Quantitative performance comparisons using MSE, SSIM, and PSNR are presented in Tables 1 and 2. Table 1 lists the results for the SNR levels of +30, +20, and +10, whereas Table 2 presents the results for SNR levels of −10, −20, and −30. The results in Tables 1 and 2 indicate that the ResU-Net model exhibited the best performance in enhancing speech in various noisy environments.

In the case of SNR +30, the ResU-Net model demonstrated the best performance in six metrics, whereas the Nested U-Net model achieved high performance, with an SSIM score of 0.999 for the IV. When SNR = +20, the Nested U-Net model exhibited the best performance in terms of the SSIM metric for the IV, whereas the ResU-Net model outperformed other metrics. In the case of SNR = +10, the Nested U-Net model achieved the best performance in the SSIM metric for both the mel spectrogram and IV, whereas the ResU-Net model yielded the best performance for the other metrics.

TABLE 2. Evaluation results for speech enhancement (SNR = −10, −20, −30).

		Spectrogram			Intensity vector		
		MSE	SSIM	PSNR	MSE	SSIM	PSNR
SNR +30	Noise	0.732	0.137	49.863	0.663	0.091	50.267
	U-Net	0.178	0.438	55.924	0.171	0.192	56.069
	Nested U-Net	0.176	0.433	55.951	0.168	0.160	56.109
	ResU-Net	0.133	0.505	57.228	0.126	0.265	57.443
SNR +20	Noise	1.155	0.051	47.879	1.090	0.028	48.123
	U-Net	0.344	0.290	53.081	0.333	0.096	53.177
	Nested U-Net	0.358	0.267	52.826	0.344	0.071	52.966
	ResU-Net	0.283	0.351	53.847	0.277	0.166	53.944
SNR +10	Noise	1.485	0.013	46.744	1.448	0.005	46.858
	U-Net	0.660	0.150	50.267	0.639	0.054	50.357
	Nested U-Net	0.650	0.126	50.277	0.640	0.044	50.319
	ResU-Net	0.584	0.198	50.742	0.583	0.106	50.753

For the cases where SNR was equal to −10, −20, and −30, the ResU-Net model yielded the best performance for all six metrics, whereas the Nested U-Net and U-Net models exhibited similar performance levels.

From these experiments, we observed that the ResU-Net model achieved the best performance in environments with minimal noise and noisy environments. The Nested U-Net and U-Net models exhibited relatively lower performance than the ResU-Net model. The proposed U-Net model exhibited a better performance overall when noise data and numerical values in noisy environments were compared. This confirms that the structure of the proposed SELD U-Net model can enhance speech. These findings suggest that the proposed SELD U-Net model with its U-Net structure can effectively reduce noise, thus leading to improved performance in SELD tasks.

Tables 3, 4, and 5 present a performance comparison of the SELD models in environments with different SNRs (equal to +30, +20, and +10). In low-noise environments, the proposed models exhibited superior performance compared with the CRNN model. However, they fell short of the performance achieved by the models that incorporated residual blocks and transformer encoders.

When comparing the performances of the proposed models, the Nested U-Net model yielded the best performance for SNR = +30 and +20, whereas the ResU-Net model performed the best for SNR = +10. In noise-free environments, the models demonstrated good performance even without separate speech enhancement. Therefore, the proposed models that involve additional noise removal processes exhibit

TABLE 3. Sound event localization and detection results for SNR = +30.

Model	Error rate	F1-score	Localization error	Localization recall	Sound event localization and detection (SELD) score
CNN + Bi-GRU	0.71	0.30	17.66	0.37	0.54
Residual block + Transformer	0.66	0.40	13.51	0.51	0.46
SELD U-Net (U-Net)	0.71	0.32	17.28	0.34	0.51
SELD U-Net (Nested U-Net)	0.69	0.35	15.64	0.46	0.49
SELD U-Net (ResU-Net)	0.70	0.34	16.69	0.46	0.50

TABLE 4. Sound event localization and detection results for SNR = +20.

Model	Error rate	F1-score	Localization error	Localization recall	SELD score
CNN + Bi-GRU	0.72	0.30	17.45	0.36	0.54
Residual block + Transformer	0.69	0.38	14.13	0.48	0.48
SELD U-Net (U-Net)	0.73	0.31	17.36	0.43	0.52
SELD U-Net (Nested U-Net)	0.71	0.32	17.80	0.44	0.51
SELD U-Net (ResU-Net)	0.72	0.31	17.99	0.44	0.52

higher complexity and may have a higher risk of overfitting owing to excessive data modification. For these reasons, the proposed SELD U-Net model may result in a lower performance than the SELDnet model without noise removal in noise-free environments.

Tables 6, 7, and 8 present performance comparison outcomes at the SNR levels of −10, −20, and −30, respectively.

TABLE 5. Sound event localization and detection results for SNR = +10.

Model	Error rate	F1-score	Localization error	Localization recall	SELD score
CNN + Bi-GRU	0.75	0.27	17.18	0.35	0.56
Residual block + Transformer	0.70	0.37	15.06	0.47	0.49
SELD U-Net (U-Net)	0.73	0.29	18.35	0.39	0.54
SELD U-Net (Nested U-Net)	0.72	0.31	17.53	0.41	0.52
SELD U-Net (ResU-Net)	0.71	0.34	15.88	0.44	0.50

TABLE 6. Sound event localization and detection results for SNR = −10.

Model	Error rate	F1-score	Localization error	Localization recall	SELD score
CNN + Bi-GRU	0.84	0.17	23.20	0.24	0.63
Residual block + Transformer	0.78	0.27	18.14	0.35	0.57
SELD U-Net (U-Net)	0.79	0.25	18.66	0.33	0.58
SELD U-Net (Nested U-Net)	0.80	0.24	20.11	0.35	0.58
SELD U-Net (ResU-Net)	0.78	0.26	18.98	0.34	0.57

In the relatively low-noise environment of SNR −10, the models incorporating ResU-Net with the residual block and Transformer encoder demonstrated the best performance. For the case where SNR = −20, the models employing U-Net and Nested U-Net yielded the best performance, whereas for the case where SNR = −30, the models using U-Net and ResU-Net performed the best. In contrast to noise-free environments, the proposed models generally outperformed existing SELD models. These findings indicate that the proposed models are more robust to noisy environments, and demonstrate significant results in terms of robustness under high-noise conditions.

TABLE 7. Sound event localization and detection results for SNR = −20.

Model	Error rate	F1-score	Localization error	Localization recall	SELD score
CNN + Bi-GRU	0.79	0.11	27.10	0.17	0.69
Residual block + Transformer	0.87	0.16	22.81	0.22	0.65
SELD U-Net (U-Net)	0.86	0.19	19.41	0.24	0.63
SELD U-Net (Nested U-Net)	0.85	0.19	22.25	0.26	0.63
SELD U-Net (ResU-Net)	0.85	0.18	21.54	0.24	0.64

TABLE 8. Sound event localization and detection results for SNR = −30.

Model	Error rate	F1-score	Localization error	Localization recall	SELD score
CNN + Bi-GRU	0.97	0.05	43.88	0.10	0.76
Residual block + Transformer	0.94	0.07	65.72	0.13	0.78
SELD U-Net (U-Net)	0.95	0.09	26.86	0.13	0.72
SELD U-Net (Nested U-Net)	0.94	0.09	36.42	0.12	0.73
SELD U-Net (ResU-Net)	0.94	0.08	27.91	0.12	0.72

V. CONCLUSION

In this study, we proposed a SELD U-Net model that combined a U-Net model for speech enhancement with a SELDnet model to construct a robust SELD model in noisy environments. Various U-Net model architectures, such as U-Net, Nested U-Net, and ResU-Net have been applied for speech enhancement. Additionally, we used feature maps from the U-Net decoder as input features for SELDnet to perform SELD. The performance of the proposed model was analyzed by comparing speech enhancement and SELD results using the respective evaluation metrics. The evaluation metrics demonstrated that the proposed model exhibited relatively lower restoration performance in highly noisy environments but generally performed well in restoration. In SELD, the proposed model achieved a lower performance than the identically structured model without noise removal in almost-noise-free environments. However, in highly noisy environments, the proposed model outperformed the comparative models. In future research, we aim to enhance the overall performance by investigating more effective speech enhancement techniques and achieving superior performance, even in noise-free environments. Additionally, we aim to further enhance the performance of SELDnet and conduct evaluations through comparisons with the latest technologies, with the goal of improving the performance of SELDnet itself.

REFERENCES

[1] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 1, pp. 279–288, Jan. 2016.

[2] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," in *Proc. IEEE Int. Conf. Robot. Autom.*, vol. 1, May 2004, pp. 1033–1038.

[3] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1120–1124, Sep. 2014.

[4] T. K. Chan and C. S. Chin, "A comprehensive review of polyphonic sound event detection," *IEEE Access*, vol. 8, pp. 103339–103373, 2020.

[5] N. K. Kim and H. K. Kim, "Polyphonic sound event detection based on residual convolutional recurrent neural network with semi-supervised loss function," *IEEE Access*, vol. 9, pp. 7564–7575, 2021.

[6] A. Kumar, R. M. Hegde, R. Singh, and B. Raj, "Event detection in short duration audio using Gaussian mixture model and random forest classifier," in *Proc. 21st Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2013, pp. 1–5.

[7] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *Proc. 18th Eur. Signal Process. Conf.*, Aug. 2010, pp. 1267–1271.

[8] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 559–563.

[9] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 6440–6444.

[10] K. Zhang, Y. Cai, Y. Ren, R. Ye, and L. He, "MTF-CRNN: Multiscale time-frequency convolutional recurrent neural network for sound event detection," *IEEE Access*, vol. 8, pp. 147337–147348, 2020.

[11] M. Neri, F. Battisti, A. Neri, and M. Carli, "Sound event detection for human safety and security in noisy environments," *IEEE Access*, vol. 10, pp. 134230–134240, 2022.

[12] R. Alsina-Pagès, J. Navarro, F. Alías, and M. Hervás, "HomeSound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring," *Sensors*, vol. 17, no. 4, p. 854, Apr. 2017.

[13] S. Goetze, J. Schroder, S. Gerlach, D. Hollosi, J.-E. Appell, and F. Wallhoff, "Acoustic monitoring and localization for social care," *J. Comput. Sci. Eng.*, vol. 6, no. 1, pp. 40–50, Mar. 2012.

[14] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 943–956, 2001.

[15] Q. Yan, J. Chen, G. Ottoy, and L. De Strycker, "Robust AOA based acoustic source localization method with unreliable measurements," *Signal Process.*, vol. 152, pp. 13–21, Nov. 2018.

[16] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Process. Lett.*, vol. 18, no. 1, pp. 71–74, Jan. 2011.

[17] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.
[18] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2386–2390.
[19] C. Pang, H. Liu, and X. Li, "Multitask learning of time-frequency CNN for sound source localization," *IEEE Access*, vol. 7, pp. 40725–40737, 2019.
[20] S. E. Chazan, H. Hammer, G. Hazan, J. Goldberger, and S. Gannot, "Multi-microphone speaker separation based on deep DOA estimation," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2019, pp. 1–5.
[21] S. V. A. Garí, W. Lachenmayr, and E. Mommertz, "Spatial analysis and auralization of room acoustics using a tetrahedral microphone," *J. Acoust. Soc. Amer.*, vol. 141, no. 4, pp. EL369–EL374, Apr. 2017.
[22] T. Butko, F. Pla, C. Segura, C. Nadeu, and J. Hernando, "Two-source acoustic event detection and localization: Online implementation in a smart-room," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2011, pp. 1317–1321.
[23] C. J. Grobler, C. P. Kruger, B. J. Silva, and G. P. Hancke, "Sound based localization and identification in industrial environments," in *Proc. IEEE Ind. Electron. Soc. (IECON)*, Sep. 2017, pp. 6119–6124.
[24] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Proc. Audio Eng. Soc. Conv.*, 2015, p. 138.
[25] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 445–455, Jun. 1998.
[26] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech*, 2013, pp. 555–559.
[27] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, 2019, pp. 1–16.
[28] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 34–48, Mar. 2019.
[29] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Accdoa: Activity-coupled Cartesian direction of arrival representation for sound event localization and detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 915–919.
[30] 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," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 316–320.
[31] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of CRNN models," 2019, *arXiv:1908.00766*.
[32] Y. Cao, T. Iqbal, Q. Kong, M. Galindo, W. Wang, and M. D. Plumbley, "Two-stage sound event localization and detection using intensity vector and generalized cross-correlation," in *Proc. Detection Classification Acoustic Scenes Events (DCASE) Challenge*, Jun. 2019, pp. 1–4.
[33] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," 2020, *arXiv:2006.01919*.
[34] T. N. T. Nguyen, K. N. Watcharasupat, N. K. Nguyen, D. L. Jones, and W.-S. Gan, "SALSA: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 30, pp. 1749–1762, 2022.
[35] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 716–720.
[36] X. Dang and T. Nakai, "Noise reduction using modified phase spectra and Wiener filter," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, Sep. 2011, pp. 1–5.
[37] J.-S. Hu and C.-H. Yang, "Speech enhancement using transfer function ratio beamformer and matched filter array," in *Proc. Int. Conf. Inf. Autom.*, Jun. 2009, pp. 1161–1166.
[38] Y. Wang, J. An, V. Sethu, and E. Ambikairajah, "Perceptually motivated pre-filter for speech enhancement using Kalman filtering," in *Proc. 6th Int. Conf. Inf., Commun. Signal Process.*, 2007, pp. 1–4.
[39] L. Gondara, "Medical image denoising using convolutional denoising autoencoders," in *Proc. Int. Conf. Data Mining Workshops (ICDMW)*, 2016, pp. 241–246.
[40] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.
[41] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested U-net architecture for medical image segmentation," in *Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, 2018, pp. 3–11.
[42] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-net: A multi-scale neural network for end-to-end audio source separation," 2018, *arXiv:1806.03185*.
[43] X. Xiang, X. Zhang, and H. Chen, "A nested U-Net with self-attention and dense connectivity for monaural speech enhancement," *IEEE Signal Process. Lett.*, vol. 29, pp. 105–109, 2022.
[44] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-U-net," 2018, *arXiv:1811.11307*.
[45] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.
[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[48] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-based joint azimuth and elevation localization with the ambisonics intensity vector," in *Proc. 16th Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Sep. 2018, pp. 241–245.
[49] T. Dozat, "Incorporating Nesterov momentum into Adam," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2016, pp. 1–4.
[50] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 30, pp. 829–852, Dec. 2022.
[51] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[52] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2019, pp. 333–337.

YEONGSEO SHIN received the B.S. and M.S. degrees from the Department of Computer Engineering, Chosun University, South Korea, in 2022 and 2023, respectively. Since July 2023, she has been a Research Engineer with LIG Nex1, South Korea. Her current research interests include speech and audio signal processing and deep learning.

YONG GUK KIM received the B.S. degree in electronics and computer engineering from Chonnam National University, South Korea, in 2006, and the M.S. and Ph.D. degrees in electrical engineering and computer science from the Gwangju Institute of Science and Technolgy (GIST), South Korea, in 2008 and 2023, respectively. Since January 2011, he has been a Senior Research Engineer with the Maritime Research and Development Center, LIG Nex1, South Korea. His current research interests include sensor signal processing and sonar system design.

CHANG-HO CHOI received the B.S. and M.S. degrees in control and instrumentation engineering from Kwangwoon University, South Korea, in 1999 and 2001, respectively. Since December 2000, he has been a Senior Research Engineer with the Maritime Research and Development Center, LIG Nex1, South Korea. His current research interest includes sonar system design.

DAE-JOONG KIM received the B.S. degree in electrical instrumentation engineering from the Kumoh National Institute of Technology, South Korea, in 1996, and the M.S. degree in electrical engineering from Sogang University, South Korea, in 2012. He was with LIG Nex1, South Korea, in September 1995. He is currently a Principal Member of the Engineering Staff with the Maritime Research and Development Center, LIG Nex1. His current research interests include AUV, UUV, and maritime security systems.

CHANJUN CHUN (Member, IEEE) received the B.S. degree in electronics engineering from the Korea University of Technology and Education, South Korea, in 2009, and the M.S. and Ph.D. degrees in electrical engineering and computer science from the Gwangju Institute of Science and Technology (GIST), in 2011 and 2017, respectively. From 2017 to 2021, he was a Senior Researcher with the Korea Institute of Civil Engineering and Building Technology (KICT), South Korea. Since March 2021, he has been an Assistant Professor with the Department of Computer Engineering, Chosun University, Gwangju, South Korea. His current research interests include speech and audio signal processing and deep learning.