DGSD: Dynamical Graph Self-Distillation for EEG-Based Auditory Spatial Attention Detection

Cunhang Fan, *Member, IEEE,* Hongyu Zhang, Wei Huang, Jun Xue, Jianhua Tao, *Senior Member, IEEE,* Jiangyan Yi, *Member, IEEE,* Zhao Lv, *Member, IEEE,* Xiaopei Wu, *Member, IEEE*

Abstract—Auditory Attention Detection (AAD) aims to detect target speaker from brain signals in a multi-speaker environment. Although EEG-based AAD methods have shown promising results in recent years, current approaches primarily rely on traditional convolutional neural network designed for processing Euclidean data like images. This makes it challenging to handle EEG signals, which possess non-Euclidean characteristics. In order to address this problem, this paper proposes a dynamical graph self-distillation (DGSD) approach for AAD, which does not require speech stimuli as input. Specifically, to effectively represent the non-Euclidean properties of EEG signals, dynamical graph convolutional networks are applied to represent the graph structure of EEG signals, which can also extract crucial features related to auditory spatial attention in EEG signals. In addition, to further improve AAD detection performance, self-distillation, consisting of feature distillation and hierarchical distillation strategies at each layer, is integrated. These strategies leverage features and classification results from the deepest network layers to guide the learning of shallow layers. Our experiments are conducted on two publicly available datasets, KUL and DTU. Under a 1-second time window, we achieve results of 90.0% and 79.6% accuracy on KUL and DTU, respectively. We compare our DGSD method with competitive baselines, and the experimental results indicate that the detection performance of our proposed DGSD method is not only superior to the best reproducible baseline but also significantly reduces the number of trainable parameters by approximately 100 times.

Index Terms—Auditory attention detection, electroencephalography (EEG), dynamical graph convolutional network, self-distillation.

I. INTRODUCTION

The cocktail party problem [1], [2] is an intriguing scenario where multiple speakers' voices are mixed together, much like in a noisy social gathering. The challenge in this problem arises when multiple sound sources are present simultaneously, and we need to find a way to separate and extract the sound source of interest, namely, the target speaker. Multi-speaker speech separation techniques [3], [4] are used to address the aforementioned issue. These techniques aim to decompose mixed speech into different sound sources, allowing us to individually extract the speech of each speaker. However, these techniques cannot extract the target speech without the prior information of the target speaker. To tackle this problem, auditory attention detection (AAD) [5]–[8] has emerged as a highly promising solution. AAD is designed to emulate the "attention" process in the human auditory system using brain signals. With AAD technology, we can identify and locate the target speaker, i.e., the speaker who has captured the listener's attention in a multi-speaker environment. Hearing-impaired individuals often struggle to differentiate between target and interfering sounds in noisy environments due to their hearing disabilities, leading to communication difficulties and emotional challenges. Modern hearing aids [9] incorporate advanced AAD algorithms to assist hearing impaired people in more accurately capturing target sounds, ignoring background noise, and enhancing their listening performance in multi-speaker scenarios.

Electroencephalography (EEG) provides a non-invasive and low-cost technique. Various studies indicate that using EEG for AAD is feasible [10]–[15]. AAD relies on extracting EEG features from the EEG signals, which can be done in the time domain [16]–[18] and frequency domain. Extracting EEG features from the frequency domain allows for a more comprehensive reflection of signal characteristics compared to the time domain. The extraction of frequency domain features is used to identify different frequency bands of brainwave rhythms, such as δ (1-3 Hz), θ (4-7 Hz), α (8-13 Hz), β (14-30 Hz), and γ (31-50 Hz) [19]–[23]. This helps describe the spatial characteristics and functional states of the EEG signals. Subsequently, EEG features can be extracted from each frequency band, including power spectral density (PSD) [24], [25] features, rational asymmetry (RASM) [26] features, differential entropy (DE) [27], [28] features, and so on.

Research on AAD primarily focuses on two paradigms [9]: speaker identification and tracking spatial attention. The former requires both EEG signals and clean auditory stimuli as input [29], [30], while the latter only relies on EEG signals [15], [31]. In this paper, we focus on models that use only EEG signals as input, choosing not to use auditory stimuli for practical reasons. Previous studies [7], [10] often

use clean stimulus as input, but in real-world environments, listeners typically receive mixed speech containing voices from multiple speakers. This makes obtaining clean auditory stimuli challenging, as models might face various challenges when dealing with mixed speech, such as speech separation. Therefore, considering the complexity and feasibility of real-world applications, we choose not to use this approach to ensure the practicality and effectiveness of our model in real-world scenarios.

In recent years, due to findings in neuroscience suggesting that the brain processes auditory stimuli through nonlinear mappings [32], traditional linear AAD methods struggle to handle the nonlinear mappings in the brain, and their decoding performance deteriorates significantly with shorter time windows [33]. Consequently, research has gradually shifted towards nonlinear methods based on EEG [9], with convolutional neural networks (CNNs) [15], [31], [34] being the most commonly used nonlinear approach. When performing auditory spatial attention detection, selecting an appropriate method to model EEG signals is of paramount importance. Unlike Euclidean data such as image pixels, EEG signals exhibit non-uniform sampling due to the uneven measurement locations on the scalp and varying distances between electrodes. Additionally, during the data collection process, electrodes are distributed discretely on the scalp, forming a discrete electrode network rather than a continuous Euclidean space. The primary reason why CNN is unsuitable for processing EEG signals is that CNN is designed to handle Euclidean data, relying on spatial relationships between pixels during the convolutional kernel sliding process. However, the non-uniform sampling points and non-Euclidean spatial characteristics of EEG signals make it challenging to model such spatial relationships effectively [35]. Relatively speaking, graph structures can naturally represent these non-uniform connections, allowing for a better capture of interactions between different electrodes. They are not constrained by a fixed grid structure and are better suited to adapt to the characteristics of EEG signals, thus providing more accurate feature extraction [36].

In this paper, a novel dynamical graph self-distillation (DGSD) method is proposed for auditory spatial attention detection. Initially, dynamical graph convolutional networks (DGCN) are employed to represent EEG signals with non-Euclidean features as a graph structure, where each node corresponds to an electrode location, and the adjacency matrix represents the connectivity between electrodes. Subsequently, graph convolution operations within the network are utilized to extract essential features related to auditory spatial attention. These operations propagate information between electrodes and dynamically update the feature representation of electrodes using information from neighboring electrodes. Additionally, self-distillation methods are integrated, applying feature distillation and hierarchical distillation strategies after each layer of graph convolution operations. This involves using features and classification results from the deepest layer to guide the learning of shallower layers, further enhancing the model's performance.

The main contributions of this paper lie in two aspects.

Firstly, the DGCN is applied to represent EEG signals with non-Euclidean characteristics and capture crucial features related to auditory spatial attention. Secondly, the self-distillation is integrated to further improve the model's performance, which consists of feature distillation and hierarchical distillation strategies at each layer. Experiments are conducted on publicly available datasets from KUL and DTU. The experimental results demonstrate that the proposed DGSD approach not only outperforms state-of-the-art reproducible AAD methods but also significantly reduces the number of trainable parameters by approximately 100 times.

The rest of this paper is organized as follows. Section II presents a brief overview of the relevant work related to this paper. Section III introduces the proposed DGSD method. The experimental setup is stated in Section IV. Section V shows experimental results. Section VI shows the discussions. Section VII draws conclusions.

II. RELATED WORK

A. Nonlinear methods for AAD

Currently, advanced AAD research methods can be categorized into two types. One type is related to spatial localization detection, such as [15], [31], [37], both of which use EEG signals as input and employ CNN for spatial feature extraction. In [31], the authors propose the SSF-CNN method for auditory spatial attention detection, which combines spectral spatial features (SSF) constructed by analyzing the topographical specificity of α band power in EEG with CNN to enhance detection performance. In [15], the authors found that using SSF-CNN with only the α band power spectrum couldn't fully reflect the spatial information of EEG, so they extracted multi-band differential entropy features as input to CNN, utilizing information from different frequency bands in EEG signals to improve performance. In [37], the authors propose an EEG-graphs convolutional network that incorporates a neural attention mechanism. It takes EEG data from a single β frequency band as input, and this mechanism simulates the topological structure of the human brain based on the spatial patterns of EEG signals. The other type is speaker identification, as in [38]–[40], which uses both EEG signals and auditory stimuli as input. In [38]–[40], the authors introduce a joint CNN-LSTM model, which takes EEG signals and stimulus spectrograms as inputs for identity recognition, improving performance by capturing long-term dependencies between EEG responses and auditory stimuli using long short-term memory (LSTM).

B. Spectral graph filtering

Spectral filtering [43], also known as graph convolution, is a widely used signal processing technique in graph data operations. The basic idea is to represent graph data as a signal in a specific domain, such as the frequency domain or spectral domain. In these domains, the graph fourier transform (GFT) [48] can be used for graph signal analysis. Recently, spectral filtering has been widely applied in graph neural network (GNN) to form graph convolutional network (GCN) [41], [42], [44], [49], [50], which can extract features of
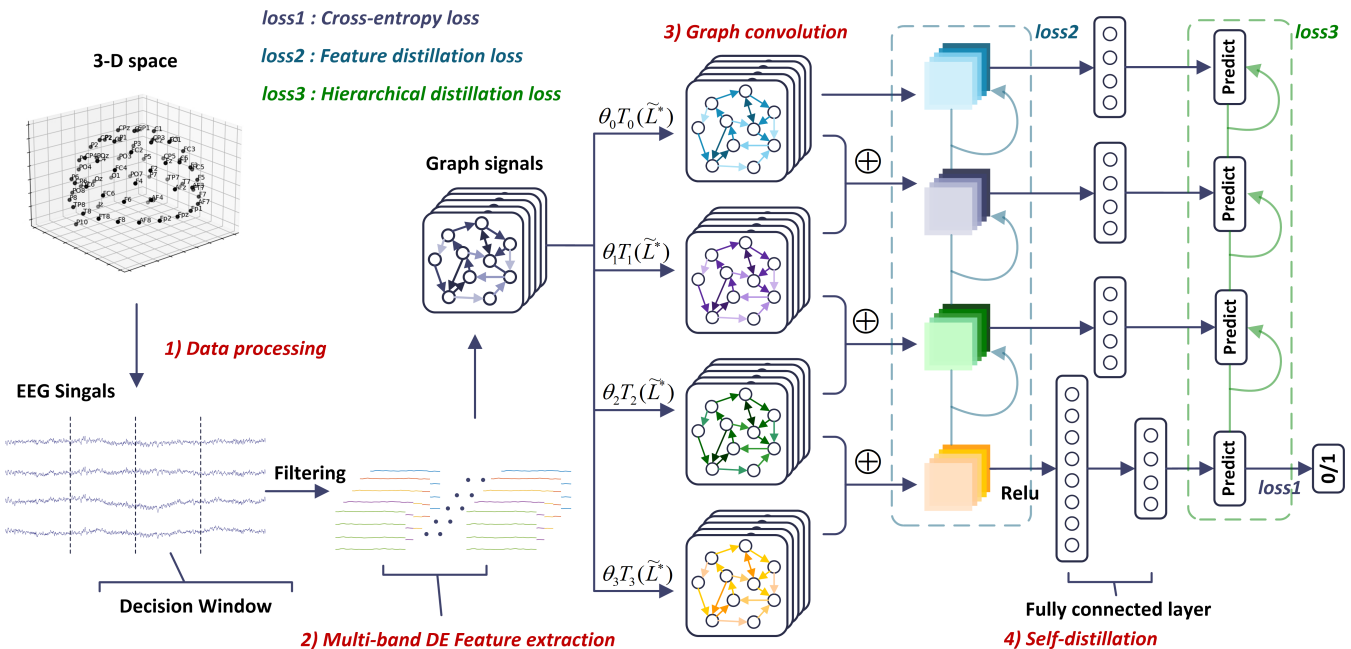
Fig. 1. DGSD model architecture. DGSD consists of four modules :1) EEG data processing, 2) Multi-band DE feature extraction, 3) Graph convolution operation and 4) Self-distillation. The fusion of the latter two modules is the core of our model. The graph structure is used to represent EEG signals, and then the graph convolution of each layer is used to extract features about auditory spatial attention from EEG signals, while the features and classification results of the deepest network are used to guide the learning of the shallow network.

nodes and edges as a convolution operation. In [51], [54], the authors first propose a domain-based hierarchical clustering or graph Laplacian spectrum GCN, which can handle signals on irregular graph structures such as social networks and brain connectomes. In [23], the authors introduce a dynamical graph convolutional neural networks (DGCNN) approach that incorporates GCN into an emotion recognition system based on multi-channel EEG. This method dynamically learns the intrinsic relationships between different EEG channels. This indicates GCN have great potential in extracting features of discrete spatial domain signals [41].

C. Self-distillation

Self-distillation, as an emerging method, has been applied in various fields such as speech recognition and computer vision [45], [46]. In [52], the proposed instance segmentation network is trained and its detection accuracy is improved by applying self-distillation. In [53], an elegant self-distillation mechanism is proposed to directly obtain high-precision models. In [47], a self-distillation method for fake speech detection is proposed, which uses the deepest network to guide and enhance the shallow network, and builds a distillation path between the features of the deepest and shallow networks to reduce feature differences. This method can significantly improve the performance of FSD. These methods demonstrate that self-distillation, as an effective knowledge transfer and model training method, has broad application prospects in different fields and tasks.

III. THE PROPOSED DGSD METHOD

In this section, we introduce our proposed DGSD model. This method not only effectively represents EEG data with

non-Euclidean characteristics as graph signals but also extracts crucial features related to auditory spatial attention using graph convolution operations. Furthermore, by combining with self-distillation, the model can enhance detection accuracy, which enables the features and classification results from the deepest network layer to guide the shallow network learning through feature distillation and hierarchical distillation in self-distillation. The framework of our proposed model is illustrated in Fig. 1, which consists of four modules: EEG data processing, multi-band DE feature extraction, graph convolution operation and self-distillation. Next, we provide detailed descriptions of these modules.

A. EEG data processing & Multi-band DE feature extraction

Many studies use the sliding window method to segment EEG signals into a series of time periods for performance analysis of different AAD algorithms [15], [31], [33], [34], [38]. In this study, we process the data according to the final frequency of EEG preprocessing for each dataset, and perform sliding window processing for each subject's data to extract multi-band DE features in each time segment.

Next, we perform frequency band decomposition on the EEG data after sliding window. It is decomposed into five frequency bands, allowing for a comprehensive description of the spatial characteristics and functional states of the EEG signal. We then extract multi-band DE features [27], [55] from each frequency band. As a result, for the input EEG signal with 64 channels, we obtain a total of 320 DE features, consisting of 64 channels across the five frequency bands.

B. Dynamical graph convolution network

The dynamical graph convolutional network (DGCN) represents EEG signals with non-Euclidean properties as a graph structure for subsequent graph convolution operations. Then, it can utilize graph convolution operations to extract important features related to auditory spatial attention. This network can obtain more discriminative features by dynamically updating the adjacency matrix within the graph structure. Further information about the dynamic updating of the adjacency matrix can be found in Section III-C2.

1) Graph representation of EEG signals: At this point, we can construct a graph $\mathbf{G}$ using the outputs of the multi-band differential entropy extraction module, which serves as the input for the graph distillation module. In $\mathbf{G} = \{\mathbf{V}, \mathbf{W}\}$, $\mathbf{V} = \{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_N}\}$ is the node set where each node $\mathbf{v_i}$ corresponds to an electrode and $\mathbf{N}$ is the number of electrodes in the EEG recording equipment. $\mathbf{W} \in \mathbb{R}^{\mathbf{N} \times \mathbf{N}}$ is the adjacency matrix of $\mathbf{G}$, with non-negative elements $\mathbf{w_{ij}}$ representing the strength of functional connection between $\mathbf{v_i}$ and $\mathbf{v_j}$. Each node is associated with $\mathbf{d}$ features, i.e., the feature matrix $\mathbf{x} \in \mathbb{R}^{\mathbf{N} \times \mathbf{d}}$ of the nodes. Each column of $\mathbf{x}$ represents a signal defined on the node. Our next step is to perform operations on $\mathbf{x}$.

2) Graph convolution: Specifically, the Laplacian matrix of graph $\mathbf{G} = \{\mathbf{V}, \mathbf{W}\}$ is $\mathbf{L} = \mathbf{D} - \mathbf{W}$ (where $\mathbf{D}$ is a diagonal matrix with elements $\mathbf{D_{ii}} = \sum_{\mathbf{j=1}}^{\mathbf{N}} \mathbf{w_{ij}}$), its eigenvector matrix is $\mathbf{U} = [\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_N}]$ and eigenvalue matrix is $\mathbf{\Lambda} = \mathrm{diag}([\lambda_1, \lambda_2, \ldots, \lambda_N])$, which can be obtained through the singular value decomposition of $\mathbf{L}$, i.e., $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U^T}$. Then the Fourier transform of $\mathbf{x}$ in the graph domain can be expressed as $\hat{\mathbf{x}} = \mathbf{U^T}\mathbf{x}$. The graph convolution operator is defined in the graph domain as:

$$\mathbf{x} * \mathbf{y} = \mathbf{U}\left[(\mathbf{U^T}\mathbf{x}) \odot (\mathbf{U^T}\mathbf{y})\right] \qquad (1)$$

where $\odot$ denotes the element-wise Hadamard product.

The key to graph convolution (also known as spectral filtering) is how to choose the filter $\mathbf{g}$ to adjust the Fourier coefficients of the signal $\mathbf{x}$ in the spectral domain, and thus control the response of the signal at different frequencies. Typically, the filter $\mathbf{g}$ is a diagonal matrix whose diagonal elements represent the weights at different frequencies, i.e., $\mathbf{g}(\mathbf{\Lambda}) = \mathrm{diag}([\theta_1, \theta_2, \ldots, \theta_N])$, where $\{\theta_i\}_{\mathbf{i=1}}^{\mathbf{N}}$ is the vector of Fourier coefficients. Therefore, for a signal $\mathbf{x}$ that has been processed by the filter $\mathbf{g}(\mathbf{L})$, its Fourier coefficients can be represented as:

$$\mathbf{y} = \mathbf{g}(\mathbf{L})\mathbf{x} = \mathbf{g}(\mathbf{U}\mathbf{\Lambda}\mathbf{U^T})\mathbf{x} = \mathbf{U}\mathbf{g}(\mathbf{\Lambda})\mathbf{U^T}\mathbf{x} \qquad (2)$$

This process can be seen as a convolution operation in the spectral domain, that is:

$$\begin{aligned} \mathbf{y} &= \mathbf{U}\mathbf{g}(\mathbf{\Lambda})\mathbf{U^T}\mathbf{x} = [\mathbf{U}\mathbf{g}(\mathbf{\Lambda})] \odot (\mathbf{U^T}\mathbf{x}) \\ &= \mathbf{U}\{\mathbf{U^T}[\mathbf{U}\mathbf{g}(\mathbf{\Lambda})]\} \odot (\mathbf{U^T}\mathbf{x}) = \mathbf{x} * [\mathbf{U}\mathbf{g}(\mathbf{\Lambda})] \end{aligned} \qquad (3)$$

where $\mathbf{U}\mathbf{g}(\mathbf{\Lambda})$ is the convolution kernel, called graph convolution operator. In this way, the signal $\mathbf{x}$ can be transformed from the graph spatial domain to the graph spectral domain, and then convolved with the graph convolution operator, to obtain the signal $\mathbf{y}$ processed by the filter $\mathbf{g}$.

We adopt a similar graph convolution representation as in [23], which is the graph convolution modified by K-order Chebyshev polynomials. The convolution kernel formula is as follows:

$$\mathbf{g}(\mathbf{\Lambda}) \approx \sum_{\mathbf{k=0}}^{\mathbf{K-1}} \theta_{\mathbf{k}} \mathbf{T_k}(\tilde{\mathbf{\Lambda}}) \qquad (4)$$

where $\theta_{\mathbf{k}}$ is the coefficient of the Chebyshev polynomial, $\tilde{\mathbf{\Lambda}}$ is normalized by $\mathbf{\Lambda}$. $\mathbf{T_k}(\tilde{\mathbf{\Lambda}})$ is the K-order Chebyshev polynomial used for evaluating $\tilde{\mathbf{\Lambda}}$.

Currently, our graph convolution is capable of extracting features related to auditory spatial attention. In this module, we design a 4-layer graph convolution, and each layer of signal $\mathbf{x}$ undergoes graph convolution as $\mathbf{x_i} = \mathbf{x_{i-1}} + \mathbf{DGCN}(\mathbf{x}, \mathbf{A_i})$, where $\mathbf{A}$ is an $\tilde{\mathbf{\Lambda}}$ matrix generated by K-order Chebyshev polynomial. Next, we provide a detailed explanation of our self-distillation strategy incorporated in the graph convolution layers and the dynamic updating process of the adjacency matrix.

C. Self-distillation

To further enhance the AAD detection performance, we incorporate the self-distillation method, which consists of feature distillation and hierarchical distillation after each DGCN layer. It guides the learning of shallow networks using features and classification results extracted by the deepest network, thereby extracting more suitable classification features for the AAD task.

1) The calculation of the loss composition: We use the cross-entropy loss function to calculate the loss between the convolution result of the four-layer graph and the true label. This loss is the main loss function for classification, expressed as $\mathbf{loss1}$:

$$\mathbf{loss1} = \mathbf{CrossEntropy}(\mathbf{p_n}, \mathbf{y}) \qquad (5)$$

where $\mathbf{p_n}$ is the output of the deepest DGCN, with $\mathbf{n}$ being set to 4 in this paper. $\mathbf{y}$ is the label of the training dataset. The process of applying average pooling after each graph convolution layer to extract task-related essential features can be described as follows:

$$\mathbf{F_i} = \mathbf{avgpool}(\mathbf{conv}(\mathbf{x_i})) \qquad (6)$$

Utilizing the features extracted from the shallow and deepest layers of DGCN, a feature distillation loss is generated using the L2 function. This loss encourages the shallow-layer features to adapt to the deepest layer features, while using the deepest layer features to guide the learning of shallow-layer features. As shown in Fig. 1, this results in $\mathbf{loss2}$. In this way, when predicting classification results, the shallow DGCN can better align with the outcomes of the deepest DGCN. The calculation formula for $\mathbf{loss2}$ is as follows:

$$\mathbf{loss2} = \sum_{\mathbf{i=1}}^{\mathbf{n-1}} \mathbf{L2}(\mathbf{F_i}, \mathbf{F_n}) \qquad (7)$$

where $\mathbf{F_i}$ is the output feature of each shallow DGCN, and $\mathbf{F_n}$ is the output feature of the deepest DGCN. $\mathbf{L2}$ is the L2 loss

function. In addition, we design a classifier for $\mathbf{x}$ after each layer of graph convolution, which generates $\mathbf{M}$ classification results for subsequent guidance of the deepest DGCN on the shallow DGCN in a global sense. It is worth noting that these classifiers are only used for training and not used in validation and testing phases. We take the deepest DGCN (i.e., the fourth layer) as the teacher model and the first three DGCNs as the student model. Then we use KL divergence to calculate the hierarchical distillation loss in the teacher-student model, i.e., **loss3** in Fig. 1, which can obtain the difference between the two output distributions and better guide the shallow network in learning features. The calculation formula of **loss3** is as follows:

$$\mathbf{loss3} = \sum_{i=1}^{n-1} \mathbf{KL}(\mathbf{p_i}, \mathbf{p_n}) \tag{8}$$

where $\mathbf{p_i}$ represents the output of each layer in the network after the fully connected classifier, and $\mathbf{KL}$ refers to the KL divergence function. At this point, the training loss consists of three components, where α and β are hyperparameters that balance these three sources of loss. Both hyperparameters have values between 0 and 1. The final **loss** is:

$$\mathbf{loss} = \alpha\mathbf{loss1} + (1-\alpha)\mathbf{loss2} + \beta\mathbf{loss3} \tag{9}$$

2) Dynamic learning of the adjacency matrix $\mathbf{W}$*:* We use the back-propagation (BP) method to iteratively update network parameters during model training to achieve optimal or suboptimal solutions. To dynamically learn the optimal adjacency matrix $\mathbf{W}$ of the DGSD model using the BP method, we must compute the partial derivative of the loss function with respect to $\mathbf{W}$, which is expressed as follows:

$$\frac{\partial \mathbf{loss}}{\partial \mathbf{W}} = \begin{pmatrix} \frac{\partial \mathbf{loss}}{\partial \mathbf{w_{11}}} & \frac{\partial \mathbf{loss}}{\partial \mathbf{w_{12}}} & \cdots & \frac{\partial \mathbf{loss}}{\partial \mathbf{w_{1N}}} \\ \vdots & \vdots & \frac{\partial \mathbf{loss}}{\partial \mathbf{w_{ij}}} & \vdots \\ \frac{\partial \mathbf{loss}}{\partial \mathbf{w_{N1}}} & \frac{\partial \mathbf{loss}}{\partial \mathbf{w_{N2}}} & \cdots & \frac{\partial \mathbf{loss}}{\partial \mathbf{w_{NN}}} \end{pmatrix} \tag{10}$$

where $\mathbf{w_{ij}}$ denotes the element in the $\mathbf{i}$-th row and $\mathbf{j}$-th column of $\mathbf{W}$. By applying the chain rule, we can express the computation of $\mathbf{w_{ij}}$ as follows:

$$\frac{\partial \mathbf{loss}}{\partial \mathbf{w_{ij}}} = \alpha\frac{\partial \mathbf{loss1}}{\partial \mathbf{w_{ij}}} + (1-\alpha)\frac{\partial \mathbf{loss2}}{\partial \mathbf{w_{ij}}} + \beta\frac{\partial \mathbf{loss3}}{\partial \mathbf{w_{ij}}} \tag{11}$$

After obtaining the partial derivative of $\frac{\partial \mathbf{loss}}{\partial \mathbf{W}}$, we can update the optimal adjacency matrix $\mathbf{W}$ using the following rule:

$$\mathbf{W} = (1-\rho)\mathbf{W} + \rho\frac{\partial \mathbf{loss}}{\partial \mathbf{W}} \tag{12}$$

where ρ is the learning rate hyperparameter we set during network training.

The detailed DGSD training algorithm is summarized in Algorithm 1. For the ablation study, we perform it by removing the feature distillation or hierarchical distillation from the self-distillation method.

Algorithm 1 Training Algorithm for Optimal DGSD Model

Input: Graph $\mathbf{G} = \{\mathbf{V}, \mathbf{W}\}$ representing multi-channel EEG signals associated with multiple frequency bands, auditory spatial attention labels $\mathbf{y}$ corresponding to EEG, the number of layers of DGSD model $\mathbf{m}$, the number of Chebyshev polynomial order $\mathbf{K}$, the learning rate ρ;

Output: The optimal adjacency matrix $\mathbf{W}$ and the optimal model parameters of DGSD;

1: Initialize the adjacency matrix $\mathbf{W}$ and model parameters;

2: **repeat**

3:　　Apply ReLU operation to normalize the elements $\mathbf{w_{ij}}$ in $\mathbf{W}$, ensuring $\mathbf{w_{ij}} \geq \mathbf{0}$ for all;

4:　　Calculate and normalize the Laplacian matrix $\mathbf{L}$;

5:　　Calculate the Chebyshev polynomials;

6:　　Extract EEG signal features through each layer of graph convolutional layers;

7:　　Using average pooling to extract more representative EEG features from each layer;

8:　　Calculate binary classification probabilities using a fully connected classifier for each layer;

9:　　Calculate cross-entropy loss (**loss1**), feature distillation loss (**loss2**), and hierarchical distillation loss (**loss3**) using Eq. (5), (7), and (8) respectively;

10:　　Calculate the loss function using Eq. (9);

11:　　Update $\mathbf{W}$ and other model parameters using back-propagation;

12: **until** the iterations satisfy the predefined algorithm convergence condition;

IV. EXPERIMENTS

In this section, we present the experimental details of DGSD. AAD datasets and EEG data preprocessing are briefly described in Section IV-A and Section IV-B, respectively. The evaluation metrics are described in Section IV-C, and implementation details and baseline descriptions are provided in Section IV-D. Additionally, EEG data and their attention direction labels are read from two original EEG public dataset files.

A. AAD Datasets

We validate our proposed method on the following two publicly available datasets, as shown in Table I, with detailed information about KUL and DTU available in [56]–[59].

1) KUL dataset: This dataset contains 64-channel EEG data from 16 subjects, with an equal gender distribution (half male, half female). The data were recorded using the BioSemi ActiveTwo system with a sampling rate of 8192 Hz and an electrode layout conforming to the international 10/20 system. Auditory stimuli consisted of four Dutch short stories narrated by a male speaker. During the experiment, each subject was instructed to focus their attention on one of two competing male speakers narrating a story while ignoring the other. Auditory stimuli were presented at a volume of 60 dB through in-ear headphones and filtered with a low-pass cutoff frequency of 4 KHz. Each subject completed 20 trials,

TABLE I
DETAILS OF THE TWO DATASETS.

Dataset	Number of subjects	EEG duration (per subject)	Speakers	Stimulus language	Direction of stimulus
KUL	16	48 min	Male	Dutch short stories	± 90°
DTU	18	50 min	Male and female	Danish audiobooks	± 60°

each lasting 6 minutes. There were two stimulus conditions: "HRTF" or "dry", resulting in a total of eight trials. Auditory stimuli were presented from the left at 90° and from the right at 90° by the two speakers. The presentation order was randomized across subjects. A total of 8 trials, each lasting 6 minutes, were collected for each subject, resulting in 48 minutes of EEG data. More detailed information about this dataset can be found in references [56], [57].

2) DTU dataset: This dataset contains 64-channel EEG data from 18 subjects. The data were recorded using the Biosemi system with a sampling rate of 512 Hz and an electrode layout conforming to the international 10/20 system. Auditory stimuli consisted of Danish audiobooks narrated by both male and female speakers. During the experiment, each subject was required to focus on one of two competing speakers (one male and one female) and ignore the other. To simulate low-reverberation conditions, the recordings of the two competing speakers were interfered with by six additional background speakers (three male and three female). The voices of the respective two speakers were presented from +60° and -60° relative to the subject as sound stimuli at a volume of 65 dB using ER-2 insert earphones with a sampling rate of 48 KHz. Each subject completed a total of 60 trials under three different conditions, with each trial lasting 50 seconds. Consequently, each subject collected 50 minutes of EEG data. For more detailed information about this dataset, please refer to references [58], [59].

B. EEG data preprocessing

Preprocessing of EEG data is different from the EEG data processing in section III-A. This section refers to a series of processing and correction of the raw EEG signals, which can improve the quality of the EEG signals and extract more effective features. Specifically, for the KUL dataset, the EEG signals are bandpass filtered from 0.1 Hz to 50 Hz, downsampled to 128 Hz, and then the brainwave data channels are normalized to ensure zero mean and unit variance across trials. For the DTU dataset, firstly, the line noise and harmonics at 50 Hz in the EEG signals are removed. Secondly, a resampling method based on the fast fourier transform (FFT) is used to downsample the EEG data to 128 Hz. Then, a joint decorrelation framework is used to remove eye artifacts, and a fourth-order forward Butterworth filter is applied to high-pass filter the EEG data at 1.0 Hz [58]. Finally, each trial's EEG data are Z-normalized (also known as Z-score normalization) to ensure that they have unit variance and zero mean for each channel. The purpose of this process is to eliminate scale differences between different channels, allowing all EEG data from each channel to be compared and analyzed on the same scale.

C. Evaluation metrics

Since our task involves detecting spatial direction, specifically left/right (i.e., 0/1), we can think of it as a binary classification task. In our study, we use two evaluation metrics to evaluate the model, the first is accuracy and standard deviation, and the second is precision and recall.

1) Accuracy and standard deviation: Paired t-tests are used to compare the performance differences between two different models at a significance level of 0.05. The mean classification accuracy and standard deviation of all subjects in each dataset are computed under different time windows (0.5s, 1s, 2s, 5s).

2) Precision and recall: In each dataset, under a chance level of 50%, precision and recall are calculated for each subject, followed by the calculation of the mean precision and recall for all subjects.

D. Implementation details

1) Training, validation and testing: We implement the entire experiment using Python 3.7.0 and PyTorch 1.12.1. All experiments are conducted on NVIDIA GeForce RTX 3090 GPU. Our research is evaluated within the subject. After sliding window processing, the data of each subject is randomly divided into training, validation and test sets at a ratio of 8:1:1, and then each subject is trained and tested separately. The data in each table in the paper is the average of all subjects in the dataset. The random seed is set to 1111, batch size is set to 32, and the number of epochs is 200, with Adam used as the optimizer. To adapt to different datasets, the learning rates during the training process for KUL and DTU are set to 0.004 and 0.007, respectively. Two hyperparameters, α and β, are set to 0.7 and 0.3, respectively.

2) Baselines: We use some baselines to evaluate the performance of the DGSD model. In order to ensure the fairness and validity of the performance comparison, the baseline models we compare are also tested on multiple datasets in their respective papers, which guarantees the generalization ability of the baseline models. All [15], [31] models are open-source implementations, and we indicate with an asterisk (*) after the model name in the experimental results, where our replicated results are provided before "/", and the results reported in the baseline model papers are provided after "/". If there is only one result, it is the one we have reproduced because their paper does not contain any experiments on this dataset, and the baseline models we reproduce is performed under the

TABLE II
THE PROPOSED MODEL ACHIEVES AAD ACCURACY (%) ON THE KUL DATASET COMPARED TO THE BASELINE. IN THE EXPERIMENTAL RESULTS OF THE BASELINE MODELS MARKED WITH ”*”, THE VALUE BEFORE ”/” REPRESENTS THE RESULTS WE HAVE REPRODUCED, WHILE THE VALUE AFTER ”/” REPRESENTS THE RESULTS REPORTED IN THE BASELINE MODEL PAPER. ”-” NOTES THAT THERE ARE NO EXPERIMENTS CONDUCTED ON THIS DATASET IN THE PAPER.

Dataset	Model	Use auditory stimuli	Time Window			
			0.5-second	1-second	2-second	5-second
	S-R [10]	Yes	53.9	58.1	61.3	67.5
	CCA [60]	Yes	55.4	59.2	62.4	-
	DNN [7]	Yes	64.9	70.7	74.5	-
	BIAnet [61]	Yes	84.1	84.4	88.1	-
KUL	CNN [34]	No	73.4	80.8	82.1	83.6
	NI-AAD [62]	No	79.4	82.8	87.1	91.2
	SSF-CNN* [31]	No	80.5 ± 8.34 / -	81.9 ± 9.86 / 81.7	87.3 ± 8.79 / 84.7	91.6 ± 7.40 / 90.5
	MBSSFCC* [15]	No	85.0 ± 7.50 / -	88.8 ± 7.80 / 89.2	90.3 ± 7.62 / 91.5	92.8 ± 5.32 / 93.9
	DGSD (ours)	**No**	**86.3 ± 7.89**	**90.3 ± 7.29**	**93.3 ± 6.53**	**94.8 ± 4.61**

TABLE III
THE PROPOSED MODEL ACHIEVES AAD ACCURACY (%) ON THE DTU DATASET COMPARED TO THE BASELINE. IN THE EXPERIMENTAL RESULTS OF THE BASELINE MODELS MARKED WITH ”*”, THE VALUE BEFORE ”/” REPRESENTS THE RESULTS WE HAVE REPRODUCED, WHILE THE VALUE AFTER ”/” REPRESENTS THE RESULTS REPORTED IN THE BASELINE MODEL PAPER. ”-” NOTES THAT THERE ARE NO EXPERIMENTS CONDUCTED ON THIS DATASET IN THE PAPER.

Dataset	Model	Use auditory stimuli	Time Window			
			0.5-second	1-second	2-second	5-second
	S-R [10]	Yes	-	51.8	55.3	-
	CCA [60]	Yes	51.2	53.5	58.9	-
	DNN [7]	Yes	56.8	61.7	62.8	-
	BIAnet [61]	Yes	78.1	79.0	80.6	-
DTU	CNN [34]	No	-	55.9	57.8	58.5
	NI-AAD [62]	No	60.2	61.6	63.2	61.5
	SSF-CNN* [31]	No	63.3 ± 6.42 / -	64.0 ± 7.21 / -	65.5 ± 7.47 / -	68.4 ± 13.89 / -
	MBSSFCC* [15]	No	71.3 ± 5.84 / -	75.2 ± 7.43 / 76.9	78.7 ± 7.86 / 80.6	80.2 ± 8.64 / 82.9
	DGSD (ours)	**No**	**75.6 ± 6.72**	**79.6 ± 6.76**	**82.4 ± 6.86**	**85.6 ± 7.36**

same conditions as our model. As the implementation of [7], [10], [34], [60]–[62] is not yet formally open-sourced, their performance comes from their original papers, and we do not mark these results.

V. RESULTS

A. Low-latency DGSD

To evaluate the feasibility of the DGSD model in practical applications, research is being conducted on the proposed DGSD model with four different time windows. The detection accuracy of the DGSD model on two datasets is reported in Table II and Table III, covering time windows ranging from relatively short durations of 0.5-second to relatively longer durations of 5-second. Table IV also presents the metrics (precision and recall) of the DGSD model for four time windows on both datasets. Additionally, Fig. 2 illustrates the performance of the DGSD model with different time windows for each subject in the two datasets. The horizontal axis is sorted by subject ID, and the vertical axis represents the accuracy of auditory spatial attention detection (starting at 60%). It is observed that in the KUL dataset (Fig. 2a) and the DTU dataset (Fig. 2b).
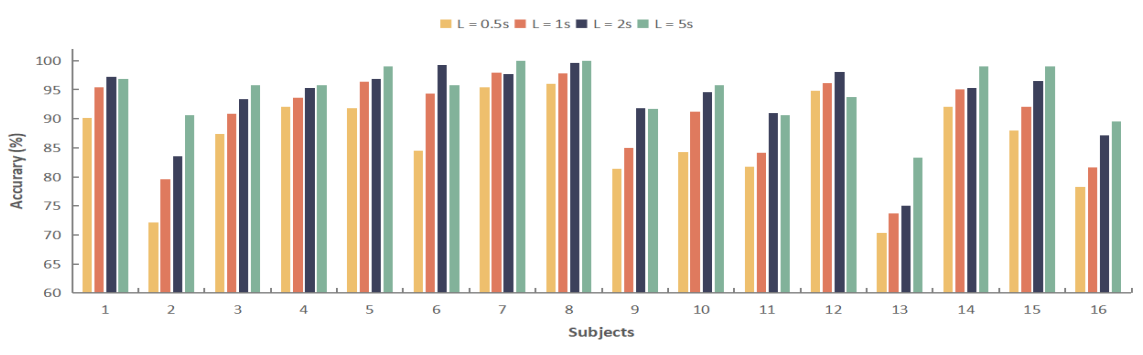
From Table II, it can be observed that in the KUL dataset, the DGSD exhibits excellent auditory attention detection performance under 1-second time window (mean: 90.3%, SD: 7.29%), 2-second time window (mean: 93.3%, SD: 6.53%), and 5-second time window (mean: 94.8%, SD: 4.61%). Moreover, as the time window shortens, the performance of the DGSD model under the 0.5-second time window (mean: 86.3%, SD: 7.89%) decreases with the reduction of EEG signal information, but still maintains a very high detection performance. It can be inferred from the research results that as the time window increases, the detection accuracy of the DGSD model significantly improves, which is consistent with the findings of [15], [61], [62].

From Table III, it can be observed that in the DTU dataset, the DGSD model achieves average accuracies of 75.6% (SD: 6.72%), 79.6% (SD: 6.76%), 82.4% (SD: 6.86%), and 85.6% (SD: 7.36%) for time windows of 0.5-second, 1-second, 2-second, and 5-second, respectively. Similar to the results on the KUL dataset, the trend in results is consistent, indicating an improvement in auditory spatial attention detection accuracy with the increase in time window size.
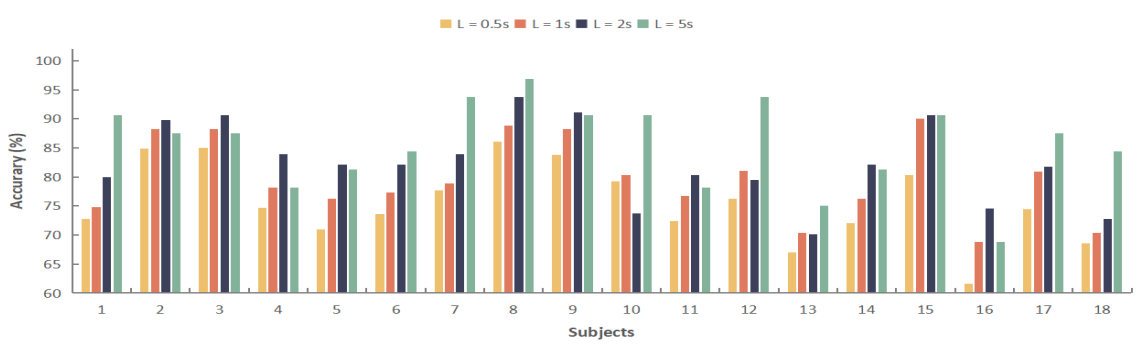
Table IV also showcases the metrics (precision and recall) of the DGSD model with four time windows on both datasets. As

TABLE IV
METRICS (ACCURACY AND RECALL) FOR DIFFERENT TIME WINDOW LENGTHS (0.5-SECOND, 1-SECOND, 2-SECOND, 5-SECOND) IN THE TWO DATASETS.
THE VALUE (%) IN THIS TABLE REPRESENTS THE AVERAGE FOR ALL SUBJECTS IN EACH DATASET.

Dataset	Model	Time Window							
		0.5-second		1-second		2-second		5-second	
		precision	recall	precision	recall	precision	recall	precision	recall
KUL	SSF-CNN* [31]	81.5	78.7	82.1	81.8	86.6	89.3	92.5	90.8
	MBSSFCC* [15]	85.2	84.9	89.1	88.6	90.1	90.9	94.1	91.6
	DGSD (ours)	**86.8**	**85.3**	**89.4**	**89.3**	**93.4**	**93.2**	**94.6**	**95.4**
DTU	SSF-CNN* [31]	63.5	60.2	64.7	58.1	66.6	61.4	69.6	64.9
	MBSSFCC* [15]	70.9	70.6	73.5	73.2	76.0	80.4	79.1	76.9
	DGSD (ours)	**71.3**	**77.0**	**79.6**	**78.6**	**81.2**	**81.7**	**83.6**	**84.1**

(a) KUL dataset

(b) DTU dataset

Fig. 2. Detection accuracy (%) of the DGSD model implemented on each subject in the KUL and DTU datasets for different decision time window lengths (0.5-second, 1-second, 2-second, 5-second). Sort the horizontal axis by the subject IDs. (a) KUL dataset. (b) DTU dataset.

the time window increases, these two metrics also improve. It can be observed that in the KUL dataset (Fig. 2a) and the DTU dataset (Fig. 2b), as the time window L increases from 0.5-second to 5-second, although there are some exceptions, the detection accuracy of most subjects gradually rises. This suggests that as L becomes longer, more information is captured in the EEG signals after sliding window processing, allowing our DGSD model to extract more useful features for auditory attention detection.

However, a noteworthy observation is that the detection performance of the DGSD model on the DTU dataset is lower compared to the KUL dataset, aligning with the findings in studies [15], [34], [61], [62]. Through a analysis of the publicly available descriptions of these two datasets, we suggest that

this may be due to the direction of the auditory stimulus or the gender of the speaker. The primary distinctions between these datasets are as follows:

1) Directional bias of attention (±90° vs ±60°): In the DTU dataset, the two auditory stimuli are distributed at ±60° angles relative to the subjects, while in the KUL dataset, they are distributed at ±90° angles. Subjects might naturally exhibit a more pronounced attention bias towards the ±90° direction, making the auditory stimuli from the KUL dataset potentially more attention-grabbing and thus yielding higher performance.

2) Gender-related influence (Male & female vs Male): The auditory stimuli in the DTU dataset are presented by both male and female speakers, whereas in the KUL dataset, they are presented only by male speakers. Variations in tone and

TABLE V
AAD ACCURACY AND STANDARD DEVIATION (%) OBTAINED FROM THE ABLATION STUDY OF LOSS FUNCTIONS ON DGCN.

Dataset	Loss Function	Time Window			
		0.5-second	1-second	2-second	5-second
KUL	loss1 (DGCN)	85.1 ± 8.18	89.0 ± 7.22	92.6 ± 6.49	93.8 ± 5.49
	loss1 + loss2	85.8 ± 7.49	89.5 ± 7.60	92.2 ± 7.06	93.8 ± 5.82
	loss1 + loss3	86.0 ± 8.14	89.5 ± 8.23	92.1 ± 7.49	94.6 ± 5.08
	loss1 + loss2 + loss3 (DGSD)	**86.3 ± 7.89**	**90.3 ± 7.29**	**93.3 ± 6.53**	**94.8 ± 4.61**
DTU	loss1 (DGCN)	73.8 ± 6.95	79.1 ± 7.12	81.3 ± 6.68	83.6 ± 9.26
	loss1 + loss2	73.1 ± 8.55	78.9 ± 6.13	81.7 ± 7.48	84.9 ± 6.36
	loss1 + loss3	73.0 ± 9.03	78.8 ± 6.36	80.9 ± 6.90	83.7 ± 7.55
	loss1 + loss2 + loss3 (DGSD)	**75.6 ± 6.72**	**79.6 ± 6.76**	**82.4 ± 6.86**	**85.6 ± 7.36**

frequency may exist between male and female voices. The auditory stimuli from both male and female speakers in the DTU dataset could possess distinct voice characteristics. This divergence might impact subjects' attention biases towards stimuli of different genders.

Therefore, we consider that researching the DTU dataset could present more challenges.

B. Ablation study

In order to evaluate the effectiveness of self-distillation (SD) in our DGSD model, we conduct an ablation study on the loss functions, investigating the impact of different combinations of loss functions. During the study, considering the components of the loss function "loss" in our approach: loss1 (cross-entropy loss), loss2 (feature distillation loss), and loss3 (hierarchical distillation loss), we employ loss1 as the primary loss and examine the performance when combined with loss2 or loss3 at specific proportions. Specific results are shown in Table V. It can be observed that:

1) Only DGCN: When using only DGCN (i.e., using loss1 as the sole loss function), its detection accuracy across different time windows surpasses the baseline models in Table II and Table III. This suggests that the DGCN can suit to use graph structure to represent with a nature of non-Euclidean EEG signals, and can effectively extract and utilize the feature information in EEG signals, resulting in higher detection accuracy.

2) Combination of loss functions: From experimental results, it can be seen that combining loss1 with either of the other two loss functions has little impact on detection accuracy. However, when all three loss functions are combined proportionally, our DGSD model outperforms DGCN by approximately 1%. This indicates that the combination of the three loss functions is the optimal choice for the task, as they collaborate to provide enhanced performance and effectiveness. The experimental results show that our SD approach, which combines feature distillation and hierarchical distillation, pays more attention to the multi-level representation of features and labels, and can use the features and classification results of the deepest network to guide the learning of shallow networks, so that shallow networks are more helpful to extract the features of auditory spatial attention and get the correct classification

results. It is helpful to improve the classification accuracy of auditory attention detection.

C. Selection of hyperparameters

Studies show that it takes approximately 1-second to 2-second for a normal person to shift attention to another speaker [63]. Therefore, suitable parameter combinations for Equation 9 are being sought through parameter tuning of hyperparameters α and β within 1-second to 2-second time windows to achieve optimal detection accuracy. The accuracy (%) under 1-second and 2-second time windows for different parameter combinations is depicted in Fig. 3, and extensive experiments are conducted by separately fixing the values of α and β to obtain these results. Overall, these experiments can be classified into two types:

1) $\alpha = 0.7$: On the KUL dataset (Fig. 3a) and the DTU dataset (Fig. 3c), we fix α at 0.7 and vary the value of β.

2) $\beta = 0.3$: On the KUL dataset (Fig. 3b) and the DTU dataset (Fig. 3d), we fix β at 0.3 and vary the value of α.

It can be seen that, regardless of the subfigure, the detection accuracy of the two datasets in the 1-second and 2-second time windows reaches the optimum when α is 0.7 and β is 0.3. This indicates that our model can effectively utilize spatial information (both local and global) in EEG signals for auditory attention detection under the above-mentioned hyperparameter settings.

VI. DISCUSSION

We believe that our proposed DGSD model not only effectively represents the channels of EEG signals but also adeptly extracts and classifies the relevant auditory attention information within the EEG signals. In this section, we begin by comparing the performance of our proposed DGSD model with models incorporating auditory stimuli. Subsequently, we also compare its performance with models without auditory stimuli, as depicted in Table II and Table III. Moreover, regarding the reproduction of state-of-the-art open-source models, namely SSF-CNN [31] and MBSSFCC [15], we analyze the precision and recall of these models in comparison with the DGSD model under various time windows for the KUL and DTU datasets, as presented in Table IV. We further analyze the trainable parameter counts of the aforementioned models
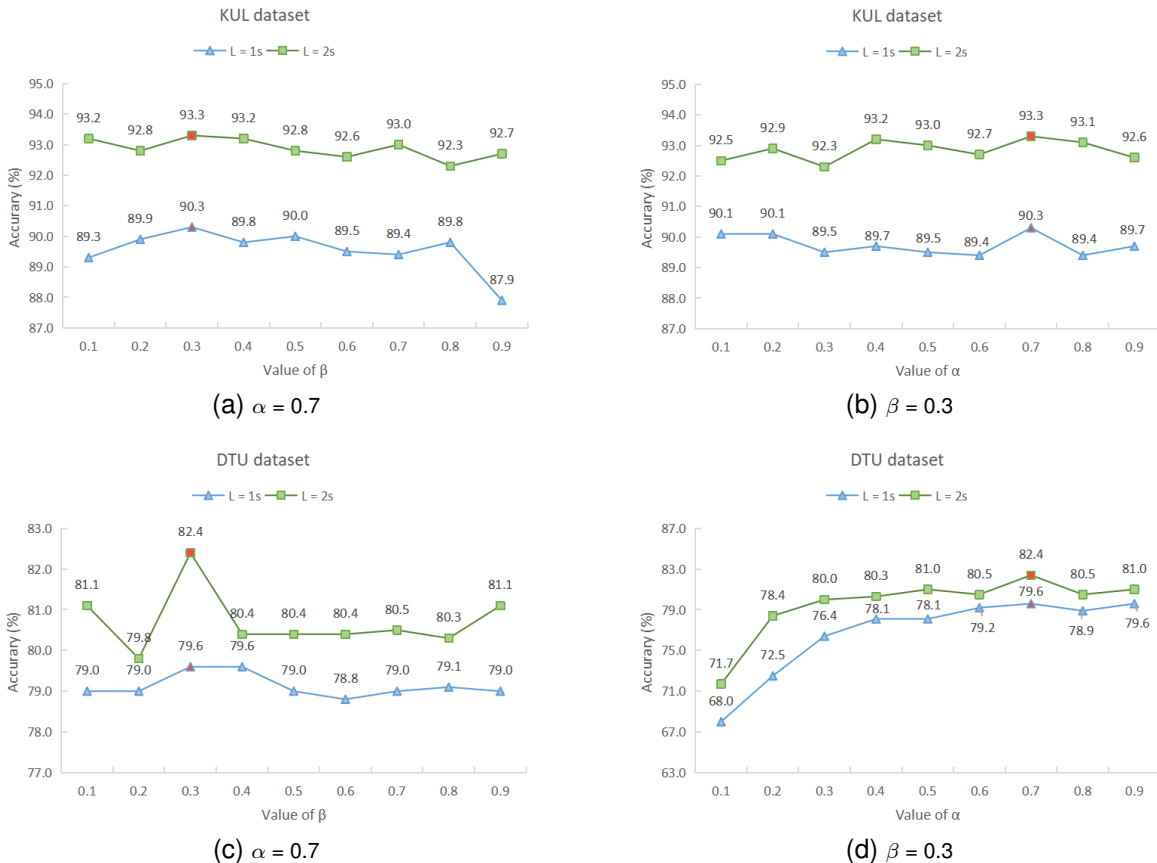
Fig. 3. In two public datasets (KUL, DTU), the impact of different parameter combinations on accuracy (%). Fig. 3a and Fig. 3c depict the accuracy under varying values of β when α is set to 0.7. Fig. 3b and Fig. 3d illustrate the accuracy under different values of α when β is fixed at 0.3. (a) $\alpha = 0.7$. (b) $\beta = 0.3$ (c) $\alpha = 0.7$. (d) $\beta = 0.3$.

as well as our DGSD model, which are detailed in Table VI. Lastly, we provide an interpretation of the results from our conducted ablation experiments on the loss functions. These experimental outcomes are available in Table V.

A. Performance comparison

1) DGSD vs Models (use auditory stimuli): We compare the DGSD model with models that utilize auditory stimuli, which are models with "Use auditory stimuli" values set to "Yes" in Table II and Table III (S-R [10], CCA [60], DNN [7], BIAnet [61]). The results for these models are derived from their respective papers, where a "-" indicates that the experiment for that specific time window is not conducted in the model paper. While this comparison is conducted under different AAD paradigms, their objectives are the same — to identify and enhance auditory stimuli that the listener pays attention to, while attenuating other auditory stimuli that the listener neglects.

From Table II, we observe that on the KUL dataset, DGSD accuracy (mean_0.5-second: 86.3%, mean_1-second: 90.3%, mean_2-second: 93.3%) for time windows of 0.5-second, 1-second, and 2-second significantly surpasses other models using auditory stimuli. Compared to the state-of-the-art auditory stimulus model, BIAnet (mean_0.5-second: 84.1%, mean_1-second: 84.4%, mean_2-second: 88.1%), the DGSD model

achieves an average accuracy improvement of 2.2%, 5.9%, and 5.2% respectively. From Table III, the effectiveness of the model is also verified on the DTU dataset. Although the DGSD model's accuracy is lower than BIAnet for the 0.5-second time window, it outperforms the BIAnet model for the 1-second and 2-second time windows.

The experimental results demonstrate that our proposed DGSD model can achieve higher AAD accuracy without utilizing auditory stimuli, making it more suitable for real-life scenarios.

2) DGSD vs Models (do not use auditory stimuli): We compare models that share the same AAD paradigm with the DGSD model. These models do not require auditory stimuli as inputs, which is more aligned with practical applications. The models not using auditory stimuli are those listed with "Use auditory stimuli" values set to "No" in Table II and Table III (CNN [34], NI-AAD [62], SSF-CNN [31], MBSSFCC [15]), where the results for CNN and NI-AAD models are derived from their respective papers. We focus on comparing the DGSD model with the state-of-the-art open-source SSF-CNN and MBSSFCC models that we have reproduced. The comparison is structured as our experiment results followed by the results from the respective papers.

The results of this comparison on the KUL dataset can be seen in Table II (p <0.001). In the 1-second time window,

TABLE VI
THE TRAINABLE PARAMETER COUNTS OF OUR PROPOSED DGSD MODEL
AND TWO SUPERIOR BASELINE MODELS ARE BEING PRESENTED. "M"
REPRESENTS THE "MILLION" SCALE IN NUMERICAL MAGNITUDE, WHICH
IS EQUIVALENT TO 10^6.

Model	Trainable parameters
SSF-CNN* [31]	4.21M
MBSSFCC* [15]	16.87M
DGSD (ours)	**0.15M**

the DGSD model achieves significantly higher detection accuracy (mean: 90.3%, SD: 7.29%) compared to SSF-CNN and MBSSFCC models, with average improvements of 8.4% and 1.5% respectively. In the 2-second time window, the DGSD model's detection accuracy (mean: 93.3%, SD: 6.53%) is on average 6.0% and 3.0% higher than SSF-CNN and MBSS-FCC. Similarly, the validation on the DTU dataset, as shown in Table III (p <0.001), demonstrates that in the 1-second time window, the DGSD model achieves significantly higher detection accuracy (mean: 79.6%, SD: 6.76%) compared to SSF-CNN and MBSSFCC models, with average improvements of 15.6% and 4.4% respectively. In the 2-second time window, the DGSD model's detection accuracy (mean: 82.4%, SD: 6.86%) is on average 16.9% and 3.7% higher than SSF-CNN and MBSSFCC.

It is evident that even without the use of auditory stimuli, our DGSD model achieves optimal classification detection results. This outcome underscores the effectiveness of our model in auditory spatial attention detection. Additionally, we compute the precision and recall of DGSD, SSF-CNN, and MBSSFCC under different time windows, as shown in Table IV. Across different time windows, the DGSD model outperforms SSF-CNN and MBSSFCC models on both metrics. This shows that the DGSD model is more accurate than the other two models in predicting the left-right spatial direction.

Finally, we compare the trainable parameter counts of the proposed DGSD model, SSF-CNN, and MBSSFCC. As shown in Table VI, the DGSD model achieves higher classification accuracy compared to SSF-CNN and MBSSFCC models, while requiring approximately 28 times fewer parameters than SSF-CNN and 100 times fewer parameters than MBSSFCC. This indicates that under the same setting, our model offers faster training speed and reduced storage requirements. This suggests that our DGSD method is more suitable for practical applications such as hearing AIDS, as it is faster at the same time with high accuracy.

B. Analysis of Self-distillation

Our self-distillation consists of feature distillation and hierarchical distillation. As indicated by Table V, when using only feature distillation (loss2) or hierarchical distillation (loss3) individually, the model performance is moderate. However, when they are combined, the effect improves. We believe this outcome is due to the following reasons:

1) loss1 is combined with loss2/loss3: When using feature distillation alone, since there is no label assistance, the extracted features related to auditory attention may not necessarily be what we desire. When using hierarchical distillation alone, as it lacks the support of features, the obtained labels may not be desirable.

2) loss1 is combined with loss2 and loss3: The combination of both compensates for their respective shortcomings. Feature distillation, through multi-level feature propagation, corrects inaccurate labels that hierarchical distillation might produce. The labels obtained from hierarchical distillation effectively guide feature distillation, emphasizing the extraction of features related to auditory attention direction.

VII. CONCLUSION

This paper introduces a DGSD model that combines self-distillation with dynamic graph convolutional networks. This model does not require auditory stimuli as input and relies solely on EEG signals for auditory spatial attention detection, making it more practical in real-world scenarios. Furthermore, the DGSD model effectively extracts crucial feature information related to auditory spatial attention from the EEG signals, with the self-distillation strategy enhancing the detection performance. The experimental results show that the DGSD model not only outperforms the linear model, but also outperforms the advanced reproducible nonlinear model while reducing the number of trainable parameters by about 100 times, demonstrating the effectiveness of our model in detecting auditory spatial attention. In summary, our DGSD model enhances the performance of EEG-based auditory attention detection and opens up endless possibilities for the development of various hearing devices in the future. Our research is conducted based on within-subject, and there is a lack of cross-subject research. Future work will be extended to cross-subject studies, which will help verify the consistency and robustness of the model.

REFERENCES

[1] E.C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, pp. 975–979, 1953.

[2] S. Haykin and Z. Chen, "The cocktail party problem," *Neural computation*, vol. 17, no. 9, pp. 1875-1902, 2005.

[3] M. Hosseini, L. Celotti, É. Plourde and S. Pillai, "End-to-End Brain-Driven Speech Enhancement in Multi-Talker Conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1718-1733, 2022.

[4] C. Fan, J. Tao, B. Liu, J. Yi, Z. Wen and X. Liu, "End-to-End Post-Filter for Speech Separation With Deep Attention Fusion Features," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1303-1314, 2020.

[5] N. Ding and J.Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 9, pp. 11854–11859, 2012.

[6] N. Mesgarani and E. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233-236, 2012.

[7] G. Ciccarelli, M. Nolan, J. Perricone, and et al,"Comparison of Two-Talker Attention Decoding from EEG with Nonlinear Neural Networks and Linear Methods," *Scientific Reports*, vol. 9, pp. 11538, 2019.

[8] S. Cai, E. Su, L. Xie and H. Li,"EEG-Based Auditory Attention Detection via Frequency and Channel Neural Attention," in *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 2, pp. 256-266, 2022.

[9] C. Puffay, B. Accou, L. Bollens, and et al,"Relating EEG to continuous speech using deep neural networks: a review," *Journal of Neural Engineering*, vol. 20, no. 4, pp. 041003, 2023.

[10] J.A. O'Sullivan, A.J. Power, N. Mesgarani, S. Rajaram, J.J. Foxe, B.G. Cunningham, M. Slaney, S.A. Shamma and E.C. Shamma,"Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG," *Cerebral cortex*, vol. 25, no. 7, pp. 1697-1706, 2015.

[11] I. Choi, S. Rajaram, L.A. Varghese and B.G. Shinn-Cunningham,"Quantifying attentional modulation of auditory-evoked cortical responses from single-trial electroencephalography," *Frontiers in Human Neuroscience*, vol. 7, pp. 115, 2013.

[12] B. Mirkovic, S. Debener, M. Jaeger and M. De Vos,"Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications," *Journal of Neural Engineering*, vol. 12, no. 4, 2015.

[13] S. Van Eyndhoven, T. Francart, and A. Bertrand,"EEG-Informed Attended Speaker Extraction From Recorded Speech Mixtures With Application in Neuro-Steered Hearing Prostheses," *IEEE transactions on bio-medical engineering*, vol. 64, no. 5, pp. 1045-1056, 2017.

[14] A. Bednar, and E.C. Lalor,"Where is the cocktail party? Decoding locations of attended and unattended moving sound sources using EEG," *Neuroimage*, vol. 205, 2020.

[15] Y. Jiang, N. Chen, and J. Jin,"Detecting the locus of auditory attention based on the spectro-spatial-temporal analysis of EEG," *Journal of neural engineering*, vol. 19, no. 5, 2022.

[16] B. Hjorth,"EEG analysis based on time domain properties," *Electroencephalography Clinical Neurophysiology*, vol. 29, no. 3, pp. 306-310, 1970.

[17] Y. Liu, and O. Sourina, "Real-Time Fractal-Based Valence Level Recognition from EEG," *Transactions on Computational Science XVIII*, pp. 101–120, 2013.

[18] P.C. Petrantonakis, and L.J. Hadjileontiadis, "Emotion recognition from EEG using higher order crossings," *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 14, no. 2, pp. 186-197, 2010.

[19] L.I. Aftanas, N.V. Reva, A.A. Varlamov, S.V. Pavlov, and V.P. Makhnev, "Analysis of evoked EEG synchronization and desynchronization in conditions of emotional activation in humans: temporal and topographic characteristics," *Neuroscience Behavioral Physiology*, vol. 34, no. 8, pp. 859-867, 2004.

[20] R.J. Davidson, "What does the prefrontal cortex "do" in affect: perspectives on frontal EEG asymmetry research," *Biological Psychology*, vol. 67, no. 1, pp. 219-234, 2004.

[21] W.L. Zheng, and B.L. Lu,"Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks," *IEEE transactions on autonomous mental development*, vol. 7, no. 3, pp. 162-175, 2015.

[22] Y.J. Liu, M. Yu, G. Zhao, J. Song, Y. Ge, and Y. Shi, "Real-time movie-induced discrete emotion recognition from EEG signals," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 550-562, 2018.

[23] T.F. Song, W.M. Zheng, P. Song, and Z. Cui, "EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 532-541, 2020.

[24] C.A. Frantzidis, C. Bratsas, C.L. Papadelis, E. Konstantinidis, C. Pappas, and P.D. Bamidis, "Toward Emotion Aware Computing: An Integrated Approach Using Multichannel Neurophysiological Recordings and Affective Visual Stimuli," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 3, pp. 589-597, 2010.

[25] N.K. Desiraju, S. Doclo, M. Buck, and T. Wolff, "Joint Online Estimation of Early and Late Residual Echo PSD for Residual Echo Suppression," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 333-344, 2023.

[26] Y.P. Lin, C.H. Wang, T.P. Jung, T.L. Wu, S.K. Jeng, J.R. Duann, and J.H. Chen, "EEG-Based Emotion Recognition in Music Listening," *IEEE Transactions on Biomedical Engineering*, pp. 1798–1806, 2010.

[27] L.C. Shi, Y.Y. Jiao, and B.L. Lu, "Differential entropy feature for EEG-based vigilance estimation," *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6627-6630, 2013.

[28] W.L. Zheng, J.Y. Zhu, Y. Peng, and B.L. Lu, "EEG-based emotion classification using deep belief networks," *2014 IEEE International Conference on Multimedia and Expo (ICME)*, 2014.

[29] T. Taillez, B. Kollmeier, and B.T. Meyer, "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech," *European Journal of Neuroscience*, vol. 51, no. 5, pp. 1234-1241, 2017.

[30] S. Cai, P. Li, E. Su, and L. Xie, "Auditory Attention Detection via Cross-Modal Attention," *Frontiers in neuroscience*, vol. 51, 2021.

[31] S. Cai, P. Sun, T. Schultz, and H. Li, "Low-latency auditory spatial attention detection based on spectro-spatial features from EEG," *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 5812-5815, 2021.

[32] H. Fastl, and E. Zwicker, "Psychoacoustics: facts and models," *Choice Reviews Online*, vol. 28, no. 10, 2013.

[33] S. Geirnaert, S. Vandecappelle, E. Alickovic, and et al, "Electroencephalography-Based Auditory Attention Decoding: Toward Neurosteered Hearing Devices," *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 89-102, 2021.

[34] S. Vandecappelle, L. Deckers, N. Das, A.H. Ansari, A. Bertrand, and T. Francart, "EEG-based detection of the locus of auditory attention with convolutional neural networks," *Elife*, vol. 10, 2021.

[35] Z. Jie, C. Ganqu, H. Shengding, Z. Zhengyan, Y. Cheng, and et al, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57-81, 2020.

[36] A.A. Nurul, S. Yeahia, K.C. Ripon, J.R. Michael, H.A. Md, and et al, "Graph Neural Network: A Comprehensive Review on Non-Euclidean Space," *IEEE Access*, vol. 9, pp. 60588-60606, 2021.

[37] S. Cai, T. Schultz, and H. Li, "Brain Topology Modeling With EEG-Graphs for Auditory Spatial Attention Detection," *IEEE Transactions on Biomedical Engineering*, pp. 1-11, 2023.

[38] I. Kuruvila, J. Muncke, E. Fischer, and U. Hoppe, "Extracting the Auditory Attention in a Dual-Speaker Scenario From EEG Using a Joint CNN-LSTM Model," *Frontiers in Physiology*, vol. 12, 2021.

[39] M. Monesi, B. Accou, T. Francart, and H. hamme, "Extracting Different Levels of Speech Information from EEG Using an LSTM-Based Model," *arXiv: Audio and Speech Processing*, 2021.

[40] Y. Lu, M. Wang, L. Yao, H. Shen, W. Wu, Q. Zhang, and et al, "Auditory attention decoding from electroencephalography based on long short-term memory networks," *Biomedical Signal Processing and Control*, vol. 70, 2021.

[41] F.P. Such, S. Sah, M.A. Dominguez, S. Pillai, C. Zhang, A. Michael, N.D. Cahill, and R. Ptucha, "Robust Spatial Filtering With Graph Convolutional Neural Networks," *IEEE Journal of Selected Topics in Signal Processing*, pp. 884-896, 2017.

[42] L. Zhao, Y.J. Song, C. Zhanf, Y. Liu, P. Wang, T. Lin, M. Deng, and H.F. LI, "T-GCN: A Temporal Graph Convolutional Network for Traffic Prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 9, pp. 3848-3858, 2020.

[43] L. Jacob, "Review of Spectral Graph Theory," *ACM SIGACT News*, vol. 30, no. 2, 1999.

[44] W.F. Liu, M.G. Gong, Z.D. Tang, A.K. Qin, K. Sheng, and M.L. Xu, "Locality preserving dense graph convolutional networks with graph context-aware node representation," *Neural Networks*, vol. 143, pp. 108-120, 2021.

[45] Q. Xu, T. Song, L. Wang, H. Shi, Y. Lin, Y. Lv, M. Ge, Q. Yu, and J. Dang, "Self-Distillation Based on High-level Information Supervision for Compressing End-to-End ASR Model," *Interspeech*, pp. 1716-1720, 2022.

[46] B. Liu, H. Wang, Z. Chen, S. Wang, and Y. Qian, "Self-Knowledge Distillation via Feature Enhancement for Speaker Verification," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7542-7546, 2022.

[47] J. Xue, C. Fan, J. Yi, C. Wang, Z. Wen, D. Zhang, and Z. Lv, "Learning From Yourself: A Self-Distillation Method for Fake Speech Detection," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5, 2023.

[48] D.I. Shuman, S.K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, pp. 83-98, 2013.

[49] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering," *Curran Associates Inc.*, pp. 3844–3852, 2016.

[50] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, and et al, "Convolutional Networks on Graphs for Learning Molecular Fingerprints," in *Advances in Neural Information Processing Systems (NIPS2015)*, pp. 2224-2232, 2015.

[51] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral Networks and Locally Connected Networks on Graphs," in *International Conference on Learning Representations (ICLR2014), CBLS, April 2014*, 2014.

[52] P. Yang, G. Yang, X. Gong, P. Wu, X. Han, J. Wu, and C. Chen, "Instance Segmentation Network With Self-Distillation for Scene Text Detection," *IEEE Access*, vol. 8, pp. 45825-45836, 2020.

[53] T.B. Xu, and C.L. Liu, "Deep Neural Network Self-Distillation Exploiting Data Representation Invariance," *IEEE Transactions on Neural Networks and Learning System*, vol. 33, no. 1, pp. 257-269, 2022.

[54] F.R. Chung, "Spectral Graph Theory," *Rhode, Island: American Mathematical Soc*, 1997.

[55] R.N. Duan, J.Y. Zhu, and B.L. Lu, "Differential entropy feature for EEG-based emotion classification," *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 81-84, 2013.

[56] N. Das, W. Biesmans, A. Bertrand, and T. Francart, "The effect of head-related filtering and ear-specific decoding bias on auditory attention detection," *Journal of Neural Engineering*, vol. 13, no. 5, 2016.

[57] N. Das, T. Francart, and A. Bertrand, "Auditory Attention Detection Dataset KULeuven," *Zenodo*, 2020, DOI: 10.5281/zenodo.3997352.

[58] S.A. Fuglsang, T. Dau, and J. Hjortkjær, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *NeuroImage*, vol. 156, pp. 435-444, 2017.

[59] S.A. Fuglsang, D. Wong, and J. Hjortkjær, "EEG and audio dataset for auditory attention decoding," *Zenodo*, 2018, DOI: 10.5281/zenodo.1199011.

[60] A. de Cheveigné, D.D. Wong, G.M. Di Liberto, J. Hjortkjaer, m. Slaney, and E. Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206-216, 2018.

[61] P. Li, S. Cai, E. Su, and L. Xie, "A Biologically Inspired Attention Network for EEG-Based Auditory Attention Detection," *IEEE Signal Processing Letters*, vol. 29, pp. 284-288, 2022.

[62] S. Cai, P. Li, E. Su, Q. Liu, and L. Xie, "A Neural-Inspired Architecture for EEG-Based Auditory Attention Detection," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 4, pp. 668-676, 2022.

[63] R. Zink, A.G. Baptist, A. Bertrand, S.V. Huffel, and M.D. Vos, "Online detection of auditory attention in a neurofeedback application," *Proc. 8th International Workshop on Biosignal Interpretation*, pp. 1-4, 2016.

Jianhua Tao (Senior Member, IEEE) received the M.S. degree from Nanjing University, Nanjing, China, in 1996, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2001. He is currently a Professor with Department of Automation, Tsinghua University, Beijing, China. He has authored or coauthored more than 300 papers on major journals and proceedings including the IEEE TASLP, IEEE TAFFC, IEEE TIP, IEEE TSMCB, Information Fusion, etc. His current research interests include speech recognition and synthesis, affective computing, and pattern recognition. He is the Board Member of ISCA, the chairperson of ISCA SIG-CSLP, the Chair or Program Committee Member for several major conferences, including Interspeech, ICPR, ACII, ICMI, ISCSLP, etc. He was the subject editor for the Speech Communication, and is an Associate Editor for Journal on Multimodal User Interface and International Journal on Synthetic Emotions. He was the recipient of several awards from the important conferences, including Interspeech, NCMMSC, etc.

Jiangyan Yi received the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2018, and the M.A. degree fromthe Graduate School of Chinese Academy of Social Sciences, Beijing, China, in 2010. She was a Senior R&D Engineer with Alibaba Group during 2011 to 2014. She is currently an Assistant Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. Her current research interests include speech processing, speech recognition, distributed computing, deep learning, and transfer learning.

Cunhang Fan received the Ph.D degree with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2021, and the B.S. degree from the Beijing University of Chemical Technology (BUCT), Beijing, China, in 2016. He is currently a associate professor with the School of Computer Science and Technology, Anhui University, Heifei, China. His current research interests include auditory attention detection, speech enhancement, speech recognition and speech processing.

Zhao Lv received his Ph.D. degree in Computer Application Technology from Anhui University, Hefei, China, in 2011. He was a visiting scholar with the University of Utah, Salt Lake City, USA, from 2017 to 2018. He is currently a professor in the School of Computer Science and Technology at Anhui University, Hefei, China. His research interests include intelligent information processing and pattern recognition regarding biomedical signal (EEG, EOG, etc.) as well as speech signal processing.

Hongyu Zhang graduated from Jinan University in 2022 with a Bachelor's degree in Computer Science and Technology. She is currently pursuing a Master's degree at the School of Computer Science and Technology, Anhui University. Her current research interests include auditory attention detection and brain-computer interfaces.

Xiaopei Wu is a professor at the School of Computer Science and Technology, Anhui University, and a doctoral/master's supervisor. He received his bachelor's, master's and Doctor's degrees from Anhui University, University of Electronic Science and Technology of China and University of Science and Technology of China in 1985, 1988 and 2002, respectively. 2003-2006 Postdoctoral research at the Signal and Information Processing Postdoctoral Mobile Station of the University of Science and Technology of China, 2004.4-2004.10 Study at the University of California, San Diego. Main research areas: Machine learning and brain-computer interface; Voice and intelligent video analysis; Human biological signal monitoring and special human-computer interaction technology.

Jun Xue received the B.S. degree from the Anhui Science And Technology University, in 2020, and the M.S. degree at Anhui University from 2021 to the present. His research interests include: Fake speech detection, Knowledge distillation and Self-supervised learning.