Automatic Bird Sound Source Separation Based on Passive Acoustic Devices in Wild Environment

Jiangjian Xie, Yuwei Shi, Dongming Ni, Manuel Milling, Shuo Liu, Junguo Zhang, Kun Qian, *Senior Member, IEEE*, and Björn W. Schuller, *Fellow, IEEE*

***Abstract*—The Internet of Things (IoT)-based passive acoustic monitoring (PAM) has shown great potential in large-scale remote bird monitoring. However, field recordings often contain overlapping signals, making precise bird information extraction challenging. To solve this challenge, first, the interchannel spatial feature is chosen as complementary information to the spectral feature to obtain additional spatial correlations between the sources. Then, an end-to-end model named BACPPNet is built based on Deeplabv3plus and enhanced with the polarized self-attention mechanism to estimate the spectral magnitude mask (SMM) for separating bird vocalizations. Finally, the separated bird vocalizations are recovered from SMMs and the spectrogram of mixed audio using the inverse short Fourier transform (ISTFT). We evaluate our proposed method utilizing the generated mixed data set. Experiments have shown that our method can separate bird vocalizations from mixed audio with root mean square error (RMSE), source-to-distortion ratio (SDR), source-to-interference ratio (SIR), source-to-artifact ratio (SAR), and short-time objective intelligibility (STOI) values of 2.82, 10.00 dB, 29.90 dB, 11.08 dB, and 0.66, respectively, which are better than existing methods. Furthermore, the average classification accuracy of the separated bird vocalizations drops the least. This indicates that our method outperforms other compared separation methods in bird sound separation and preserves the fidelity of the separated sound sources, which might help us better understand wild bird sound recordings.**

***Index Terms*—Bird sound separation, multichannel audio processing, polarized self-attention (PSA) mechanism.**

This work was supported in part by the National Natural Science Foundation of China under Grant 62303063, Grant 32371874, and Grant 62272044; in part by the Ministry of Science and Technology of the People's Republic of China with the STI2030-Major Projects under Grant 2021ZD0201900; and in part by the Teli Young Fellow Program from the Beijing Institute of Technology, China. *(Jiangjian Xie, Yuwei Shi, and Dongming Ni contributed equally to this work.) (Corresponding authors: Junguo Zhang; Kun Qian.)*

Jiangjian Xie and Junguo Zhang are with the School of Technology and the State Key Laboratory of Efficient Production of Forest Resources, Beijing Forestry University, Beijing 100083, China (e-mail: shyneforce@bjfu.edu.cn; zhangjunguo@bjfu.edu.cn).

Yuwei Shi and Dongming Ni are with the School of Technology, Beijing Forestry University, Beijing 100083, China (e-mail: shiyuwei@bjfu.edu.cn; nidongming@bjfu.edu.cn).

Manuel Milling is with the Chair for Health Informatics, MRI, Technical University of Munich, 80333 Munich, Germany (e-mail: manuel.milling@informatik.uni-augsburg.de).

Shuo Liu is with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany (e-mail: shuo.liu@uni-a.de).

Kun Qian is with the Key Laboratory of Brain Health Intelligent Evaluation and Intervention, Ministry of Education, and the School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: qian@bit.edu.cn).

Björn W. Schuller is with the Group on Language, Audio, & Music, Imperial College London, SW7 2AZ London, U.K., also with the Chair for Health Informatics, MRI, Technical University of Munich, 80333 Munich, Germany, and also with the Munich Center for Machine Learning, 80538 Munich, Germany (e-mail: schuller@ieee.org).

Digital Object Identifier 10.1109/JIOT.2024.3354036

I. Introduction

BIRDS communicate mostly using sound, particularly vocalizations, which are typically unique for species and can be used to identify taxonomic discrepancies between species. Thus, passive acoustic monitoring (PAM) can be employed as an automated bird monitoring tool that eliminates the time consuming and costly requirements of traditional manual surveys. Furthermore, a monitoring system with several PAMs can enable dynamic monitoring of birds over long periods of time and across large regions, which has attracted a lot of attention [1], [2]. Automatic bird species recognition is an efficient way for processing the massive PAM recordings [3]. The bird acoustic "cocktail party problem" (CCP) is a term used to describe how birds may vocalize simultaneously in real wild environments, such as during the bird dawn chorus. This causes overlaps between different sound sources [4]. The overlapping of bird sounds caused by the simultaneous vocalization of multiple bird individuals has been considered the biggest obstacle in the automated processing of natural recordings [5]. Automatic sound source separation aims to address the overlap issue by disentangling the overlapping signals into multiple components that correspond to different sound sources, which can also extract meaningful information from interactions when there is concurrent sound in a noisy environment [6]. Therefore, when isolated recordings of relevant sources are inaccessible, source separation of mixture audio has become an effective step prior to further bioacoustics research [7]. Numerous previous methods, such as independent component analysis (ICA) [8], and nonnegative matrix factorization (NMF) [9] have been proposed to address the sound source separation problem on a single channel and have shown promising results.

However, since a single-channel microphone cannot obtain the spatial details of diverse sources, the performance of these methods would degrade with overlapping sounds from several sources [10]. Setting certain microphones to collect multichannel recordings to gather spatial position information of the sound sources, and then integrating the spatial features between the different channels, allows for improved separation

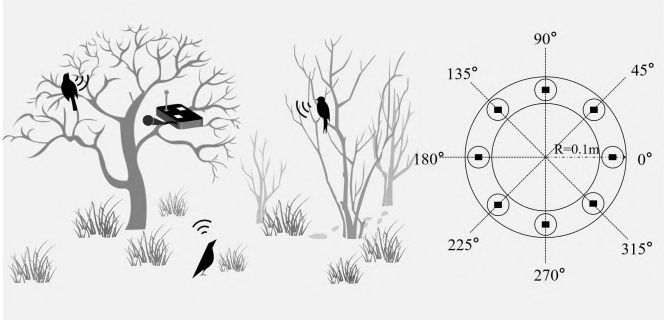


Fig. 1. Simulation of the deployment of bird monitoring equipment in the field environment and the structure of the eight-channel monitoring equipment.

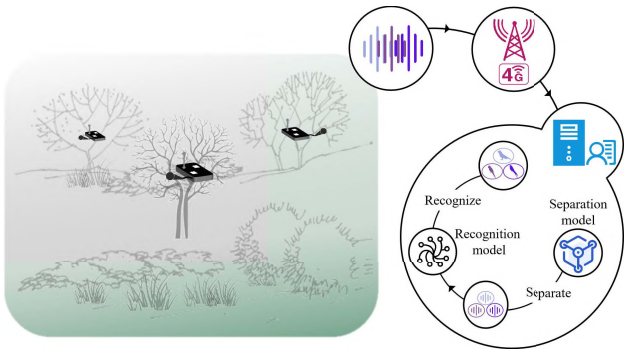


Fig. 2. Diagram of the IoT-based bird monitoring system.

performance [11], which had attracted widespread attention in recent years. Hence, this study designs a bird monitoring device with an eight-channel circular microphone array, which enables different bird sound sources to be spatially distinguished [12]. As shown in Fig. 1, each array consists of eight microphones arranged uniformly within a circular framework with a radius of 0.1 m. The distance between each microphone is 45° degrees apart. A bird monitoring system with several proposed monitoring devices, as shown in Fig. 2, can also be implemented in forest areas using the Internet of Things (IoT) technology to enable large-scale bird monitoring. The recordings are transmitted to the monitoring center, where a separation model is employed to separate the mixed sound signals into audio signals that solely contain the individual bird vocalizations. Subsequently, a bird sound recognition model is used to recognize the separated audio. The information on the birds at the monitoring site may then be analyzed.

Throughout the whole procedure, the importance of the separation is evident, as it directly affects the subsequent task of bird species recognition. Improved separation performance leads to more comprehensive understanding of the monitoring recordings. Therefore, this article focuses on addressing the task of separating mixed audio signals.

For multichannel source separation of bird sounds, conventional approaches use the cascade method, which incorporates unique functions based on array signal processing techniques [13], [14]. However, the main drawback of cascade systems is their accumulation of errors at each function block. Recently, deep learning (DL)-based multichannel source separation methods have made remarkable advancements in speech separation [15], [16], owing to the fact that DL-based methods can constitute an end-to-end method that does not use a traditional cascade system and thus reduces error accumulation [17], [18]. Although DL-based speech separation models are competitive areas of research, the bird acoustics cocktail party problem (BACPP) has received less attention, because current researches on bird sounds focus more on other aspects of machine learning-based tasks, such as automated detection and recognition of bird sounds [19], [20]. Hence, we argue that a DL-based multichannel separation method should be explored for achieving robust bird sound source separation.

This study attempts to address the BACPP using a DL-based multichannel method. There are two challenges: one is how to extract effective and meaningful representations for the sounds of various bird species and the other is how to distinguish isolated sources within a mixture of audio signals. For the first challenge, we propose the BACPPNet based on Deeplabv3plus [21]. The atrous spatial pyramid pooling (ASPP) [22], consisting of dilated convolutional layers [23] with different rates, is introduced to efficiently handle multiple features of different bird species. To learn more general and meaningful representations, the polarized self-attention (PSA) [24] mechanism is embedded at the bottleneck of the encoder and the decoder. In response to the second challenge, both the spectral and spatial properties of mixture audio signals are used as network inputs to improve the discrimination of various isolated sources. Furthermore, the spectral magnitude mask (SMM) [25] of each direction of sound sources is predicted according to the angular resolution, which can avoid over-fitting to the relationship between the direction of the sound source and the bird class. Finally, the estimated SMM and the original input mixture audio signal are used to reconstruct each source signal. A downstream classification task is designed to evaluate the fidelity of the estimated source signals, and five source separation evaluation metrics are employed to assess the quality of the sound source reconstruction.

This study's contributions are summarized as follows.

1) We propose BACPPNet, a Deeplabv3plus-based model for bird sound separation, whose inputs incorporate the spectral and spatial features of mixture audio signals to provide precise SMM predictions in each direction of the sound source, assuring the fidelity of the reconstructed sound source signal. The manner in which separation in different directions can avoid over-fitting to the relationship between the direction of the sound source and the bird class.
2) We introduce the PSA mechanism to the bottleneck of Deeplabv3plus to guide the model to preserve the high-resolution features of the encoder, resulting in better feature extraction capability of the separation model.
3) We design a downstream classification task to evaluate the quality of the reconstructed signal. The higher classification accuracy reflects the better fidelity of the estimated source signals from BACPPNet in comparison to the ground truth sources.

The remainder of this article is organized as follows. Section II describes relevant work related to our study. Section III describes the process of creating a multisource mixture data set of bird sounds. The details of the BACPPNet

and the classification model used are explained in Section IV. Section V explains the experimental setup and evaluation criteria, and presents the simulation results. The conclusion and future work are outlined in Section VI.

II. Related Work

This section mainly reviews the work related to our study, including bird sound source separation, multichannel sound source separation methods, and attention mechanisms.

A. Bird Sound Source Separation

The bird sound source separation per se is barely a common use case. Instead, it is frequently linked to the subject of bird sound classification. Traditional solutions of bird sound source separation problems are always based on "cascade" systems [13], [26], such as Kojima et al. [26]'s study, where a spatial-cue-based framework integrating sound source detection, localization, separation, and identification of bird sounds was introduced, and the result showed that the system outperformed a conventional method based on robot audition [27]. Gabriel et al. [13] introduced an original system that used a multiple signal classification (MUSIC) algorithm, a geometric high-order decorrelation-based source separation (GHDSS) algorithm and convolutional neural networks (CNNs) in sequence for sound source localization, separation, and classification, respectively. The results showed that this system can distinguish between different bird types with satisfying results. However, the independent optimization of each block causes the accumulation of multistage errors, which can result in poor performance [11].

Several studies have proposed the integration of sound source segmentation or separation with classification tasks. Shugaev et al. [28] proposed a multistep (train-segment-shift) training method that first develops pseudo labels for segmentation and then uses noise from external data to mitigate the domain mismatch. The audio segmentation model incorporated the global multihead self-attention to account for the interaction between different parts of the audio, which consequently enhanced the classification performance. Dai et al. [29] first used independent vector analysis (IVA) to separate source signals from the multichannel signal, before classifying the separated sources using an Xception-based architecture. Experiments on the BirdCLEF2020 data set showed that this model achieved a higher macro F1-score and average accuracy than state-of-the-art methods. Denton et al. [30] separated single channel bird sounds using the unsupervised sound source separation method MixIT. Over-separation occurred, leading the probability of the most prevalent species in the recordings to decrease after separation. Through combining the original mixture with the separated signals, the downstream classifier outperformed the best model in the BirdCLEF2019 competition.

Although these studies suggested that combining separation and classification can be beneficial to identify the bird sounds in the complex acoustic environments, they failed to comprehensively evaluate the contribution of the separation or segmentation model to the overall task. Contrary to these investigations, we qualitatively evaluate the separation performance of BACPPNet using spectrograms, which are visual representations of the audio signals. Furthermore, using widely applied evaluation metrics and downstream classification accuracy, we quantitatively evaluate reconstruction quality of separated sound sources estimated by a separation model.

B. Multichannel Sound Source Separation

Several DL-based multichannel sound source separation models have recently been developed to efficiently handle the CPP of human speech, music, or ambient sound [31], [32]. The mask-based approach is the most commonly used method, which is always performed in the time-frequency (TF) domain, allowing the spatial information to be incorporated efficiently [33]. Furthermore, the DL-based model is trained by minimizing the estimation error between the spectrograms of T-F masks of the ground truth and estimated signals.

Hence, the performance of the separation models highly depends on the T-F mask's estimation, which may require more elaborate separation models using effective and meaningful representations from the sound sources. Additionally, the spatial information is vital in multichannel sound source separation [34], and numerous studies have demonstrated that using spatial features between different channels as additional information to spectral features can effectively improve source separation performance [35], [36]. To use the spatial information, the interchannel features (sometimes combined with spectral features), for example, time/phase/level difference (ITD/IPD/ILD) are input to the neural network for the full-band TF mask prediction in [32] or sub-band TF mask prediction in [37].

U-Net is a popular semantic segmentation network that can be used to segment the spectrogram of mixed audio to predict the T-F masks for different sources [38]. Kadandale et al. [39] performed a singing voice separation task using a multichannel U-Net. The results of several publicly available data sets showed that this architecture outperformed other designs. Sudo et al. [18] proposed a multichannel environmental sound segmentation method using sound source localization and separation based on U-Net. The results on the developed data sets, which included 75-class environmental sounds, indicated that the proposed method achieved a smaller root mean squared error than the standard method. Wang et al. [40] utilized Multichannel U-Net to separate the cardiopulmonary sound, obtaining higher separation quality and robustness. Tan and Wang [38] added a long short-term memory recurrent neural network (LSTM-RNN) to U-net's bottleneck to improve the context modeling ability, and the results demonstrated that this method effectively improved the objective intelligibility and perceptual quality of the source signal. Although these separation methods based on the U-net have achieved promising results in their respective tasks, we suggest that the results should be improvable, because the U-net does not sufficiently combine shallow and deep features when performing feature extraction during down-sampling, which results in the loss of signal features, which is undesirable for fitting accurate T-F masks.

Aside from the U-Net-based approaches, many other segmentation methods for computer vision have been developed recently [41], [42]. An impressive example is Deeplabv3plus which has an encoder–decoder architecture, and uses depthwise separable convolutions in the encoder to improve accuracy and computational speed. An ASPP is inserted between the encoder and the decoder of Deeplabv3plus. It is a pyramidal structure with dilated convolutional layers of varying rates, which is significantly useful for gathering contextual information on semantic units and enhancing context modeling. Deeplabv3plus benefits from ASPP in challenging situations involving two interfering sources, such as when the target and interference are close in terms of angular distance [43]. Deeplabv3plus has been introduced to audio-related tasks in recent years [11], [44], [45] and has obtained excellent outcomes. In our study, we use Deeplabv3plus for bird sound separation, and expand the input features to a multichannel input to fuse the spectral and spatial features of mixed audio signals. We expect that the above-mentioned strategies will provide precise T-F masks, which could ensure the fidelity of the reconstructed sound source signal.

C. Attention Mechanism

The introduction of the attention mechanism has contributed to the success of numerous DL models, and it continuous to be a prevalent component in state-of-the-art models, including several source separation systems [46], [47], [48].

Hong et al. [46] proposed an attention mechanism for the temporal-spatial neural filter (TSNF), which realized the channel attention on merged features and the feature map of the 1-D convolution block in the temporal convolution network. According to experimental results, the proposed methods produced an SI-SNR improvement of approximately 1.2 dB for close speakers, and a slight decrease of 0.1 dB for other cases. Sun et al. [47] combined squeeze-and-excitation network (SE) [48] attention and a convolutional RNN for speech separation in monaural recordings. The feature recalibration strategy of the SE, which can explicitly construct the interdependence between feature channels, enabled the model to highlight the useful feature maps of spectrograms. Chen et al. [49] proposed a lightweight multistage network for monaural vocal and accompaniment separation. In this network, a dual-branch attention (DBA) module was used to obtain the correlation of each position pair among the channels of feature maps, respectively. The ablation experiments demonstrated that the DBA module can improve the separation performance.

Given that the attention mechanism has demonstrated its effectiveness in various source separation tasks, and that DL-based sound source separation models with an encoder–decoder architecture suffer from a bottleneck, we employ the attention mechanism to automatically choose which part should be the focus, allocate limited information processing resources to more important parts, and concentrate specifically on the important information. Compared with other attention mechanisms [48], [50], [51] the PSA can maintain high internal resolution in both polarized channel-only and spatial-only attention branches while including a nonlinear composition that fully uses the high-resolution information maintained in the channel and spatial branches. To the best of our knowledge, not much research work has been done on attention mechanisms for bird sound source separation. In this study, we explore incorporating the PSA mechanism into the bird sound source separation model to assist the model to focus more on the differences between sound source representations.

III. Data Generation and Processing

Although supervised DL has produced decent source separation results, training such approaches requires a significant number of pairs of mixed and isolated sounds. The most typical solution to this problem is to generate mixed signals from a database of separate source sounds. To create the data set, we select three different bird species living in the same habitat as the research objects, and develop a mixing system to generate the multisource overlapping bird sounds.

A. Bird Sound Data

For training of both the separation and classification models, we choose three bird species, *Great Reed-Warbler*, *Common Cuckoo*, and *Magpie* that share a habitat but have dissimilar vocal behaviors in terms of spectral and temporal properties. The 3.4-h long recordings of bird sounds used in this work are downloaded from the Xeno-Canto website (https://www.xeno-canto.org/). Because these recordings are recorded in the wild, they contain noise from wind, rain, or other environmental audio sources. Nevertheless, the most important thing is that the selected recordings do not have overlapping bird sounds. All the recordings are resampled at 16 kHz and saved in WAV format. Furthermore, we split the recordings of each species into 4.07-s segments to ensure that sufficient bird sound features are included while adjusting to the computation and memory efficiency. We use these known class bird sound segments as single-bird-sound source signals to generate multisource overlapping bird sounds.

B. Mixing System

To obtain multisource overlapping bird sounds with spatial characteristics, we simulate the eight-channel microphone array mentioned above as a hybrid system (shown in Fig. 3), with the 0° facing microphone used as the reference microphone.

In this mixing system, the single-bird-sound sources are placed 1 m from the microphone at the same height. Because birds have territorial awareness, and different bird species tend to live in different spaces [16], the direction of the single-bird-sound source is randomly selected at intervals greater than 45°. The impulse response from each source to the microphone array is calculated using the image source method [52] as follows:

$$\boldsymbol{h}(t, d, \theta) = [\boldsymbol{h_1}(t, d, \theta), \boldsymbol{h_2}(t, d, \theta), \ldots, \boldsymbol{h_8}(t, d, \theta)] \quad (1)$$

where $\boldsymbol{h_i}(t, d, \theta)$ represents the spatial impulse response generated using a microphone i at a distance d from the sound

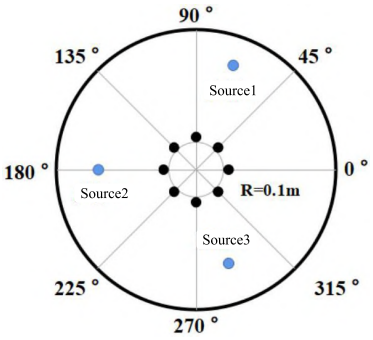


Fig. 3. Eight-channel circular microphone array. The distance between the center of the microphone array and the single-bird-sound sources is 1.0 m. The direction of the single-bird-sound source is randomly located at intervals greater than 45°.

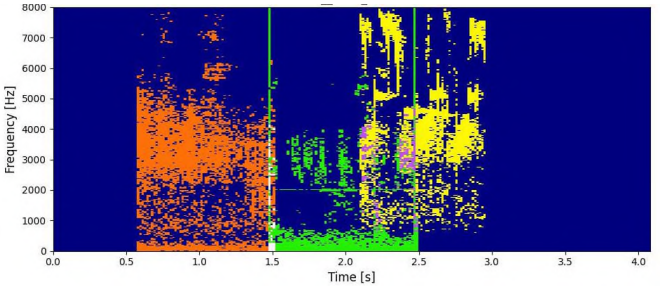


Fig. 4. Example of a spectrogram of mixed audio samples with overlap. Different colors represent the spectrum of different sound sources in the T-F domain. The overlapping part of the sound source is displayed in white and purple.

source in the θ direction, and $\boldsymbol{h}(t, d, \theta)$ represents the spatial impulse response generated using the microphone array. Because the majority of a bird's acoustic environment is an open field, we ignore the reflection of the bird sound source, therefore, the effect of reverberation on source separation is not considered. To generate mixture audio signals captured with the microphone array, the impulse response from the sound source to the microphone array is convolved with each source $s_i(t)$ as follows:

$$\boldsymbol{x}(t) = \boldsymbol{h}(t, d, \theta) * s_i(t - t_r) \tag{2}$$

where * denotes the convolution operator, $\boldsymbol{x}(t)$ represents the eight-channel multisource mixture bird sound signal, and i represents the number of single-bird-sound sources.

C. Dataset

The mixture data set is generated from the single-bird-sound sources of the three specific species, annotated according to the known classes. Asynchrony has been suggested as an important acoustic cue, that helps the animal brain solve the CPP [53]. To simulate a more practical scenario, we randomly select and mix three single-bird-sound sources with random shifts of the overlaps. Fig. 4 shows an example of a spectrogram for a mixed audio sample with overlap. The spectral information of the three birds is shown in a single spectrogram that is colored differently for each bird. Finally, 1000 mixed files with a duration of 4.07 s are generated, totaling roughly 1.13 h. During the training of the separation model, the mixture of audio signals is utilized as input data, and each single-bird-sound source signal is used as the ground truth. In each separation experiment, 80 % of the mixed data are used for training, 10 % for validation, and 10 % for testing.

For the downstream classification task, the single-bird-sound source signals are used to train a classifier. The reconstructed signals from the separation task are used to evaluate the fidelity of the sound source reconstruction.

D. Feature Extraction for the Separation Task

Combining interchannel spatial features of sound signals as complementary information to spectral features can significantly enhance speech separation quality [36]. The objective of multifeature fusion is to mathematically analyze data from different sources and create a new representation that can be used more effectively for pattern recognition and other multimedia information processing tasks [54].

In this study, we use both the spectral and spatial features proposed in [35]. For the multichannel mixture audio signals, we calculate the short-time Fourier transform (STFT) features of the reference microphone m and the nonreference microphone n using a 512-point STFT with 50% overlap. An amplitude spectrogram [55] of the reference microphone is used as the spectral feature. Then, the interchannel phase difference (IPD) between the reference microphone m and the other microphone n is calculated using (3). The amplitude of the IPD is subsequently normalized to the range [0, 1] using (4) and (5) as follows:

$$\delta_{t,f,m,n} = \angle x_{t,f,m} - \angle x_{t,f,n} \tag{3}$$

$$sinIPD(t, f, m, n) = \sin\left(\delta_{t,f,m,n}\right) \tag{4}$$

$$cosIPD(t, f, m, n) = \cos\left(\delta_{t,f,m,n}\right) \tag{5}$$

where $\delta t, f, m, n$ denotes the IPD between the STFT $\angle x_{t,f,m}$ and $\angle x_{t,f,n}$ at time t and frequency f of the signals at microphones m and n. Because we use the eight-channel microphone array, the selection of one reference microphone allows us to calculate 14 IPDs of *sinIPD* and *cosIPD*.

IV. Methodology

DL has shown superior performance in modeling masking-based source separation methods. In this study, we propose the BACPPNet masking-based method (Fig. 5). BACPPNet is essentially a Deeplabv3plus model with the PSA mechanism. During the separation stage, the spectral and spatial features are jointly used as inputs to train the separation model, which is motivated by the fact that the human auditory system employs the localization information (that is, binaural cues) to solve the CPP. The model predicts the SMM for each direction of the sound source. With each single SMM, the spectrogram of each direction can be achieved by the multiplication between the SMM and the mixture audio spectrogram of the reference microphone. Finally, the separated bird vocalization is recovered from the corresponding spectrograms of each direction by the inverse short Fourier transform (ISTFT). Besides the widely used metrics for speech separation, the

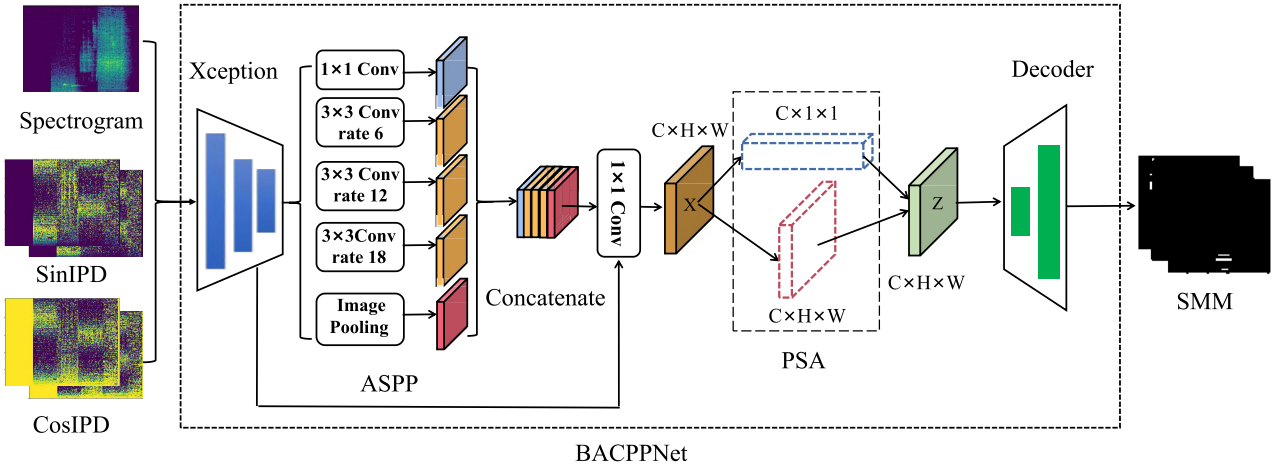


Fig. 5. Overall procedure of SMM prediction with BACPPNet. BACPPNet consists of Xception, ASPP, PSA, and decoder blocks. The inputs to the BACPPNet are the STFT spectrogram, sinIPD, and cosIPD of mixture audio signals. Xception is a DCNN aiming at an extraction of high-level features of the inputs. ASPP is a multiscale feature fusion strategy, which adopts atrous convolution with different rates to get multiscale information of the input features. PSA is inserted to the bottleneck to prevent the potential loss of high-resolution information in DCNNs by downsampling. The BACPPNet outputs the SMM for each source direction.

classification performance of the separated sound sources is evaluated to assess the accuracy of sound source reconstruction. Here the CS-CLDNN model, which we proposed in [56], is employed to classify bird vocalizations.

A. Deeplabv3plus

Deeplabv3plus is a semantic segmentation model with an encoder–decoder structure, which uses Xception [57] in the encoder to extract high-level features of the input spectrogram. Xception is a deep CNN (DCNN) that uses a depth-wise separable convolution or group convolution to achieve high performance while using less processing power and memory.

The ASPP comes after the Xception block and uses an atrous convolution with different rates to extract convolutional features at multiple scales. This operation allows for a wider range of contextual information to be extracted without increasing the number of parameters, effectively improving the contextual modeling capability. Therefore, the model is more suitable for bird sound spectrograms with different shapes. For example, the duration of a *Great Reed-Warbler* call is shorter than that of a *Magpie* call, which leads to different sizes of spectrograms.

The ASPP module uses a kernel size of 3×3, and dilation rates of 6, 12, and 18. The input features are down-sampled by the encoder and the size of the feature map is reduced to 1/16 of the original input. The encoder sequentially reduces the spatial resolution and increases the channel resolution. The number of channels is reduced to the requested size using a convolutional kernel of 1×1 after the encoder. The encoded features are first bi-linearly up-sampled by a factor of four before being concatenated with the corresponding low-level features from the network backbone. After the concatenation, we use 3×3 convolutions to refine the features before performing another simple bi-linear up-sampling by a factor of 4. The final layer of the decoder uses a Sigmoid activation to obtain the SMMs with the same size as the input spectrogram.

B. PSA Mechanism

Because the down-sampling operation of the encoder reduces the resolution of the input features, the tensor connecting the encoder and decoder has less elements than both the input tensor and the output tensor, which improves computation and memory efficiency. However, the down-sampling operation results in a partial loss of the input features' details. Moreover, high resolutions of the input and the output are preferred for the fine details of segmentation results [58]. Therefore, we insert the PSA mechanism at the bottleneck of Deeplabv3plus to decrease the potential loss of high-resolution information in the encoder caused by down-sampling.

The PSA consists of channel-only and spatial-only self-attention branches. The channel-only self-attention $A^{ch}(\boldsymbol{X}) \in \mathbf{R}^{C\times 1\times 1}$ can be presented as follows:

$$A^{ch}(\boldsymbol{X}) = F_{SG}\left[W_{z|\theta_1}(\sigma_1(W_v(\boldsymbol{X}))) \times F_{SM}\left(\sigma_2(W_q(\boldsymbol{X}))\right)\right] \quad (6)$$

where F_{SG} represents the Sigmoid operator, W_q, W_v, and W_z are convolutional layers, respectively, σ_1 and σ_2 are two tensor reshape operators, F_{SM} is the SoftMax operator, and $\times$ is the matrix dot-product operation. F_{SM} is defined as follows:

$$F_{SM}(\boldsymbol{X}) = \sum_{j=1}^{N_p} \frac{e^{x_j}}{\sum_{m=1}^{N_p} e^{x_m}} x_j. \quad (7)$$

The input feature X is first converted into features Q and V using a convolution operation. The channel of Q is significantly compressed, while the channel dimension of V remains at a high level of $C/2$, maintaining a high resolution in the channel. Then, Q and K are used for matrix dot-product operations, which is followed by applying 1×1 convolutional and LayerNorm layers to increase the dimension of the channels to the original input size C. Finally, the Sigmoid function is used to maintain all parameters between

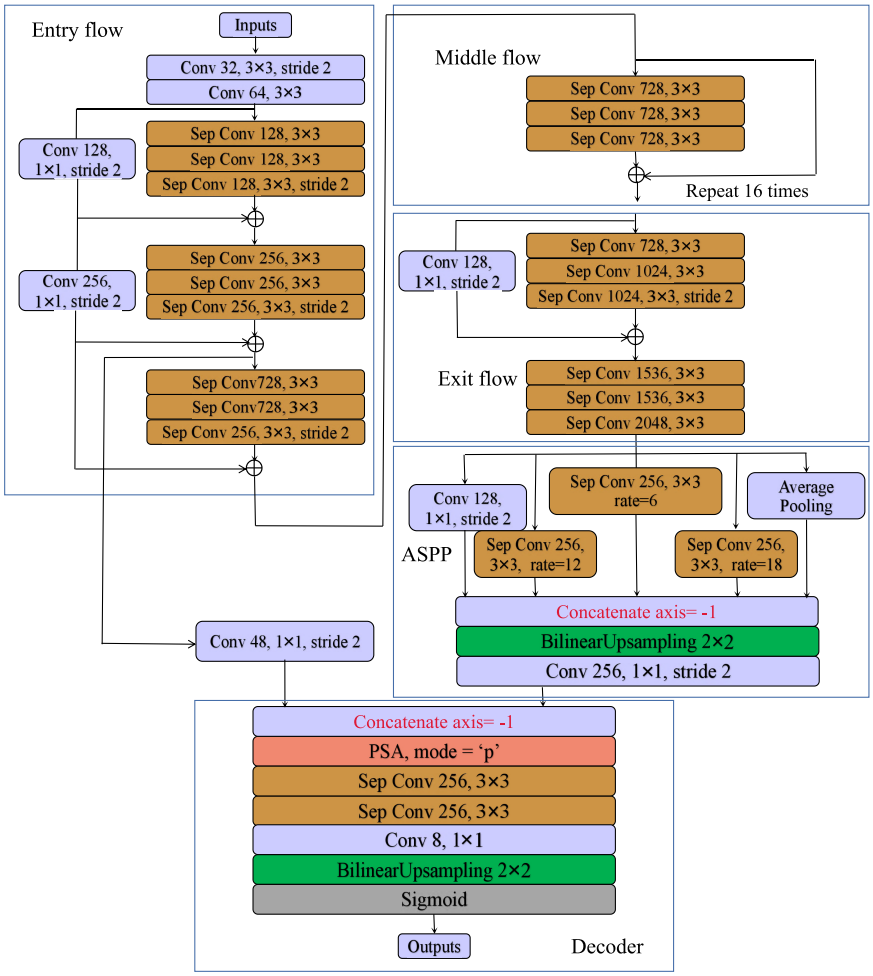


Fig. 6. BACPPNet framework. The encoder consists of an entry flow, middle flow, and exit flow. There are four blocks in the entry flow, 16 blocks in the middle flow, and two blocks in the exit flow. Each block consists of a 2-D Separable convolution of stride 2 and kernel size 3 × 3, followed by a batch normalization and an ReLU activation after each 3 × 3 depth-wise separable convolution. The ASPP consists of a 2-D convolution of stride 2 with kernel size 1 × 1, three atrous convolutions with a dilation rate of 6, 12, and 18, respectively, and an average pooling layer. The PSA is added to the concatenated low-level features from the entry flow and the output of the ASPP. After the concatenation, 3 × 3 separable convolutions are applied to refine the features. Then, a 2-D convolution with eight filters followed by another simple bi-linear up-sampling by a factor of four is used to obtain mask spectrograms of the same size as the input spectrogram.

0 and 1. The output of the channel attention branch is shown in the following:

$$Z^{ch}(\boldsymbol{X}) = A^{ch}(\boldsymbol{X}) \otimes^{ch} \boldsymbol{X} \in \mathbf{R}^{C\times H\times W}. \tag{8}$$

The spatial-only self-attention branch is shown in the following:

$$A^{sp}(\boldsymbol{X}) = F_{SG}\left[\sigma_3(F_{SM}\left(\sigma_1(F_{GP}(W_q(\boldsymbol{X}))) \times \sigma_2(W_v(\boldsymbol{X}))\right)\right] \tag{9}$$

where F_{GP} denotes a global pooling operator defined as follows:

$$F_{GP}(\boldsymbol{X}) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{X}(:,i,j). \tag{10}$$

Similar to the channel-only self-attention, the spatial-only Self-Attention branch first converts the input features into the features Q and V using a 1 × 1 convolution operation. Then, the spatial resolution of Q is compressed using global pooling, while maintaining the spatial resolution of V at a high level of $H \times W$. Because the spatial resolution of Q is compressed, Q is augmented using the Softmax operator. Finally, the Sigmoid function is used to maintain all parameters between 0 and 1. The output of the spatial-only self-attention branch is shown in the following:

$$Z^{sp}(\boldsymbol{X}) = A^{sp}(\boldsymbol{X}) \otimes^{sp} \boldsymbol{X} \in \mathbf{R}^{C\times H\times W}. \tag{11}$$

The outputs of the two above branches are composed under the parallel layout as follows:

$$\mathrm{PSA}_p(\boldsymbol{X}) = \boldsymbol{X}^{ch} + \boldsymbol{X}^{sp}. \tag{12}$$

C. BACPPNet

Fig. 6 shows a detailed description of BACPPNet. The encoder is an Xception block consisting of three flows, i.e., entry flow, middle flow, and exit flow. There are four blocks in the entry flow, 16 blocks in the middle flow, and two blocks in the exit flow. An ASPP is used following the exit flow. Each block consists of a 2-D separable convolution

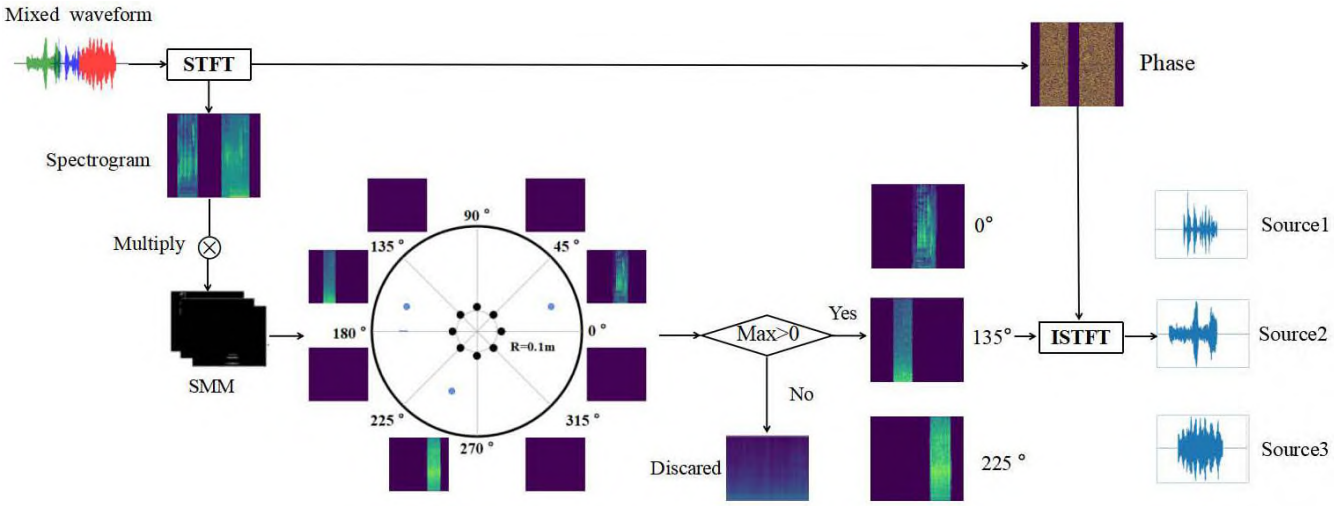


Fig. 7. Reconstruction of the separated single-bird-sound source signals. The spectrograms with energy values of 0 are discarded, while the rest is kept. The time domain signals of kept spectrograms are recovered from the phase of the reference microphone m by ISTFT.

of stride 2, kernel size 3 × 3, batch normalization, and ReLU activation [59]. Skip connections are used to pass low level features directly from the encoder to the decoder. The PSA mechanism is added to the concatenation of the low-level features from the Entry flow and the output of the ASPP. The final layer uses a Sigmoid activation. BACPPNet predicts the SMM of each direction of sound sources according to the angular resolution N. The SMM is defined as follows: follows [15]:

$$\boldsymbol{SMM}(t,f) = \frac{|\,\boldsymbol{S}(t,f)\,|}{|\,\boldsymbol{Y}(t,f)\,|} \tag{13}$$

where $|\,\boldsymbol{S}(t,f)\,|$ and $|\,\boldsymbol{Y}(t,f)\,|$ represent spectral magnitudes of clean source and mixture sound, respectively. $|\,\boldsymbol{S}(t,f)\,|$ is the spectral magnitudes of a sound source in a specific direction, and $|\,\boldsymbol{Y}(t,f)\,|$ is the spectral magnitude of mixture sound signals. Because the input features are normalized, the range of $|\,\boldsymbol{SMM}(t,f)\,|$ is bounded to [0, 1].

The training aims to optimize the reconstruction of separated spectrograms from the input mixture signals. The mean square error (MSE) between the ground truth and separated spectrogram is selected as the loss function using (14) as follows:

$$L(\boldsymbol{X},\boldsymbol{Y}) = ||\,f(\boldsymbol{X}) \otimes \boldsymbol{X}_{\mathbf{mag}} - \boldsymbol{Y}\,||_2 \tag{14}$$

where $\otimes$ denotes the element-wise product, $f(\boldsymbol{X})$ is the prediction of SMM, $\boldsymbol{X}_{\mathbf{mag}}$ denotes the mixture spectrogram of the reference microphones m, and $\boldsymbol{Y}$ denotes the ground truth spectrogram.

D. Sound Source Signal Reconstruction

In this study, the angular resolution N is set to 45°, and the model predicts the SMMs of 360 °N directions as shown in Fig. 7. Therefore, we can obtain the spectrograms for all eight directions. Considering the actual application scenario, if two or more bird sound sources are particularly closer than 45°, we can improve the separation quality by decreasing the angular resolution. We abandon the directions where the spectral magnitudes are zero, which implies that there is no bird sound source in that direction. For the other directions, the time domain signal is reconstructed from the predicted spectrograms and the phase of the reference microphone m using the ISTFT.

E. Classification Models

We develop a classification model to identify the separated sound sources as a downstream task to provide a more physical interpretation in support of the widely used metrics for speech separation. This interpretation is based on the assumption that classification accuracy reflects the fidelity of the separated signal relative to the ground truth source. This makes it a good proxy for assessing the accuracy of sound source reconstruction.

We employ the CS-CLDNN model to classify bird vocalization [56], which introduces the convolutional block attention module (CBAM) [51] and Swish [60] activation functions to improve the CLDNN model. Higher performance can be achieved with the assistance of the attention module [55], [56]. The detailed structure of the CS-CLDNN is shown in Fig. 8. The 40-D MFCCs (given a frame length of 520, and an overlap between frames of 260) of the sound source segments are calculated as the inputs. To extract TF features, a two-layer 2-D convolutional layer with shortcut connection is utilized, followed by the CBAM and another two-layer 2-D convolutional layer with shortcut connection. Subsequently, the time–frequency features are transposed and split according to the time dimension, and progressively input into the LSTM network to extract time sequence features. Finally, concatenating the later layer with the output of the LSTM improves its feature representation ability. The DNN layer maps features to a low-dimensional classification space and outputs the classification results.

V. Experiments and Results

We conduct experiments to evaluate our proposed method on the test set of the generated data set. We first compare

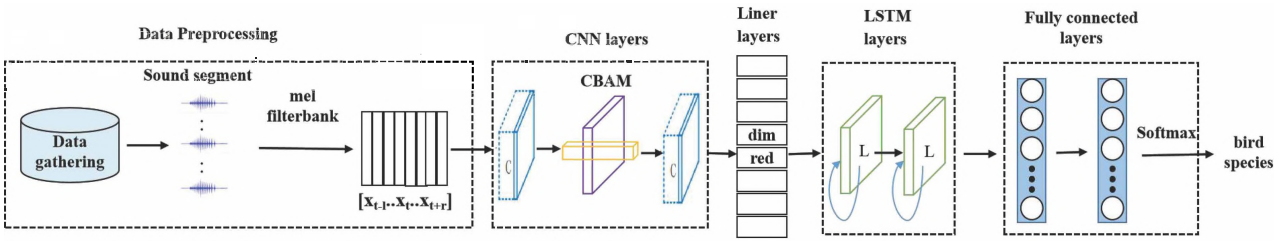


Fig. 8. Overall framework for bird species classification based on the CS-CLDNN. As inputs, the 40-D MFCCs of the sound source segments are computed. The CNN layers (inserted CBAM), linear layers, LSTM layers, and fully connected layers sequency make up the CS-CLDNN. The classification results can be obtained from the Softmax layer [56].

the performances of our proposed model with four existing separation models. Additionally, we conduct several experiments to analyze the importance of PSA and multifeature inputs. Finally, classification experiments are conducted on the reconstructed signals derived from different separation models. Different classification performances are used to evaluate the fidelity of the reconstructed signals.

A. Experimental Setup

The separation experiments are performed on the Tensorflow 1.10 and Keras 2.2.0 frameworks with a programming environment of Python 3.6.0 and a hardware environment of Intel i5, NVIDIA 1080 Ti (11 GB). During the training process, the same experimental conditions and hyper parameter settings are used for all models. The MSE is introduced as the loss function. Adam [61] is used as the optimizer with a learning rate of $1e{-}3$, a decay rate of $\beta_1 = 0.9$ and a batch size of four.

The classification experiments are conducted on the same framework and programming platforms as the separation experiments. We minimize the cross entropy loss using the Adam optimizer with a learning rate of 1e–4 and a batch size of 32.

B. Evaluation Metrics

To evaluate the performances of the source separation methods, five objective metrics, including root MSE (RMSE), source-to-distortion ratio (SDR), source-to-interference ratio (SIR), source-to-artifact ratio (SAR) [62], and short-time objective intelligibility (STOI) [63] are considered. The value of RMSE shows the absolute error on the spectral magnitudes of the ground truth and estimation as shown in the following:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n=0}^{N}\left(Y_n - \hat{Y}_n\right)^2} \tag{15}$$

where Y_n and $\hat{Y}_n$ represent the magnitude spectra of the ground truth and estimation, respectively. N represents the number of TF bins. The SDR, SIR, and SAR are calculated in the BSS EVAL toolbox [62], and the STOI is the intelligibility measure used in the evaluations of the separation method. STOI ranges from 0 to 1 and indicates the correlation of short-time temporal envelopes of the clean anechoic target and the mixture/segregated signal.

C. Comparison With Other Multichannel Separation Models

Since our source separation method employs a semantic segmentation model with an encoder–decoder structure, we compare four different typical semantic segmentation models that have recently been applied to audio source separation related tasks, including DFANet [41], CRNN [38], U-Net [64], and Deeplabv3plus [21]. In addition, another model, Denseaspp [42], is selected as a comparison model due to the fact that it also contains the ASPP structure. Above five separation models are all trained with the same hyperparameter settings. The spectrogram, sinIPD, and cosIPD of the mixture audio signals are used as input features of the separation model. The results on the test set are shown in Table II.

As shown in Table I, the Deeplabv3plus model outperforms DAFNet, Denseaspp, U-Net, and CRNN with respect to all the evaluation metrics, suggesting that the Deeplabv3plus model can learn deeper contextual features. Therefore, it seems more suitable for bird sound source separation. BACPPNet outperforms all other compared models. Because BACPPNet's RMSE is the lowest in comparison, its predictions can be interpreted as the closest to the ground truth. The SDR, SIR, and SAR of BACPPNet are also better than for the other models, indicating that the energy values of the true source part are greater than those of interference and algorithmic artifacts. Furthermore, BACPPNet shows the highest STOI, demonstrating a stronger correlation between the clean anechoic target and a short time envelope of the mixture/separated signal. The above results demonstrate that applying the Deeplabv3plus model to bird sound source separation is effective. The BACPPNet outperforms all compared models in terms of feature extraction capability.

D. Effect of the PSA Mechanism

We conduct some experiments to explore whether introducing the PSA module will enhance the performance. In this section, two other commonly used attention modules—that is, the SE [48] module and CBAM modules [51]—are selected for comparison. SE is a channel-only attention mechanism that does not depend on the position and geometric structure of the microphone array, and it can capture information about input features across channels and models the interdependencies between input channels [65]. CBAM is an efficient attention mechanism that includes a channel attention module and a spatial attention module. As a dual attention mechanism, CBAM can focus on important information in both the channel

TABLE I
GLOSSARY

Abbr.	Definition	Abbr.	Definition
ASPP	Atrous Spatial Pyramid Pooling	LSTM-RNN	Long-Short Term Memory Recurrent Neural Network
BACPP	Bird Acoustics Cocktail Party Problem	MUSIC	MUltiple SIgnal Classification
CBAM	Convolutional Block Attention Module	NMF	Non-negative Matrix Factorization
CCP	Cocktail Party Problem	Orig-Acc	Original-Accuracy
CNN	Convolutional Neural Network	PAM	Passive Acoustic Monitoring
CRNN	Convolutional Recurrent Neural Network	PSA	Polarized Self-Attention
DBA	Dual-Branch Attention	RMSE	Root Mean Square Error
DCNN	Deep Convolutional Neural Network	SAR	Source-to-Artifact Ratio
DenseASPP	Densely connected Atrous Spatial Pyramid Pooling	SDR	Source-to-Distortion Ratio
DFANet	Deep Feature Aggregation Network	SE	Squeeze-and-Excitation networks
DL	Deep Learning	SIR	Source-to-Interference Ratio
GHDSS	Geometric High-order Decorrelation-based Source Separation	SMM	Spectral amplitude Mask
ICA	Independent Component Analysis	STOI	Short-Time Objective Intelligibility
IoT	Internet of Things	TDOA	Time Difference Of Arrival
IPD	Inter-channel Phase Difference	TF	Time-Frequency
IVA	Independent Vector Analysis	TSNF	Temporal-Spatial Neural Filter
LSTM	Long-Short Term Memory		

TABLE II
COMPARISON OF RESULTS OF DIFFERENT MODELS

Model	RMSE	SDR	SIR	SAR	STOI
DFANet[41]	6.44	7.15	25.24	8.46	0.57
Denseaspp[42]	3.57	6.29	27.16	8.26	0.59
U-Net[64]	3.97	9.42	29.15	10.28	0.64
CRNN[38]	2.90	9.29	29.43	10.12	0.65
Deeplabv3plus[21]	2.90	9.92	29.83	11.05	0.66
BACPPNet(proposed)	**2.82**	**10.00**	**29.90**	**11.08**	**0.66**

and spatial dimensions of the input features. Table II displays the experimental results on the test set.

As shown in Table III, after introducing SE and CBAM to Deeplabv3plus, the values of some evaluation metrics improve while others decrease. The CBAM module outperforms the SE module in terms of overall performance. This might be because the SE module solely focuses on channel features, whereas the CBAM module focuses on features in both the channel and spatial dimensions. PSA significantly collapses features in one direction while keeping high resolution in the orthogonal direction in the self-attention branch. The channel-only and spatial-only self-attention branches can thus attain higher channel and spatial resolutions. Higher resolutions facilitate better pixel-wise feature quality than lower resolutions [24]. As a result, PSA outperforms CBAM in our task, and the BACPPNet performs better than all of those compared methods.

Fig. 9, which analyzes a typical separation example, displays the ground truth and the separated spectrograms predicted by different separation models in the first four directions (0°, 45°, 90°, and 135°). As shown in Fig. 9, the U-Net and CRNN projected spurious components in the 0°–90° and 135°–180° ranges. In 90°–135°, however, certain parts of the spectrogram are missing. There are no spurious components predicted for the other models. However, when we focus on the spectrograms' edges, we see tiny differences between the predicted spectrograms. The edges of predicted spectrograms of BACPPNet are more exact than those of other models, especially in the 90°–180° range, where they are closest to the spectrograms of ground truth among the comparison models. In conclusion, the BACPPNet outperforms other compared methods in terms of bird sound source separation.

E. Comparison of Different Input Features

We further investigate the use of diverse inputs to validate the improvement of separation performance owing to the spatial features of bird sound sources. Table IV compares the separation performances of different models using different features.

According to Table IV, when spectrogram, sinIPD, and cosIPD features are fused as inputs, the performance of all the models is better than when only spectral features are used. The RMSE of the Deeplabv3plus model with fused features is reduced by 67.34% compared with the spectral features. The RMSE of the BACPPNet model with fused features is reduced by 65.61% compared with the spectral features. Additionally, the SDR, SIR, SAR, and STOI values are all improved compared with the spectral features. The results show that sinIPD and cosIPD can help models to predict SMMs more accurately. Because the eight-channel mixture signals contain the position relationship between the sound sources, the sinIPD and cosIPD appropriately represent these spatial associations. The experimental results show that combining interchannel spatial features of sound signals as complementary information to spectral features can effectively improve the quality of bird sound source separation.

F. Results of Classification Task

To evaluate the fidelity of the separated sources relative to the ground truth sources, we first use the ground truth sources (single-bird-sound sources) of the training set to train a classification model with good performance. Then, we test the trained CS-CLDNN model with the ground truth sources of the test set. Finally, we compare the classification accuracy of the ground truth sources of the test set and the separated sources estimated using different separation models. The classification accuracy reflects the fidelity of the separated signal relative to the ground truth sources. We

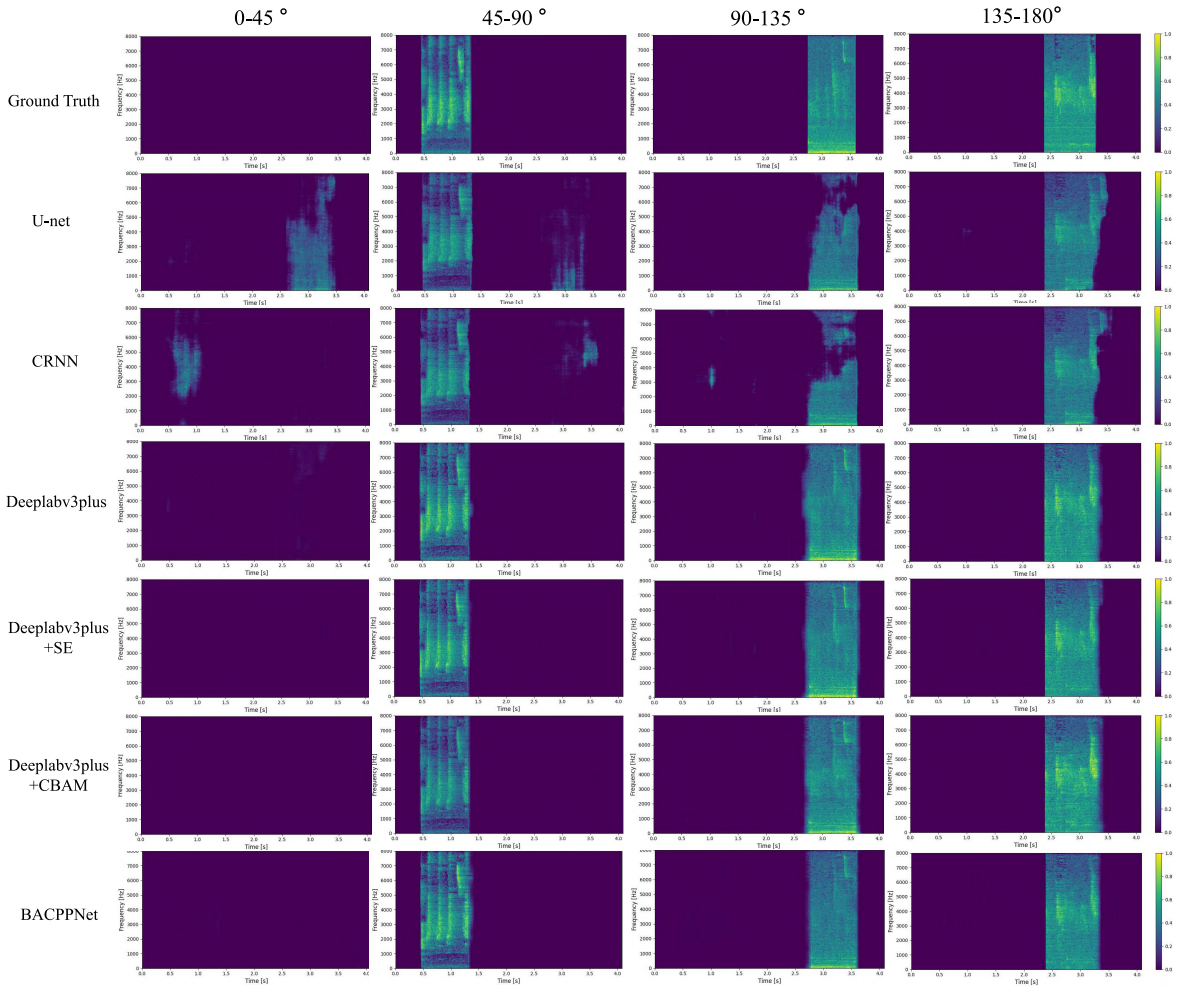


Fig. 9. Visualizations of separation results: The first row is the ground truth spectrogram in the first four directions [0, 180°]; The second to seventh rows are the spectrograms estimated by U-Net, CRNN, Deeplabv3plus, Deeplabv3plus+SE, Deeplabv3plus+CBAM, and BACPPNet, respectively.

TABLE III
COMPARISON OF RESULTS WITH DIFFERENT ATTENTION MODULES

Model	RMSE	SDR	SIR	SAR	STOI
Deeplabv3plus	2.90	9.92	29.83	11.05	0.66
Deeplabv3plus+SE	2.86	8.58	28.07	10.01	0.65
Deeplabv3plus+CBAM	2.87	9.49	29.46	10.63	0.66
BACPPNet(proposed)	**2.82**	**10.00**	**29.90**	**11.08**	**0.66**

present the classification accuracy of the separated sound sources predicted using our separation model. The results of the classification accuracies are shown in Table IV.

As presented in Table V, the average classification accuracy of the ground truth (cf. to Orig-Acc in Table V) is 96.18%, and the average classification accuracy of the BACPPNet decreases by 5.38% compared with the ground truth. According to the classification accuracy of specific species, the classification accuracy of the *Great Reed-Warbler* produced by the BACPPNet model is even better than that based on the unmixed vocalization ground truth. As shown in Fig. 10(a), the spectrum of the *Great Reed-Warbler* is distributed over a wide range of frequency bands. Moreover, because the two main call types operate in different frequency bands, the distortion from the separation has less effect on the distinguishing features of the species.

As expected, the BACPPNet model reduces the classification performances of *Common Cuckoos* and *Magpies*. According to Fig. 10(b), the call frequency of the *Common Cuckoo* is the lowest of the three classes of birds, and its spectrum is primarily distributed in the low-frequency band, resulting in a significant loss of spectrum features during the separation step. Consequently, the classification accuracy is reduced the most. Conversely, for the *Magpie*, the energy value of spectrograms is more intensive than for other species; this requires more accurate mask predictions from the separation models, which also slightly reduces the classification accuracy after separation. However, these two bird species still have higher classification accuracies than the other models.

Overall, BACPPNet outperforms other compared separation models in bird sound separation and maintains the fidelity of the separated sound sources.

VI. CONCLUSION AND FUTURE WORK

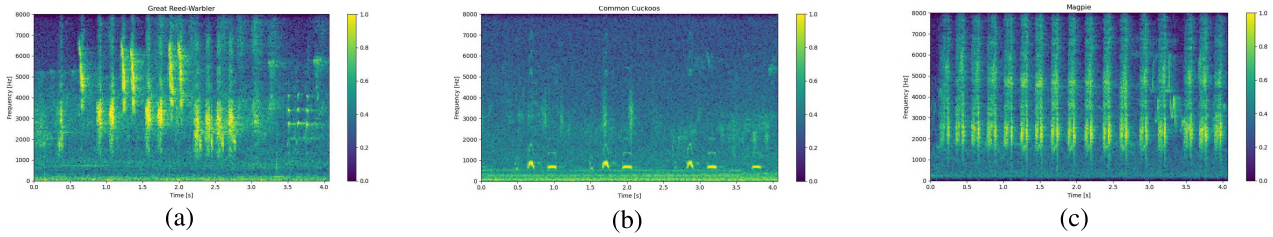
This study proposed a BACPPNet to realize bird sound source separation of multichannel mixture audio signals. First,

TABLE IV
COMPARISON OF RESULTS OF DIFFERENT INPUT FEATURES

Model	Input features	RMSE	SDR	SIR	SAR	STOI
U-Net	spectrogram	15.04	7.55	22.51	9.32	0.59
	spectrogram+sinIPD+cosIPD	**3.97**	**9.42**	**29.15**	**10.28**	**0.64**
CRNN	spectrogram	10.01	8.20	29.41	10.09	0.59
	spectrogram+sinIPD+cosIPD	**2.90**	**9.29**	**29.43**	**10.12**	**0.65**
Deeplabv3plus	spectrogram	8.88	5.33	26.72	9.31	0.60
	spectrogram+sinIPD+cosIPD	**2.90**	**9.92**	**29.83**	**11.05**	**0.66**
BACPPNet	spectrogram	8.20	8.63	29.89	10.48	0.61
	spectrogram+sinIPD+cosIPD	**2.82**	**10.00**	**29.90**	**11.08**	**0.66**

TABLE V
COMPARISON OF CLASSIFICATION ACCURACIES OF SOUND SOURCES ESTIMATED BY DIFFERENT SEPARATION MODELS. WE LIST THE CLASSIFICATION ACCURACIES OF EACH BIRD SPECIES AS WELL AS THEIR AVERAGE. THE PERFORMANCE DROPS (⇓) AND PERFORMANCE RAISES (⇑) ARE OBTAINED WITH RESPECT TO THE ORIGINAL PERFORMANCE (ORIG-ACC). THE BEST VALUES ARE **HIGHLIGHTED**

Model	*Great Reed-Warbler*	*Common Cuckoos*	*Magpie*	Average
Orig-Acc	95.31	97.37	95.88	96.18
DFANet	⇓ 3.56	⇓ 59.03	⇓32.92	⇓ 31.77
U-Net	⇓ 0.05	⇓ 30.87	⇓ 19.84	⇓ 17.01
CRNN	⇓ 0.98	⇓ 35.00	⇓ 17.69	⇓ 17.88
Denseaspp	⇓ 2.05	⇓ 13.69	⇓ 9.20	⇓ 11.63
Deeplabv3plus	⇑ 1.79	⇓ 12.11	⇓ 13.92	⇓ 8.07
Deeplabv3plus+SE	⇓ 2.15	⇓ 17.37	⇓ 10.02	⇓ 9.89
Deeplabv3plus+CBAM	⇓ 0.07	⇓ 18.09	⇓ 11.34	⇓ 9.89
BACPPNet(proposed)	⇑ **2.60**	⇓ **10.30**	⇓ **6.80**	⇓ **5.38**



(a) (b) (c)

Fig. 10. Spectrograms of (a) *Great Reed-Warbler*, (b) *Common Cuckoo*, and (c) *Magpie*.

we employed Deeplabv3plus as the backbone model, which uses depth-wise separable convolutions in the encoder to reduce the computational cost and the number of parameters while maintaining superior performance. Because bird sound sources are strongly connected to their locations, the spatial features of multichannel audio are used as inputs. Second, the ASPP structure enabled the model to retain more low-level features, which was useful to extract multiscale features. Furthermore, the PSA module was inserted into the Deeplabv3plus‘s bottleneck, which can maintain higher resolutions in both the channel and the spatial dimensions. This considerably reduced the loss of high resolution features caused by the encoder's down-sampling operations. According to the experimental results, BACPPNet outperforms existing methods and maintained species category features in the reconstructed signals.

However, in addition to the good performances, the proposed separation method has several limitations. We did not investigate situations when several sources occur in the same direction in the current study, which normally makes high-performance source separation more challenging. In future studies, we will investigate our own end-to-end method as a solution to this problem with more bird species utilizing the time difference of arrival (TDOA) of different sound sources.

Overall, our proposed solution is a promising first step toward resolving the BACPP, and it can help extract comprehensive details from sound recordings collected by the developed IoT-based bird monitoring system. Although further research is required before this method is used in practical scenario, its potential for biodiversity studies is clear. We are optimistic that our approach, along with the monitoring system, would be widely utilized in other bio-acoustic studies with overlapping signals.

REFERENCES

[1] T. A. Rhinehart, L. M. Chronister, T. Devlin, and J. Kitzes, "Acoustic localization of terrestrial wildlife: Current practices and future opportunities," *Ecol. Evol.*, vol. 10, no. 13, pp. 6794–6818, 2020.

[2] J. Xie, Y. Zhong, J. Zhang, S. Liu, C. Ding, and A. Triantafyllopoulos, "A review of automatic recognition technology for bird vocalizations in the deep learning era," *Ecol. Informat.*, vol. 73, Mar. 2023, Art. no. 101927.

[3] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "BirdNET: A deep learning solution for avian diversity monitoring," *Ecolog. Inform.*, vol. 61, Mar. 2021, Art. no. 101236.

[4] C. Masco, S. Allesina, D. J. Mennill, and S. Pruett-Jones, "The song overlap null model generator (SONG): A new tool for distinguishing between random and non-random song overlap," *Bioacoustics*, vol. 25, no. 1, pp. 29–40, 2016.

[5] C. Bergler, M. Schmitt, A. Maier, R. X. Cheng, V. Barth, and E. Nöth, "ORCA-PARTY: An automatic killer whale sound type separation toolkit using deep learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 1046–1050.

[6] P. C. Bermant, "BioCPPNet: Automatic bioacoustic source separation with deep neural networks," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, 2021.

[7] M. R. Izadi, R. Stevenson, and L. N. Kloepper, "Separation of overlapping sources in bioacoustic mixtures," *J. Acoust. Soc. Amer.*, vol. 147, no. 3, pp. 1688–1696, 2020.

[8] J.-T. Chien and H.-L. Hsieh, "Convex divergence ICA for blind source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 302–313, 2011.

[9] C. Rohlfing, J. M. Becker, and M. Wien, "NMF-based informed source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2016, pp. 474–478.

[10] Y. Sudo, K. Itoyama, K. Nishida, and K. Nakadai, "Multi-channel environmental sound segmentation," in *Proc. IEEE/SICE Int. Symp. Syst. Integr. (SII)*, 2020, pp. 820–825.

[11] Y. Sudo, K. Itoyama, K. Nishida, and K. Nakadai, "Multichannel environmental sound segmentation," *Appl. Intell.*, vol. 51, no. 11, pp. 8245–8259, 2021.

[12] T.-H. Lin and Y. Tsao, "Source separation in ecoacoustics: A roadmap towards versatile soundscape information retrieval," *Remote Sens. Ecol. Conserv.*, vol. 6, no. 3, pp. 236–247, Sep. 2020.

[13] D. Gabriel, R. Kojima, K. Hoshiba, K. Itoyama, K. Nishida, and K. Nakadai, "2D sound source position estimation using microphone arrays and its application to a VR-based bird song analysis system," *Adv. Robot.*, vol. 33, nos. 7–8, pp. 403–414, 2019.

[14] K. Nakamura, K. Nakadai, and H. G. Okuno, "A real-time super-resolution robot audition system that improves the robustness of simultaneous speech recognition," *Adv. Robot.*, vol. 27, no. 12, pp. 933–945, 2013.

[15] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2020, pp. 6394–6398.

[16] Y. Du et al., "Semi-supervised multichannel speech separation based on a phone-and speaker-aware deep generative model of speech spectrograms," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, 2021, pp. 870–874.

[17] Y. Sudo, K. Itoyama, K. Nishida, and K. Nakadai, "Sound event aware environmental sound segmentation with mask U-Net," *Adv. Robot.*, vol. 34, no. 20, pp. 1280–1290, 2020.

[18] Y. Sudo, K. Itoyama, K. Nishida, and K. Nakadai, "Multi-channel environmental sound segmentation utilizing sound source localization and separation U-Net," in *Proc. IEEE/SICE Int. Symp. Syst. Integr. (SII)*, 2021, pp. 382–387.

[19] Y.-C. Tseng, B. N. I. Eskelson, K. Martin, and V. LeMay, "Automatic bird sound detection: Logistic regression based acoustic occupancy model," *Bioacoustics*, vol. 30, no. 3, pp. 324–340, 2021.

[20] R. Rajan, J. Johnson, and N. A. Kareem, "Bird call classification using dnn-based acoustic modelling," *Circuits, Syst., Signal Process.*, vol. 41, no. 5, pp. 2669–2680, 2022.

[21] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[22] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

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

[24] H. Liu, F. Liu, X. Fan, and D. Huang, "Polarized self-attention: Towards high-quality pixel-wise mapping," *Neurocomputing*, vol. 506, pp. 158–167, Sep. 2022.

[25] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[26] R. Kojima, O. Sugiyama, R. Suzuki, K. Nakadai, and C. E. Taylor, "Semi-automatic bird song analysis by spatial-cue-based integration of sound source detection, localization, separation, and identification," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2016, pp. 1287–1292.

[27] H. Nakajima, G. Ince, K. Nakadai, and Y. Hasegawa, "An easily-configurable robot audition system using histogram-based recursive level estimation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2010, pp. 958–963.

[28] M. Shugaev, N. Tanahashi, P. Dhingra, and U. Patel, "BirdCLEF 2021: Building a birdcall segmentation model based on weak labels," in *Proc. CLEF (Working Notes)*, 2021, pp. 1649–1658.

[29] Y. Dai, J. Yang, Y. Dong, H. Zou, M. Hu, and B. Wang, "Blind source separation-based iva-xception model for bird sound recognition in complex acoustic environments," *Electron. Lett.*, vol. 57, no. 11, pp. 454–456, 2021.

[30] T. Denton, S. Wisdom, and J. R. Hershey, "Improving bird classification with unsupervised sound separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 636–640.

[31] J. Yu et al., "Audio-visual multi-channel integration and recognition of overlapped speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2067–2082, 2021.

[32] Z.-Q. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 457–468, Feb. 2019.

[33] X. Li and R. Horaud, "Multichannel speech enhancement based on time-frequency masking using subband long short-term memory," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 2019, pp. 298–302.

[34] D. Takeuchi, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Invertible DNN-based nonlinear time-frequency transform for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2020, pp. 6644–6648.

[35] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2018, pp. 1–5.

[36] S. Gul, M. S. Khan, and S. W. Shah, "Integration of deep learning with expectation maximization for spatial cue-based speech separation in reverberant conditions," *Appl. Acoust.*, vol. 179, Aug. 2021, Art. no. 108048.

[37] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2112–2121, Dec. 2014.

[38] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proc. Interspeech*, 2018, pp. 3229–3233.

[39] V. S. Kadandale, J. F. Montesinos, G. Haro, and E. Gómez. "Multi-channel u-net for music source separation," in *Proc. IEEE 22nd Int. Workshop Multimed. Signal Process. (MMSP)*, 2020, pp. 1–6.

[40] W. Wang, D. Qin, S. Wang, Y. Fang, and Y. Zheng, "A multi-channel unet framework based on snmf-dcnn for robust heart-lung-sound separation," *Comput. Biol. Med.*, vol. 164, Sep. 2023, Art. no. 107282.

[41] H. Li, P. Xiong, H. Fan, and J. Sun, "DFANet: Deep feature aggregation for real-time semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9522–9531.

[42] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3684–3692.

[43] A. Bosca, A. Guerin, L. Perotin, and S. Kitic. "Dilated u-net based approach for multichannel speech enhancement from first-order ambisonics recordings," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, 2021, pp. 216–220.

[44] C. Jin, M. Kim, S. Jang, and D.-G. Paeng, "Semantic segmentation-based whistle extraction of Indo-Pacific bottlenose dolphin residing at the coast of Jeju island," *Ecolog. Indic.*, vol. 137, Apr. 2022, Art. no. 108792.

[45] J. Li, P. Wang, and Y. Zhang, "DeepLabV3+ vision transformer for visual bird sound denoising," *IEEE Access*, vol. 11, pp. 92540–92549, 2023.

[46] Q.-B. Hong, C.-H. Wu, T. B. Nguyen, and H.-M. Wang, "Improvement of spatial ambiguity in multi-channel speech separation using channel attention," in *Proc. Asia-Pac. Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, 2021, pp. 619–623.

[47] C. Sun et al., "A convolutional recurrent neural network with attention framework for speech separation in monaural recordings," *Sci. Rep.*, vol. 11, no. 1, pp. 1–14, 2021.

[48] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[49] Y. Chen, Y. Hu, L. He, and H. Huang, "Multi-stage music separation network with dual-branch attention and hybrid convolution," *J. Intell. Inf. Syst.*, vol. 59, pp. 1–22, Jun. 2022.

[50] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[51] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[52] E. A. P. Habets, "Room impulse response generator," Technische Universiteit Eindhoven, Eindhoven, The Netherlands, Rep. 2.2.4, 2010. [Online]. Available: https://www.researchgate.net/profile/Emanuel-Habets/publication/259991276_Room_Impulse_Response_Generator/links/5800ea5808ae1d2d72eae2a0/Room-Impulse-Response-Generator.pdf

[53] M. A. Bee and C. Micheyl, "The 'cocktail party problem': What is it? How can it be solved? And why should animal behaviorists study it?" *J. Comp. Psychol.*, vol. 122, no. 3, p. 235, 2008.

[54] L. Gao, R. Zhang, L. Qi, E. Chen, and L. Guan, "The labeled multiple canonical correlation analysis for information fusion," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 375–387, Feb. 2019.

[55] B. Wu and X.-P. Zhang, "Environmental sound classification via time–frequency attention and framewise self-attention-based deep neural networks," *IEEE Internet Things J.*, vol. 9, no. 5, pp. 3416–3428, Mar. 2022.

[56] J. Xie, S. Zhao, X. Li, D. Ni, and J. Zhang, "KD-CLDNN: Lightweight automatic recognition model based on bird vocalization," *Appl. Acoust.*, vol. 188, Jan. 2022, Art. no. 108550.

[57] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1251–1258.

[58] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1925–1934.

[59] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models." in *Proc. ICML*, vol. 30, 2013, p. 3.

[60] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," 2017, *arXiv:1710.05941*.

[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[62] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[63] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[64] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.

[65] Y. Jianming and Z. Sheng, "A multichannel speech separation method for Ad-Hoc microphones," *J. Signal Process.*, vol. 37, no. 5, pp. 757–762, 2021.

Jiangjian Xie received the B.S. degree from China Agricultural University, Beijing, China, in 2007, and the Ph.D. degree from Beijing Jiaotong University, Beijing, in 2013.

He is currently an Associate Professor with Beijing Forestry University, Beijing. His research interest includes intelligent progressing of forestry ecological environment information.

Yuwei Shi received the B.S. degree from the University of Shanxi Agricultural University, Jinzhong, China, in 2021. He is currently pursuing the M.S. degree with the Beijing Forestry University, Beijing, China.

His research interest includes automatic recognition of bird sound.

Dongming Ni received the B.S. degree from the University of Science and Technology Liaoning, Anshan, China, in 2020. He is currently pursuing the M.S. degree with the Beijing Forestry University, Beijing, China.

His research interest includes automatic separation of bird sound.

Manuel Milling received the first Bachelor of Science degree in physics, the second Bachelor of Science degree in computer science, and the Master of Science degree in physics from the University of Augsburg, Augsburg, Germany, in 2014, 2015, and 2018, respectively. He is currently pursuing the Ph.D. degree in computer science from the Chair for Health Informatics, MRI, Technical University of Munich, Munich, Germany.

His research interests include machine learning with a particular focus on the development and application of deep learning methodologies.

Shuo Liu received the bachelor's degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2012, and the M.Sc. degree from the Technical University of Darmstadt, Darmstadt, Germany, in 2017. He is currently pursuing the Ph.D. degree with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany.

He worked as a Researcher with the Sivantos group for hearing aids solutions. His research focuses are deep learning for audio processing, mobile computing, digital health, and affective computing.

Junguo Zhang received the B.S. and M.S. degrees from the China University of Mining and Technology, Xuzhou, China, in 2000 and 2003, respectively, and the Ph.D. degree from Beijing Forestry University, Beijing, China, in 2009.

He visited the Forest Product Laboratory, Madison, WI, USA, in 2012. He is currently the Director of the Department of Automation. He is committed to research on forestry information collection and intelligent processing. In addition, he has led nearly ten scientific projects supported by the National Natural Science Foundation of China and State Forestry Administration.

Kun Qian (Senior Member, IEEE) received the Doctoral degree (Dr.-Ing.) in electrical engineering and information technology (for his study on automatic general audio signal classification) from Technische Universität München, Munich, Germany, in 2018.

Since 2021, he has been appointed as a (Full) Professor with the title of "Teli Young Fellow" with Beijing Institute of Technology, Beijing, China.

Björn W. Schuller (Fellow, IEEE) received the Diploma, Doctoral, and Habilitation degrees in electrical engineering and information technology from Technical University of Munich (TUM), Munich, Germany, in 1999, 2006, and 2006, respectively.

He was entitled Adjunct Teaching Professor in 2012. He is a Full Professor of Artificial Intelligence and the Head of GLAM with Imperial College London, London, U.K., the Chair of the Chair for Health Informatics, MRI, TUM, and a Full Professor with the Munich Center for Machine Learning, Munich, amongst other Professorships and Affiliations. He has (co-)authored more than 1200 publications (more than 50 000 citations, H-index=100+).

Prof. Schuller is the Golden Core Awardee of the IEEE Computer Society, the Fellow of the BCS, ELLIS, and ISCA, and the President-Emeritus of the AAAC, an Elected Full Member Sigma Xi, and a Senior Member of the ACM.