

EEG-Based Emotion Recognition Using Regularized Graph Neural Networks

Peixiang Zhong, Di Wang, *Member, IEEE*, and Chunyan Miao, *Senior Member, IEEE*

Abstract—EEG signals measure the neuronal activities on different brain regions via electrodes. Many existing studies on EEG-based emotion recognition do not exploit the topological structure of EEG signals. In this paper, we propose a regularized graph neural network (RGNN) for EEG-based emotion recognition, which is biologically supported and captures both local and global inter-channel relations. Specifically, we model the inter-channel relations in EEG signals via an adjacency matrix in our graph neural network where the connection and sparseness of the adjacency matrix are supported by the neuroscience theories of human brain organization. In addition, we propose two regularizers, namely node-wise domain adversarial training (NodeDAT) and emotion-aware distribution learning (EmotionDL), to improve the robustness of our model against cross-subject EEG variations and noisy labels, respectively. To thoroughly evaluate our model, we conduct extensive experiments in both subject-dependent and subject-independent classification settings on two public datasets: SEED and SEED-IV. Our model obtains better performance than competitive baselines such as SVM, DBN, DGCNN, BiDANN, and the state-of-the-art BiHDM in most experimental settings. Our model analysis demonstrates that the proposed biologically supported adjacency matrix and two regularizers contribute consistent and significant gain to the performance. Investigations on the neuronal activities reveal that pre-frontal, parietal and occipital regions may be the most informative regions for emotion recognition, which is consistent with relevant prior studies. In addition, experimental results suggest that global inter-channel relations between the left and right hemispheres are important for emotion recognition and local inter-channel relations between (FP1, AF3), (F6, F8) and (FP2, AF4) may also provide useful information.

Index Terms—Affective Computing, EEG, Graph Neural Network, SEED

1 INTRODUCTION

EMOTION recognition is an important subarea of affective computing, which focuses on recognizing human emotions based on a variety of modalities, such as audio-visual expressions, body language, physiological signals, etc. Compared to other modalities, physiological signals, such as electroencephalogram (EEG), electrocardiogram (ECG), electromyogram (EMG), galvanic skin response (GSR), etc., have the advantage of being difficult to hide or disguise. In recent years, due to the rapid development of noninvasive, easy-to-use and inexpensive EEG recording devices, EEG-based emotion recognition has received an increasing amount of attention in both research [1] and applications [2].

Emotion models can be broadly categorized into discrete models and dimensional models. The former categorizes emotions into discrete entities, e.g., anger, disgust, fear, happiness, sadness, and surprise in Ekman's theory [3]. The latter describes emotions using their underlying dimensions, e.g., valence, arousal and dominance [4], which measures emotions from unpleasant to pleasant, passive to active, and submissive to dominant, respectively.

EEG signals measure voltage fluctuations from the cortex in the brain and have been shown to reveal important information about human emotional states [5]. For example, greater relative left frontal EEG activity has been observed

when experiencing positive emotions [5]. The voltage fluctuations on different brain regions are measured by electrodes attached to the scalp. Each electrode collects EEG signals in one channel. The collected EEG signals are often analyzed in specific frequency bands for each channel, namely delta (1-4 Hz), theta (4-7 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma (>30 Hz).

Many existing EEG-based emotion recognition methods are primarily based on the supervised machine learning approach wherein features are extracted from preprocessed EEG signals in each channel over a time window and then a classifier is trained on the extracted features to recognize emotions. Wang *et al.* [6] compared power spectral density features (PSD), wavelet features and nonlinear dynamical features with a Support Vector Machine (SVM) classifier. Zheng and Lu [7] investigated critical frequency bands and channels using PSD, differential entropy (DE) [8] and PSD asymmetry features, and obtained robust accuracy using deep belief networks (DBN). However, most existing EEG-based emotion recognition approaches do not address the following three challenges: 1) the topological structure of EEG signals are not effectively exploited to learn more discriminative EEG representations; 2) EEG signals vary significantly across different subjects, which hinders the generalizability of the trained classifiers; and 3) participants may not always generate the intended emotions when watching emotion-eliciting stimuli. Consequently, the emotion labels in the collected EEG data are noisy and may not be consistent with the actual elicited emotions.

There have been several attempts to address the first challenge. Zhang *et al.* [9] and Zhang *et al.* [10] incorporated spatial relations in EEG signals using convolutional

- P. Zhong, D. Wang and C. Miao are with Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), Nanyang Technological University, Singapore.
P. Zhong and C. Miao are with School of Computer Science and Engineering, Nanyang Technological University, Singapore.
E-mail: peixiang001@e.ntu.edu.sg, {wangdi, ascymiao}@ntu.edu.sg

neural networks (CNN) and recurrent neural networks (RNN), respectively. However, their approaches require a 2D representation of EEG channels on the scalp, which may cause information loss during flattening because channels are actually arranged in the 3D space. In addition, their approach of using CNNs and RNNs to capture inter-channel relations has difficulty in learning long-range dependencies [11]. Graph neural networks (GNN) has been applied in [12] to capture inter-channel relations using an adjacency matrix. However, similar to CNNs and RNNs, their approach only considers relations between the nearest channels, which thus may lose valuable information between distant channels, such as PSD asymmetry between channels on the left and right hemispheres in the frontal region, which has been shown as informative in valence prediction [5]. A recent work applies RNNs to learn EEG representations in the two hemispheres separately and then adopts the asymmetric differences between them to recognize emotions [13]. However, their approach is limited to using only the bi-hemispherical discrepancies and ignores other useful features such as neuronal activities recorded from each channel.

In recent years, several studies [14], [15] investigated the transferability of EEG-based emotion recognition models across subjects. Lan *et al.* [16] compared several domain adaptation techniques such as maximum independence domain adaptation (MIDA), transfer component analysis (TCA), subspace alignment (SA), etc. They found that the subject-independent classification accuracy can be improved by around 10%. Li *et al.* [17] applied domain adversarial learning to lower the influence of individual subject on EEG data and obtained improved performance as well. However, their approaches do not exploit any graph structure and only leads to small performance improvement (see Section 7.1).

To the best of our knowledge, no attempt has been made to address the problem of noisy labels in EEG-based emotion recognition.

In this paper, we propose a regularized graph neural network (RGNN) aiming to address all three aforementioned challenges. Graph analysis for human brain has been studied extensively in the neuroscience literature [18], [19]. However, making an accurate connectome is still an open question and subject to different scales [19]. Inspired by [12], [20], we consider each channel in EEG signals as a node in our graph. Our RGNN model extends the simple graph convolution network (SGC) [21] and leverages the topological structure of EEG signals, i.e., according to the economy of brain network organization [20], we propose a biologically supported sparse adjacency matrix to capture both local and global inter-channel relations. Local inter-channel relations connect nearby groups of neurons and may reveal anatomical connectivity at macroscale [19], [22]. Global inter-channel relations connect distant groups of neurons between the left and right hemispheres and may reveal emotion-related functional connectivity [5], [17].

In addition, we propose a node-wise domain adversarial training (NodeDAT) to regularize our graph model for better generalization in subject-independent classification scenarios. Different from the domain adversarial training adopted by [17], [23], our NodeDAT gives a finer-grained regularization by minimizing the domain discrepancies be-

tween features in the source and target domains for each channel/node. Moreover, we propose an emotion-aware distribution learning (EmotionDL) method to address the problem of noisy labels in the datasets. Prior studies have shown that noisy labels can adversely impact classification accuracy [24]. Instead of learning single-label classification, our EmotionDL learns a distribution of labels of the training data and thus acts as a regularizer to improve the robustness of our model against noisy labels. Finally, we conduct extensive experiments to validate the effectiveness of our proposed model and investigate emotion-related informative neuronal activities.

In summary, the main contributions of this paper are as follows:

- 1) We propose a regularized graph neural network (RGNN) model to recognize emotions based on EEG signals. Our model is biologically supported and captures both local and global inter-channel relations.
- 2) We propose two regularizers: a node-wise domain adversarial training (NodeDAT) and an emotion-aware distribution learning (EmotionDL), which aim to improve the robustness of our model against cross-subject variations and noisy labels, respectively.
- 3) We conduct extensive experiment in both subject-dependent and subject-independent classification settings on two public EEG datasets, namely SEED [7] and SEED-IV [25]. Experimental results demonstrate the effectiveness of our proposed model and regularizers. In addition, our RGNN achieves superior performance over the state-of-the-art baselines in most experimental settings.
- 4) We investigate the neuronal activities and the results reveal that pre-frontal, parietal and occipital regions may be the most informative regions for emotion recognition. In addition, global inter-channel relations between the left and right hemispheres are important and local inter-channel relations between (FP1, AF3), (F6, F8) and (FP2, AF4) may also provide useful information.

2 RELATED WORK

In this section, we review related work in the fields of EEG-based emotion recognition, graph neural networks, unsupervised domain adaptation and learning with noisy labels.

2.1 EEG-Based Emotion Recognition

EEG feature extractors and classifiers are the two fundamental components in the machine learning approach of EEG-based emotion recognition. EEG features can be broadly divided into single-channel features and multi-channel ones [26]. The majority of existing features are single-channel features such as statistical features [27], [28], fractal dimension (FD) [29], PSD [30], differential entropy (DE) [8], and wavelet features [31]. A few features are computed on multiple channels to capture the inter-channel relations, e.g.,

the asymmetry features of PSD [7] and functional connectivity [32], [33], where common indices such as correlation, coherence and phase synchronization were used estimate brain functional connectivity between channels. However, leveraging functional connectivity require labor-intensive manual connectivity analysis for each subject and may not be ideal for real-time applications.

EEG classifiers can be broadly divided into topology-invariant classifiers and topology-aware ones. The majority of existing classifiers are topology-invariant classifiers such as SVM, k-Nearest Neighbors (KNN), DBNs [34] and RNNs [35], which do not take the topological structure of EEG features into account when learning the EEG representations. In contrast, topology-aware classifiers such as CNNs [9], [36], [37], [38] and GNNs [12] consider the inter-channel topological relations and learn EEG representations for each channel by aggregating features from nearby channels using convolutional operations either in the Euclidean space or in the non-Euclidean space. However, as discussed in Section 1, existing CNNs and GNNs have difficulty in learning the dependencies between distant channels, which may reveal important emotion-related information. Recently, Zhang *et al.* [10] and Li *et al.* [13] proposed to use RNNs to learn spatial topological relations between channels by scanning electrodes in both vertical and horizontal directions. However, their approaches do not fully exploit the topological structure of EEG channels. For example, two topologically close channels may be far away from each other in the scanning sequence.

2.2 Graph Neural Networks

Graph neural networks (GNN) is a class of neural networks dealing with data in the graph domains, e.g., molecular structures, social networks and knowledge graphs [39]. One early work on GNNs [40] aimed to learn a converged static state embedding for each node in the graph using a transition function applied to its neighborhood. Later, inspired by the convolutional operation of CNN in Euclidean domains, Bruna *et al.* [41] combined spectral graph theory [42] with neural networks and defined convolutional operations in graph domains using the spectral filters computed from the normalized graph Laplacian. Following this line of research, Defferrard *et al.* [43] proposed fast localized convolutions by using a recursive formulation of the K -order Chebyshev polynomials to approximate the filters. The resulting representation for each node is an aggregation of its K^{th} -order neighborhood. Kipf and Welling [44] further limited $K = 1$ and proposed the standard graph convolutional network (GCN) with a faster localized graph convolutional operation. The convolutional layers in GCN can be stacked K times to effectively convolve the K^{th} -order neighborhood of a node. Recently, Wu *et al.* [21] simplified GCN by removing the nonlinearities between convolutional layers in GCN and proposed the simple graph convolution network (SGC), which effectively behaves like a linear feature transformation followed by a logistic regression. SGC performs orders of magnitude faster than GCNs with comparable classification accuracy. In this paper, we extend SGC to model EEG signals and propose a biologically supported adjacency matrix and two regularizers for robust EEG-based emotion recognition.

Apart from the convolution operation used in GCNs, there are other types of operations in GNNs, such as attention [45] or RNN [46]. However, they are often trained significantly slower than SGC [21].

2.3 Unsupervised Domain Adaptation

Unsupervised domain adaptation aims to mitigate the domain shift in knowledge transfer from a supervised source domain to an unsupervised target domain. The most common approaches are instance re-weighting, domain-invariant feature learning, domain mapping and normalization statistics. Instance re-weighting methods [47] aim to infer the resampling weight directly by feature distribution matching across source and target domains in a non-parametric manner. Domain-invariant feature learning methods align features from both source and target domains to a common feature space. The alignment can be achieved by minimizing divergence [48], maximizing reconstruction [49] or adversarial training [23]. The domain mapping technique is typically applied in the computer vision field where pixel-level image-to-image translation from one domain to another domain improves domain adaptation performance [50]. Normalization statistics are based on the assumption that the batch norm statistics learn domain knowledge. Caruacci *et al.* [51] performed domain adaptation by modulating the batch norm layers' statistics from source to target domain. Our proposed NodeDAT regularizer extends the domain adversarial training [23] to graph neural networks and achieves finer-grained regularization by minimizing the discrepancies between features in source and target domains for each channel/node individually.

2.4 Learning with Noisy Labels

Commonly adopted approaches to learning with noisy labels are based on the noise transition matrix and robust loss functions. The noise transition matrix specifies the probabilities of transition from each ground true label to each noisy label and is often applied to modify the cross-entropy loss. The matrix can be pre-computed as *a priori* [52] or estimated from noisy data [53]. A few studies tackle noisy labels by using noise-tolerant robust loss functions, such as unhinged loss [54] and ramp loss [55]. Several other approaches include bootstrap that leverages predicted labels to generate training targets [56] and alternatively updating network parameters and labels during training [57]. Our proposed EmotionDL regularizer is inspired by [58], which applies distribution learning to learn labels with ambiguity in the computer vision domain.

3 PRELIMINARIES

In this section, we introduce the preliminaries of the simple graph convolution network (SGC) [21] and its spectral analysis, which is the basis of our RGNN model.

3.1 Simple Graph Convolution Network (SGC)

Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where \mathcal{V} denotes a set of nodes and \mathcal{E} denotes a set of edges between nodes in \mathcal{V} . Data on \mathcal{V} can be represented by a feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, where

n denotes the number of nodes and d denotes the input feature dimension. The edge set \mathcal{E} can be represented by a weighted adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ with self-loops, i.e., $\mathbf{A}_{ii} = 1, i = 1, 2, \dots, n$. In general, GNNs learn a feature transformation function for \mathbf{X} and produces output $\mathbf{Z} \in \mathbb{R}^{n \times d'}$, where d' denotes the output feature dimension.

Between adjacent layers in GNNs, the feature transformation can be written as

$$\mathbf{H}^{l+1} = f(\mathbf{H}^l, \mathbf{A}), \quad (1)$$

where $l = 0, 1, \dots, L-1$, L denotes the number of layers, $\mathbf{H}^0 = \mathbf{X}$, $\mathbf{H}^L = \mathbf{Z}$, and f denotes the function we want to learn. A simple definition of f would be

$$f(\mathbf{H}^{l+1}) = \sigma(\mathbf{A}\mathbf{H}^l\mathbf{W}^l), \quad (2)$$

where σ denotes a non-linear function and \mathbf{W}^l denotes a weight matrix at layer l . For each node \mathbf{x} , function f simply sums up all node features in its neighborhood including \mathbf{x} itself, followed by a non-linear transformation. However, one major limitation of f in (2) is that repeatedly applying f along multiple layers may lead to \mathbf{H}^l with overly large values due to summation. Kipf and Welling [44] alleviated this limitation by proposing the graph convolution network (GCN) as follows:

$$f(\mathbf{H}^{l+1}) = \sigma(\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{\frac{1}{2}}\mathbf{H}^l\mathbf{W}^l), \quad (3)$$

where \mathbf{D} denotes the diagonal degree matrix of \mathbf{A} , i.e., $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$. The normalized adjacency matrix $\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{\frac{1}{2}}$ prevents \mathbf{H} from growing overly large. If we ignore σ and \mathbf{W}^l temporarily and expand (3), the hidden state \mathbf{H}_i^{l+1} for node $\mathbf{x}_i, i = 1, 2, \dots, n$, can be computed via

$$\mathbf{H}_i^{l+1} \leftarrow \frac{1}{\mathbf{D}_{ii} + 1} \mathbf{H}_i^l + \sum_{j=1}^n \frac{\mathbf{A}_{ij}}{\sqrt{(\mathbf{D}_{ii} + 1)(\mathbf{D}_{jj} + 1)}} \mathbf{H}_j^l. \quad (4)$$

Note that each neighboring \mathbf{H}_j^l is now normalized by both the degrees of \mathbf{x}_i and \mathbf{x}_j . Therefore, essentially, for each node, the feature transformation function f in GCN is a non-linear transformation of the weighted sum of node features of itself and its neighborhood. Successively applying L graph convolutional layers aggregates node features within a neighborhood of size L .

To further accelerate training while keeping comparable performance, Wu *et al.* [21] proposed SGC by removing the non-linear function σ in (3) and reparameterizing all linear transformations \mathbf{W}^l across all layers into one linear transformation \mathbf{W} as follows:

$$\mathbf{Z} = \mathbf{H}^L = \mathbf{S}\mathbf{H}^{L-1}\mathbf{W}^{L-1} = \dots = \mathbf{S}^L\mathbf{X}\mathbf{W}, \quad (5)$$

where $\mathbf{S} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{\frac{1}{2}}$, and $\mathbf{W} = \mathbf{W}^{L-1}\mathbf{W}^{L-2}\dots\mathbf{W}^0$. Essentially, SGC computes a topology-aware linear transformation $\hat{\mathbf{X}} = \mathbf{S}^L\mathbf{X}$, followed by one final linear transformation $\mathbf{Z} = \hat{\mathbf{X}}\mathbf{W}$.

3.2 Spectral Graph Convolution

We analyze GCN from the perspective of spectral graph theory [42]. Graph Fourier analysis relies on the graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{A}$ or the normalized graph Laplacian $\hat{\mathbf{L}} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{\frac{1}{2}}$. Since $\hat{\mathbf{L}}$ is a symmetric positive

semidefinite matrix, it can be decomposed as $\hat{\mathbf{L}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where \mathbf{U} is the orthonormal eigenvector matrix of $\hat{\mathbf{L}}$ and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_N)$ is the diagonal matrix of corresponding eigenvalues. Given graph data \mathbf{X} , the graph Fourier transform of \mathbf{X} is $\hat{\mathbf{X}} = \mathbf{U}^T\mathbf{X}$, and the inverse Fourier transform of $\hat{\mathbf{X}}$ is $\mathbf{X} = \mathbf{U}\hat{\mathbf{X}}$. Hence, the graph convolution between \mathbf{X} and a filter \mathbf{G} is computed as follows:

$$\mathbf{X} * \mathbf{G} = \mathbf{U}((\mathbf{U}^T\mathbf{G}) \odot (\mathbf{U}^T\mathbf{X})) = \mathbf{U}\hat{\mathbf{G}}\mathbf{U}^T\mathbf{X}, \quad (6)$$

where \odot denotes element-wise multiplication, and $\hat{\mathbf{G}} = \text{diag}(\hat{g}_1, \dots, \hat{g}_N)$ denotes a diagonal matrix with n spectral filter coefficients.

To reduce the current learning complexity of $\mathcal{O}(n)$ to that of conventional CNN, i.e., $\mathcal{O}(K)$, (6) can be approximated using the K th order polynomials as follows:

$$\mathbf{U}\hat{\mathbf{G}}\mathbf{U}^T\mathbf{X} \approx \mathbf{U}(\sum_{i=0}^K \theta_i \mathbf{\Lambda}^i) \mathbf{U}^T\mathbf{X} = \sum_{i=0}^K \theta_i \hat{\mathbf{L}}^i \mathbf{X}, \quad (7)$$

where θ_i denotes coefficients. To further reduce computational cost, Defferrard *et al.* [43] proposed to use Chebyshev polynomials to approximate the filtering operation as follows:

$$\mathbf{U}\hat{\mathbf{G}}\mathbf{U}^T\mathbf{X} = \sum_{i=0}^K \theta_i T_i(\hat{\mathbf{L}}') \mathbf{X}, \quad (8)$$

where θ_i denotes learnable parameters, $\hat{\mathbf{L}}'$ denotes the scaled normalized Laplacian $\hat{\mathbf{L}}' = \frac{2}{\lambda_{max}}\hat{\mathbf{L}} - \mathbf{I}$ with its eigenvalues lying within $[-1, 1]$, and $T_i(x)$ denotes the Chebyshev polynomials recursively defined as $T_i(x) = 2xT_{i-1}(x) - T_{i-2}(x)$ with $T_0(x) = 1$ and $T_1(x) = x$.

The GCN proposed in [44] made a few approximations to simplify the filtering operation in (8): 1) use $K = 1$; 2) set $\lambda_{max} = 2$; and 3) set $\theta_0 = -\theta_1$. The resulted GCN arrives at (3). Essentially, the graph convolutional operations defined in (3) and (5) behave like a low-pass filter by smoothing the features of each node on the graph using node features in its neighborhood.

4 REGULARIZED GRAPH NEURAL NETWORK

In this section we present our regularized graph neural network (RGNN), specifically, the biologically supported adjacency matrix, and RGNN with two regularizers, i.e., node-wise domain adversarial training (NodeDAT) and emotion-aware distribution learning (EmotionDL).

4.1 Adjacency Matrix in RGNN

The adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ in RGNN represents the topological structure of EEG channels, where n denotes the number of channels in EEG signals or nodes on the graph. Each entry \mathbf{A}_{ij} in the adjacency matrix indicates the weight of connection between channels i and j . Note that \mathbf{A} contains self-loops. To reduce overfitting, we model \mathbf{A} as a symmetric matrix by using only $\frac{n(n+1)}{2}$ number of parameters instead of n^2 . Salvador *et al.* [59] observed that the strength of connection between brain regions decays as an inverse square or gravity-law function of physical distance.

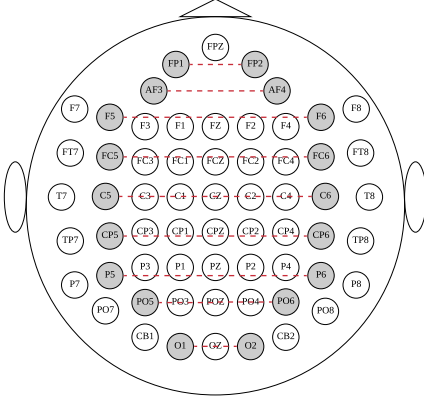


Fig. 1: The 62-channel EEG placement used to collect data in SEED and SEED-IV. Gray symmetric channels are connected globally via red dashed lines.

Hence, we initialize the local inter-channel relations in our adjacency matrix as follows:

$$\mathbf{A}_{ij} = \exp\left(-\frac{d_{ij}}{2\delta^2}\right), \quad (9)$$

where d_{ij} , $i, j = 1, 2, \dots, n$, denotes the physical distance between channels i and j , computed from the data sheet of the recording device, and δ denotes a sparsity hyper-parameter controlling the decay rate of the connection between channels.

Bullmore and Sporns [20] proposed that the brain organization is shaped by an economic trade-off between minimizing wiring costs and network running costs. Minimizing wiring costs encourages local inter-channel connections as modelled in (9). However, minimizing network running costs encourages certain global inter-channel connections for high efficiency of information transfer across the network as a whole. To this end, we add several global connections to our adjacency matrix. The global connections are subject to the specific EEG channel placement adopted in experiments. Fig. 1 depicts the global connections in both SEED [7] and SEED-IV [25]. The selection of global channels is supported by prior studies showing that the asymmetry in neuronal activities between the left and right hemispheres is informative in valence and arousal predictions [5], [60], [61]. To leverage the differential asymmetry information, we initialize the global inter-channel relations in \mathbf{A} to $[-1, 0]$ as follows:

$$\mathbf{A}_{ij} = \mathbf{A}_{ij} - 1, \quad (10)$$

where (i, j) denotes the indices of empirically selected symmetric channel pairs that balance wiring cost and global efficiency [20]: (FP1, FP2), (AF3, AF4), (F5, F6), (FC5, FC6), (C5, C6), (CP5, CP6), (P5, P6), (PO5, PO6), and (O1, O2). Note that our adjacency matrix \mathbf{A} obtained in (10) aims to represent the brain network which combines both local anatomical connectivity and emotion-related global functional connectivity.

The last step in constructing the adjacency matrix is finding an optimal value of δ to regularize the weights of connections between local channels. Achard and Bullmore [62] observed that sparse fMRI networks, comprising

around 20% of all possible connections, typically maximize the efficiency of the network topology. Thus, we choose δ such that around 20% of entries in \mathbf{A} are larger than 0.1 in absolute values. We empirically pick 0.1 as the threshold of having negligible connections between channels.

4.2 Dynamics of RGNN

Our RGNN model extends the SGC model [21]. The architecture of RGNN is illustrated in Fig. 2. Given EEG features $\mathbf{X} \in \mathbb{R}^{N \times n \times d}$ and labels $\mathbf{Y} \in \mathbb{Z}^N$, where N denotes the number of training samples, n denotes the number of nodes or channels, d denotes the input feature dimension, $\mathbf{Y}_i \in \{0, 1, \dots, C-1\}$ denotes the label index, and C denotes the number of classes. Our model aims to minimize the following cross-entropy loss:

$$\Phi = - \sum_{i=1}^N \log(p(\mathbf{Y}_i | \mathbf{X}_i, \theta)) + \alpha \|\mathbf{A}\|_1, \quad (11)$$

where θ denotes the model parameters we want to optimize, and α denotes the L1 sparse regularization strength of our adjacency matrix \mathbf{A} .

By passing each feature matrix \mathbf{X}_i into our RGNN, the output probability of class \mathbf{Y}_i can be computed as

$$\begin{aligned} \mathbf{Z}_i &= \mathbf{S}^L \mathbf{X}_i \mathbf{W} \\ p(\mathbf{Y}_i | \mathbf{X}_i, \theta) &= \text{softmax}_i(\text{pool}(\sigma(\mathbf{Z}_i)) \mathbf{W}^O), \end{aligned} \quad (12)$$

where $\mathbf{S} \in \mathbb{R}^{n \times n}$, $\mathbf{W} \in \mathbb{R}^{d \times d'}$ and L follow the definitions in (5), $\sigma(x) = \max(0, x)$, $\mathbf{W}^O \in \mathbb{R}^{d' \times C}$ denotes the output weight matrix, and $\text{pool}(\cdot)$ denotes the sum pooling across all nodes on the graph. We choose sum pooling because it demonstrated more expressive power than mean pooling and max pooling [63]. Note that we use the absolute values of \mathbf{A} to compute the degree matrix \mathbf{D} because \mathbf{A} has negative elements, e.g., global connections.

4.2.1 Node-wise Domain Adversarial Training

EEG signals vary significantly across different subjects, which hinders the generalizability of trained classifiers. To improve subject-independent classification performance, we extend the domain adversarial training [23] by proposing a node-wise domain adversarial training (NodeDAT) to reduce the discrepancies between source and target domains, i.e., training and testing sets, respectively. Specifically, a domain classifier is proposed to classify each node representation into either source domain or target domain. Compared to [23], which only regularizes the pooled representation in the last layer, our NodeDAT has finer-grained regularization because it explicitly regularizes each node representation before pooling (see Section 7.1). During optimization, our model aims to confuse the domain classifier by learning domain-invariant representations for each node.

Specifically, given source/training data $\mathbf{X}^S \in \mathbb{R}^{N \times n \times d}$ (in this subsection, we denote \mathbf{X} by \mathbf{X}^S for better clarity) and unlabelled target/testing data $\mathbf{X}^T \in \mathbb{R}^{N \times n \times d}$, where in practice \mathbf{X}^T can be either oversampled or downsampled to have the same number of samples as \mathbf{X}^S [23], the domain

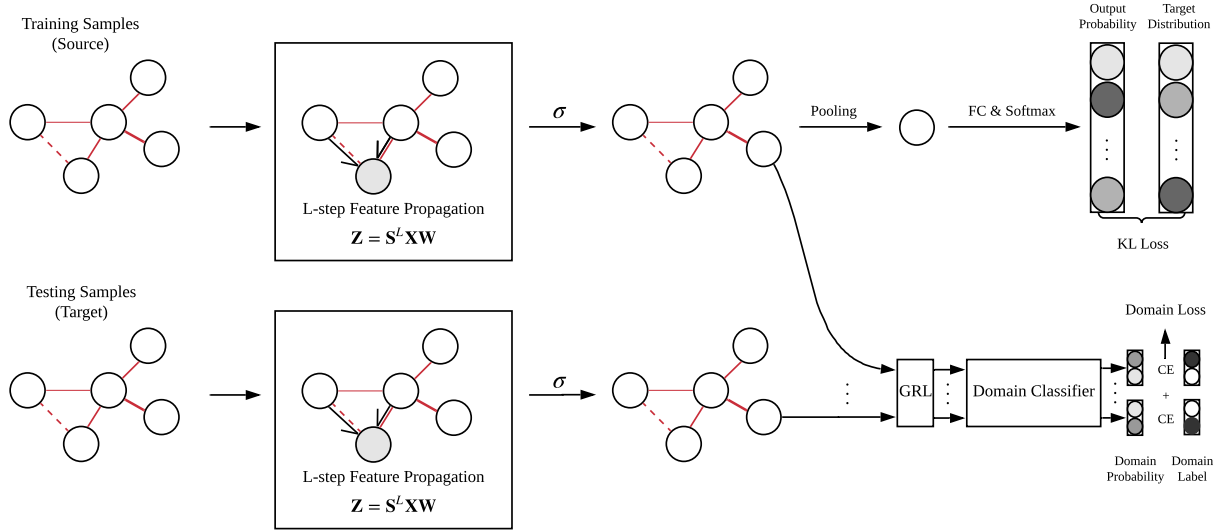


Fig. 2: The overall architecture of our RGNN model. FC denotes a fully-connected layer. CE denotes cross-entropy loss.

classifier aims to minimize the sum of the following two binary cross-entropy losses:

$$\Phi_D = - \sum_{i=1}^N \sum_{j=1}^n (\log(p_D(0|\mathbf{X}_i^S)_j) + \log(p_D(1|\mathbf{X}_i^T)_j)), \quad (13)$$

where 0 and 1 denote source and target domains, respectively. Intuitively the domain classifier aims to classify source data as 0 and target data as 1. The domain probabilities $p_D(\cdot)_j$ for node j are computed as

$$\begin{aligned} p_D(0|\mathbf{X}_i^S)_j &= \text{softmax}_0(\sigma(\mathbf{Z}_{ij}^S \mathbf{W}^D)), \\ p_D(1|\mathbf{X}_i^T)_j &= \text{softmax}_1(\sigma(\mathbf{Z}_{ij}^T \mathbf{W}^D)), \end{aligned} \quad (14)$$

where $\mathbf{Z}_{ij}^{\{S,T\}}$ denote the j th node representation in $\mathbf{Z}_i^{\{S,T\}}$, and $\mathbf{W}^D \in \mathbb{R}^{d \times 2}$ denotes the model parameters in the domain classifier. Essentially, our NodeDAT encourages learning domain invariant node presentation $\mathbf{Z}_{ij}^{\{S,T\}}$ by trying to confuse the domain classifier.

Note that our domain classifier implements a gradient reversal layer (GRL) [23] to reverse the gradients of the domain classifier during backpropagation. The gradients are further scaled by a GRL scaling factor β which gradually increases from 0 to 1 as the training progresses. The gradually increasing β allows our domain classifier to be less sensitive to noisy inputs at the early stages of the training process. Specifically, as suggested in [23], we let $\beta = \frac{2}{1+e^{-10p}} - 1$, where $p \in [0, 1]$ denotes the training progress.

4.2.2 Emotion-aware Distribution Learning

Participants may not always generate the intended emotions when watching emotion-eliciting stimuli. To address the problem of noisy emotion labels in the datasets, we propose an emotion-aware distribution learning method (EmotionDL) to learn a distribution of classes instead of one single class for each training sample. Specifically, we convert each training label $\mathbf{Y}_i \in \{0, 1, \dots, C-1\}$ into a prior probability distribution of all classes $\hat{\mathbf{Y}}_i \in \mathbb{R}^C$, where $\hat{\mathbf{Y}}_{ic}$

denotes the probability of class c in $\hat{\mathbf{Y}}_i$. The conversion is dataset-dependent. In SEED, there are three classes: negative, neutral, and positive with corresponding class indices 0, 1, and 2, respectively. We convert \mathbf{Y} as follows:

$$\hat{\mathbf{Y}}_i = \begin{cases} (1 - \frac{2\epsilon}{3}, \frac{2\epsilon}{3}, 0), & \mathbf{Y}_i = 0, \\ (\frac{\epsilon}{3}, 1 - \frac{2\epsilon}{3}, \frac{\epsilon}{3}), & \mathbf{Y}_i = 1, \\ (0, \frac{2\epsilon}{3}, 1 - \frac{2\epsilon}{3}), & \mathbf{Y}_i = 2, \end{cases} \quad (15)$$

where $\epsilon \in [0, 1]$ denotes a hyper-parameter controlling the noise level in the training labels. This conversion mechanism is based on our assumption that participants are unlikely to generate opposite emotions when watching emotion-eliciting stimuli. Therefore, the converted class distribution centers on the original class and has non-zero and zero probabilities at its nearest and opposite classes, respectively.

In SEED-IV, there are four classes: neutral, sad, fear, and happy with corresponding class indices 0, 1, 2, and 3, respectively. We can convert \mathbf{Y} as follows:

$$\hat{\mathbf{Y}}_i = \begin{cases} (1 - \frac{3\epsilon}{4}, \frac{\epsilon}{4}, \frac{\epsilon}{4}, \frac{\epsilon}{4}), & \mathbf{Y}_i = 0, \\ (\frac{\epsilon}{3}, 1 - \frac{2\epsilon}{3}, \frac{\epsilon}{3}, 0), & \mathbf{Y}_i = 1, \\ (\frac{\epsilon}{4}, \frac{\epsilon}{4}, 1 - \frac{3\epsilon}{4}, \frac{\epsilon}{4}), & \mathbf{Y}_i = 2, \\ (\frac{\epsilon}{3}, 0, \frac{\epsilon}{3}, 1 - \frac{2\epsilon}{3}), & \mathbf{Y}_i = 3. \end{cases} \quad (16)$$

The intuition behind this conversion is based on the distances between the four emotions on the valence-arousal plane. Specifically, in the self-reported ratings [25], neutral, sad, fear, and happy movie ratings cluster in the zero valence zero arousal, negative valence negative arousal, negative valence positive arousal, and positive valence positive arousal regions, respectively. Thus, we assume that participants are likely to generate emotions that have similar ratings in either valence or arousal dimensions, e.g., both angry and happy have high arousal, but unlikely to generate emotions that are far away in both dimensions, e.g., sad and happy are different in both valence and arousal.

After obtaining the converted class distributions $\hat{\mathbf{Y}}$, our model can be optimized by minimizing the following Kullback-Leibler (KL) divergence [64] instead of (11):

$$\Phi' = \sum_{i=1}^N \text{KL}(p(\mathbf{Y}|\mathbf{X}_i, \theta), \hat{\mathbf{Y}}_i) + \alpha \|\mathbf{A}\|_1, \quad (17)$$

where $p(\mathbf{Y}|\mathbf{X}_i, \theta)$ denotes the output probability distribution computed via (12). Note that our EmotionDL is different from label smoothing, which simply adds uniform noise to other classes.

4.2.3 Optimization of RGNN

Combining both NodeDAT and EmotionDL, the overall loss function Φ'' of RGNN is computed as follows:

$$\Phi'' = \Phi' + \Phi_D. \quad (18)$$

The detailed algorithm for training RGNN is presented in Algorithm 1.

Algorithm 1 The Training Algorithm for RGNN

Input: Training samples \mathbf{X} and $\hat{\mathbf{Y}}$, unlabelled testing samples \mathbf{X}^T , symmetric adjacency matrix \mathbf{A} with self-loops, learning rate η , number of epochs T , batch size B , other regularization hyper-parameters;

Output: The learned model parameters in RGNN;

- 1: Randomly initialize model parameters in RGNN;
 - 2: **for** $i = 1: T$ **do**
 - 3: **repeat**
 - 4: Draw one batch of training samples \mathbf{X}_B and $\hat{\mathbf{Y}}_B$ from \mathbf{X} and $\hat{\mathbf{Y}}$, respectively;
 - 5: Draw one batch of testing samples \mathbf{X}_B^T from \mathbf{X}^T ;
 - 6: Compute degree matrix \mathbf{D} based on (3);
 - 7: Compute normalized adjacency matrix \mathbf{S} based on (5);
 - 8: Compute output representation \mathbf{Z} based on (12);
 - 9: Use \mathbf{X}_B and $\hat{\mathbf{Y}}_B$ to compute KL loss Φ' based on (17);
 - 10: Use \mathbf{X}_B and \mathbf{X}_B^T to compute domain loss Φ_D based on (13);
 - 11: Compute GRL scaling factor β ;
 - 12: Update $\mathbf{W}^D \leftarrow \mathbf{W}^D + \eta \frac{\partial \Phi_D}{\partial \mathbf{W}^D}$;
 - 13: Update $\mathbf{W}^O \leftarrow \mathbf{W}^O + \eta \frac{\partial \Phi}{\partial \mathbf{W}^O}$;
 - 14: Update $\mathbf{W} \leftarrow \mathbf{W} + \eta (\frac{\partial \Phi}{\partial \mathbf{W}} - \beta \frac{\partial \Phi_D}{\partial \mathbf{W}})$;
 - 15: Update $\mathbf{A} \leftarrow \mathbf{A} + \eta (\frac{\partial \Phi}{\partial \mathbf{A}} - \beta \frac{\partial \Phi_D}{\partial \mathbf{A}})$;
 - 16: **until** all samples in \mathbf{X} are drawn;
-

5 EXPERIMENTAL SETTINGS

In this section, we present the datasets, classification settings and model settings in our experiments.

5.1 Datasets

We use both SEED and SEED-IV datasets in our experiments. The SEED dataset [7] comprises EEG data of 15 subjects (7 males) recorded in 62 channels using the ESI NeuroScan System¹. The EEG data was collected when

participants watch emotion-eliciting movies in three types of emotions, namely negative, neutral and positive. Each movie lasts around 4 minutes. There are three sessions of data collected and each session comprises 15 trials/movies for each subject. To make a fair comparison with existing studies, we directly use the pre-computed differential entropy (DE) features smoothed by linear dynamic systems (LDS) [7], [65] in SEED. DE extends the idea of Shannon entropy and measures the complexity of a continuous random variable. For a fixed length EEG segment, DE features are computed as the logarithm energy spectrum in a certain frequency band [8]. In SEED, DE features are pre-computed over five frequency bands (delta, theta, alpha, beta and gamma) for each second of EEG signals (without overlapping) in each channel.

The SEED-IV dataset [25] comprises EEG data of 15 subjects (7 males) recorded in 62 channels². The recording device is the same as the one used in SEED. The EEG data were collected when participants watch emotion-eliciting movies in four types of emotions, namely, neutral, sad, fear, and happy. Each movie lasts around 2 minutes. There are three sessions of data collected and each session comprises 24 trials/movies for each subject. Similar to SEED, we adopt the pre-computed DE features from SEED-IV.

5.2 Classification Settings

We conduct both subject-dependent and subject-independent classifications on both SEED and SEED-IV to evaluate our model.

5.2.1 Subject-Dependent Classification

For SEED, we follow the experimental settings in [7], [12], [17] to evaluate our RGNN model using subject-dependent classification, i.e., we evaluate our model for individual subjects. Specifically, for each subject, we train our model using the first 9 trials as the training set and the remaining 6 trials as the testing set. We evaluate the model performance by using the accuracy averaged across all subjects over two sessions of EEG data in SEED [7]. For SEED-IV, we follow the experimental settings in [13], [25] to evaluate our RGNN model using subject-dependent classification. Specifically, for each subject, the first 16 trials are used for training and the remaining 8 trials containing all emotions (each emotion with two trials) are used for testing. We evaluate our model using data from all three sessions [25].

5.2.2 Subject-Independent Classification

For SEED, we follow the experimental settings in [12], [14], [17] to evaluate our RGNN model using subject-independent classification. Specifically, we adopt leave-one-subject-out cross-validation, i.e, during each fold, we train our model on 14 subjects and test on the remaining subject. We evaluate the model performance using the accuracy averaged across all test subjects over one session of EEG data in SEED [14]. For SEED-IV, we follow the experimental settings in [13] to evaluate our RGNN model using subject-independent classification. We evaluate our model using data from all three sessions.

2. SEED-IV also contains eye movement data, which we do not use in our experiment.

1. <https://compumedicsneuroscan.com/>

TABLE 1: Subject-dependent classification accuracy (mean/standard deviation) on SEED and SEED-IV

Model	SEED						SEED-IV
	delta band	theta band	alpha band	beta band	gamma band	all bands	all bands
SVM	60.50/14.14	60.95/10.20	66.64/14.41	80.76/11.56	79.56/11.38	83.99/09.92	56.61/20.05
GSCCA [66]	63.92/11.16	64.64/10.33	70.10/14.76	76.93/11.00	77.98/10.72	82.96/09.95	69.08/16.66
DBN [7]	64.32/12.45	60.77/10.42	64.01/15.97	78.92/12.48	79.19/14.58	86.08/08.34	66.77/07.38
STRNN [10]	80.90/12.27	83.35/09.15	82.69/12.99	83.41/10.16	69.61/15.65	89.50/07.63	-
DGCNN [12]	74.25/11.42	71.52/05.99	74.43/12.16	83.65/10.17	85.73/10.64	90.40/08.49	69.88/16.29
A-LSTM [67]	-	-	-	-	-	-	69.50/15.65
BiDANN [17]	76.97/10.95	75.56/07.88	81.03/11.74	89.65/09.59	88.64/09.46	92.38/07.04	70.29/12.63
EmotionMeter [25]	-	-	-	-	-	-	70.58/17.01
BiHDM [13] (SOTA)	-	-	-	-	-	93.12/06.06	74.35/14.09
RGNN (Our model)	76.17/07.91	72.26/07.25	75.33/08.85	84.25/12.54	89.23/08.90	94.24/05.95	79.37/10.54

TABLE 2: Subject-independent classification accuracy (mean/standard deviation) on SEED and SEED-IV

Model	SEED						SEED-IV
	delta band	theta band	alpha band	beta band	gamma band	all bands	all bands
SVM	43.06/08.27	40.07/06.50	43.97/10.89	48.63/10.29	51.59/11.83	56.73/16.29	37.99/12.52
TCA [68]	44.10/08.22	41.26/09.21	42.93/14.33	43.93/10.06	48.43/09.73	63.64/14.88	56.56/13.77
SA [69]	53.23/07.47	50.60/08.31	55.06/10.60	56.72/10.78	64.47/14.96	69.00/10.89	64.44/09.46
T-SVM [70]	-	-	-	-	-	72.53/14.00	-
TPT [71]	-	-	-	-	-	76.31/15.89	-
DGCNN [12]	49.79/10.94	46.36/12.06	48.29/12.28	56.15/14.01	54.87/17.53	79.95/09.02	52.82/09.23
A-LSTM [67]	-	-	-	-	-	-	55.03/09.28
DAN [72]	-	-	-	-	-	83.81/08.56	58.87/08.13
BiDANN-S [17]	63.01/07.49	63.22/07.52	63.50/09.50	73.59/09.12	73.72/08.67	84.14/06.87	65.59/10.39
BiHDM [13] (SOTA)	-	-	-	-	-	85.40/07.53	69.03/08.66
RGNN (Our model)	64.88/06.87	60.69/05.79	60.84/07.57	74.96/08.94	77.50/08.10	85.30/06.72	73.84/08.02

5.3 Model Settings in RGNN

For our RGNN in all experiments, we empirically set the number of convolutional layers $L = 2$, dropout rate [73] of 0.7 at the output fully-connected layer, and batch size of 16. We use Adam optimization [74] with default values, i.e., $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We only tune the output feature dimension d' , label noise level ϵ , learning rate η , L1 regularization factor α , and L2 regularization for each experiment. Note that we only adopt NodeDAT in subject-independent classification experiments. We compare our model with several baselines, which are cited from published results [10], [12], [13], [17].

6 PERFORMANCE EVALUATIONS

In this section we present model evaluation results in both subject-dependent and subject-independent classification settings on both datasets. We also investigate critical frequency bands and confusion matrix of our model.

6.1 Subject-Dependent Classification

Table 1 presents the subject-dependent classification accuracy (mean/standard deviation) of our RGNN model and all baselines on both SEED and SEED-IV using the pre-computed DE features. The performance on SEED using DE feature in the individual delta, theta, alpha, beta, and gamma bands is reported as well. It is encouraging to see that our model achieves superior performance on both datasets as compared to all baselines including the state-of-the-art BiHDM when DE features from all frequency bands are used. It is worth noting that our model improves the accuracy of the state-of-the-art model on SEED-IV by around 5%. In particular, our model performs better than

DGCNN, which is another GNN-based model that leverages the topological structure in EEG signals. Besides the proposed two regularizers (see Table 3), the main performance improvement can be attributed to two factors: 1) our adjacency matrix incorporates the global inter-channel asymmetry relation between the left and right hemispheres; and 2) our model has less concern of overfitting by extending SGC, which is much simpler than ChebNet [43] used in DGCNN.

6.2 Subject-Independent Classification

Similar to Table 1, Table 2 presents the subject-independent classification results. When using features from all frequency bands, our model performs marginally worse than BiHDM on SEED but much better than BiHDM on SEED-IV (nearly 5% improvement). In addition, our model achieves the lowest standard deviation in accuracy compared to all baselines on both datasets, demonstrating the robustness of our model.

Comparing the results shown in Tables 1 and 2, we find that the accuracy obtained in subject-independent settings is consistently worse than the accuracy obtained in subject-dependent settings by around 5% to 30% for every model. This finding is unsurprising because the variability of EEG signals across subjects makes subject-independent classification more challenging. However, the interesting part is that the performance gap between these two settings is gradually decreasing from around 27% on SEED and 19% on SEED-IV using SVM to around 9% on SEED and 6% on SEED-IV using our model. One possible reason for the diminishing gap is that recent deep learning models in subject-independent settings are becoming better at leveraging a larger amount of data and learning more subject-invariant EEG representations. This observation seems to indicate

that transfer learning may be a necessary tool for emotion recognition in cross-subject settings. With the increasing amount of data available from different subjects and a proper transfer learning tool, it would not be surprising that subject-independent classification accuracy will surpass the subject-dependent classification accuracy in the future.

6.3 Performance Comparison of Frequency Bands

We further compare the performance of our model and all baselines using features from different frequency bands, as reported in Tables 1 and 2. In subject-dependent experiments on SEED, STRNN achieves the highest accuracy in delta, theta and alpha bands, BiDANN performs best in beta band, and our model performs best in gamma band. In subject-independent experiments on SEED, BiDANN-S achieves the highest accuracy in theta and alpha bands, and our model performs best in delta, beta and gamma bands.

We investigate the critical frequency bands for emotion recognition. For both subject-dependent and subject-independent settings on SEED, we compare the performance of each model across different frequency bands. In general, most models including our model achieve better performance on beta and gamma bands than delta, theta and alpha bands, with one exception of STRNN, which performs the worst on gamma band. This observation is consistent with the literature [7], [75]. One subtle difference between our model and other models is that our model performs consistently better in gamma band than beta band, whereas other models perform comparably in both bands, indicating that gamma band may be the most discriminative band for our model.

6.4 Confusion Matrix

We present the confusion matrix of our model in Fig. 3. For both subject-dependent and subject-independent settings on SEED, our model can recognize better for positive and neutral emotions than negative emotion. By combining training data from other subjects (see Fig. 3 (a) and (b)), our model is getting much worse at detecting negative emotion, indicating that participants are likely to generate distinct EEG patterns when experiencing negative emotion. Similar phenomenon is observed in SEED-IV for sad emotion as well (see Fig. 3 (c) and (d)). For SEED-IV, our model performs significantly better on sad emotion than all other emotions in both classification settings. We notice that fear is the only emotion that performs better in subject-independent classification than in subject-dependent classification. This finding indicates that participants watching horror movies may generate similar EEG patterns.

7 MODEL ANALYSIS ON RGNN

In this section we conduct ablation study and sensitivity analysis for model.

7.1 Ablation Study

We conduct ablation study to investigate the contribution of each key component in our model. Table 3 reports the results obtained in subject-independent setting on both datasets.

TABLE 3: Ablation study for subject-independent classification accuracy (mean/standard deviation) on SEED and SEED-IV. Symbols - and + indicate the following component is removed and added, respectively.

Model	SEED	SEED-IV
RGNN	85.30/06.72	73.84/08.02
- global connection	82.42/08.24	71.13/08.78
- symmetric adjacency matrix	83.69/07.92	72.02/08.66
- NodeDAT	81.92/09.35	71.65/09.43
- NodeDAT + DAT	83.51/08.11	72.40/08.54
- EmotionDL	82.27/08.81	70.76/09.22

The two major designs in our adjacency matrix \mathbf{A} , i.e., global connection and symmetric adjacency matrix designs, are helpful in recognizing emotions. The global connection models the asymmetric difference between neuronal activities in the left and right hemispheres and have been shown to reveal certain emotions [5], [60], [61]. The symmetric adjacency matrix design is mostly motivated to reduce the number of model parameters and prevent overfitting, especially in subject-dependent classifications where lesser training data is available.

Our NodeDAT regularizer has a noticeable positive impact on the performance of our model, which demonstrates that domain adaptation is significantly helpful in cross-subject classification. To further investigate the impact of our node-level domain classifier, we further experimented with replacing NodeDAT with a generic domain classifier (DAT) [23] that operates after the pooling operation, i.e., (-NodeDAT + DAT) in Table 3. The clear performance gap between (-NodeDAT + DAT) and our RGNN model indicates that our NodeDAT can better regularize the model by learning subject-invariant representation at node level than graph level. In addition, if NodeDAT is removed, the performance of our model has a greater variance, demonstrating the importance of NodeDAT in improving the robustness of our model against cross-subject variations.

Our EmotionDL regularizer improves performance of our model by around 3% in accuracy on both datasets. This performance gain validates our assumption that participants are not always generating the intended emotions when watching emotion-eliciting stimuli. In addition, our EmotionDL can be easily adopted by other deep learning models.

7.2 Sensitivity Analysis

We analyze the performance of our model across varying L1 sparsity coefficient α (see (11)) and noise coefficient ϵ in EmotionDL (see (15) and (16)), as illustrated in Fig. 4. For subject-dependent classification, increasing α from 0 to 0.1 will generally increase the model performance. However, for subject-independent classification, increasing α beyond a certain threshold, i.e, 0.01 in Fig. 4(a), will decrease the model performance. One possible explanation for the difference in model behaviors is that there is much less training data in subject-dependent classification, which requires a stronger regularization to reduce overfitting, whereas for subject-independent classification where the number of training data is less of a concern, adding stronger

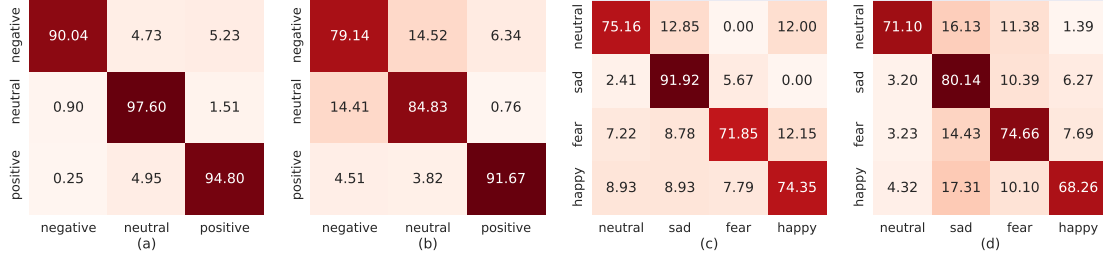


Fig. 3: Confusion matrix of RGNN. (a) subject-dependent classification on SEED. (b) subject-independent classification on SEED. (c) subject-dependent classification on SEED-IV. (d) subject-independent classification on SEED-IV.

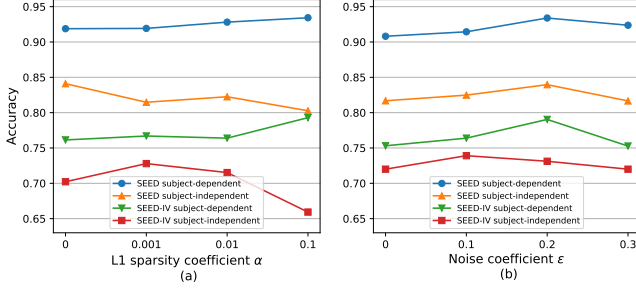


Fig. 4: Classification accuracy of RGNN with varying (a) L1 sparsity coefficient α in (11) and (b) noise coefficient ϵ in (15) and (16)

regularization may introduce bias and hinder the learning efficacy.

As illustrated in Fig. 4(b), our model behaves consistently across different experimental settings with varying noise coefficient ϵ . Specifically, by increasing ϵ , the performance of our model first increases and then decreases. In particular, our model usually performs best when ϵ is set to 0.2, demonstrating the existence of label noises and the necessity of addressing them on both datasets. Introducing excessive noise in EmotionDL causes performance drop, which is expected because excessive noise weakens the true learning signals.

8 NEURONAL ACTIVITY ANALYSIS FOR EMOTION RECOGNITION

In this section we analyze and identify important neuronal activities for emotion recognition.

8.1 Activation Maps of Channels

Fig. 5 shows the heatmap of the diagonal elements in our learned adjacency matrix \mathbf{A} . Conceptually, as shown in (4), the diagonal values in \mathbf{A} represents the contribution of each channel in computing the final EEG representation. It is clear from Fig. 5 that there are strong activations on the pre-frontal, parietal, and occipital regions, indicating that these regions may be strongly related to the emotion processing of the brain. Our finding is consistent with existing studies, which observed that asymmetrical frontal and parietal EEG activity may reflect changes on both valence and arousal [5], [30]. The synchronization between frontal and occipital regions has also been reported to be related to positive and

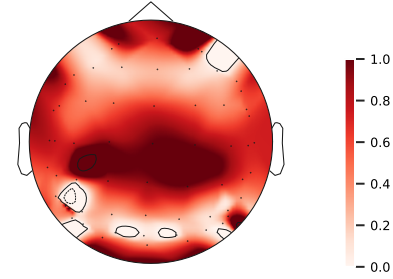


Fig. 5: Heatmap of the diagonal values in the learned adjacency matrix \mathbf{A} , which is averaged across five frequency bands in subject-dependent classification on both SEED and SEED-IV. The values for each channel is further scaled to the $[0, 1]$ interval for better visualization.

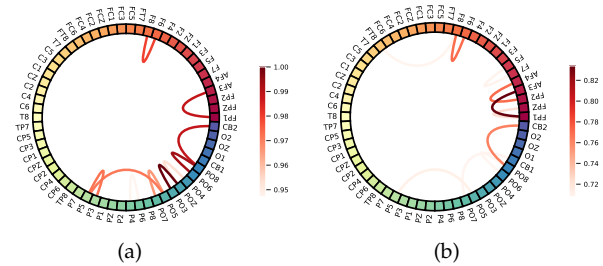


Fig. 6: Top 10 edge weights between electrodes in the adjacency matrix \mathbf{A} , excluding global connections in (10) for better clarity. (a): Initialized \mathbf{A} according to (9). (b): Learned and averaged \mathbf{A} across five frequency bands in subject-dependent classification on both SEED and SEED-IV.

fear emotion [76], [77]. The symmetry pattern on the activation map of channels indicate again that the asymmetry in EEG activity between the left and right hemispheres is critical for emotion recognition.

8.2 Inter-channel Relations

Fig. 6 shows the top 10 connections between channels having the largest edge weights in our adjacency matrix \mathbf{A} . Note that all global connections remain among the strongest connections after \mathbf{A} is learned, demonstrating again that global inter-channel relations are essential for emotion recognition. It is obvious from Fig. 6 that there are

both similarities and differences between these two plots, indicating that our initialization strategy presented in (9) can capture local inter-channel relations to a certain degree. One notable difference between the two plots is that a few strong connections are gone in Fig. 6(a), e.g., (POZ, PO3), (PO6, PO8), and (P3, P5), indicating that these connections may not be critical for emotion recognition. In addition, it is clear from Fig. 6(b) that the connection between the channel pair (FP1, AF3) is the strongest, followed by (F6, F8), (FP2, AF4), and (PO8, CB2), indicating that local inter-channel relations in the frontal region may be important for emotion recognition.

9 CONCLUSION

In this paper, we propose a regularized graph neural network for emotion recognition based on EEG signals. Our model is biologically supported to capture both local and global inter-channel relations. In addition, we propose two regularizers, namely NodeDAT and EmotionDL, to improve the robustness of our model against cross-subject EEG variations and noisy labels. We evaluate our model in both subject-dependent and subject-independent classification settings on two public datasets SEED and SEED-IV. Our model obtains better performance than a few competitive baselines such as SVM, DBN, DGCNN, BiDANN, and the state-of-the-art BiHDM in most classification settings. Notably, our model achieves accuracy of 79.37% and 73.84% in subject-dependent and subject-independent classifications on SEED-IV, respectively, outperforming the current state-of-the-art model by around 5%. Our model analysis demonstrates that our proposed biologically supported adjacency matrix and two regularizers contribute consistent and significant gain to the performance of our model. Investigations on the neuronal activities reveal that pre-frontal, parietal and occipital regions may be the most informative regions in emotion recognition. In addition, global inter-channel relations between the left and right hemispheres are important and local inter-channel relations between (FP1, AF3), (F6, F8) and (FP2, AF4) may also provide useful information.

In the future, we plan to investigate how to apply our model to EEG signals that have a smaller number of channels. A simpler version of our model may be necessary to avoid overfitting on these datasets. In addition, how to incorporate global connections on these smaller graphs may be worth exploring.

REFERENCES

- [1] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using EEG signals: a survey," *IEEE Transactions on Affective Computing*, 2017.
- [2] U. R. Acharya, V. K. Sudarshan, H. Adeli, J. Santhosh, J. E. Koh, and A. Adeli, "Computer-aided diagnosis of depression using EEG signals," *European neurology*, vol. 73, no. 5-6, pp. 329-336, 2015.
- [3] P. Ekman and D. Keltner, "Universal facial expressions of emotion," *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture*, pp. 27-46, 1997.
- [4] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, no. 4, pp. 261-292, 1996.
- [5] L. A. Schmidt and L. J. Trainor, "Frontal brain electrical activity (EEG) distinguishes valence and intensity of musical emotions," *Cognition & Emotion*, vol. 15, no. 4, pp. 487-500, 2001.
- [6] X.-W. Wang, D. Nie, and B.-L. Lu, "Emotional state classification from EEG data using machine learning approach," *Neurocomputing*, vol. 129, pp. 94-106, 2014.
- [7] 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.
- [8] L.-C. Shi, Y.-Y. Jiao, and B.-L. Lu, "Differential entropy feature for EEG-based vigilance estimation," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2013, pp. 6627-6630.
- [9] D. Zhang, L. Yao, X. Zhang, S. Wang, W. Chen, R. Boots, and B. Benatallah, "Cascade and parallel convolutional recurrent neural networks on EEG-based intention recognition for brain computer interface," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [10] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial-temporal recurrent neural network for emotion recognition," *IEEE Transactions on Cybernetics*, no. 99, pp. 1-9, 2018.
- [11] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*, 2013, pp. 1310-1318.
- [12] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Transactions on Affective Computing*, 2018.
- [13] Y. Li, W. Zheng, L. Wang, Y. Zong, L. Qi, Z. Cui, T. Zhang, and T. Song, "A novel bi-hemispheric discrepancy model for EEG emotion recognition," *arXiv preprint arXiv:1906.01704*, 2019.
- [14] W.-L. Zheng and B.-L. Lu, "Personalizing EEG-based affective models with transfer learning," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 2016, pp. 2732-2738.
- [15] X. Chai, Q. Wang, Y. Zhao, Y. Li, D. Liu, X. Liu, and O. Bai, "A fast, efficient domain adaptation technique for cross-domain electroencephalography (EEG)-based emotion recognition," *Sensors*, vol. 17, no. 5, p. 1014, 2017.
- [16] Z. Lan, O. Sourina, L. Wang, R. Scherer, and G. R. Müller-Putz, "Domain adaptation techniques for EEG-based emotion recognition: a comparative study on two public datasets," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 11, no. 1, pp. 85-94, 2018.
- [17] Y. Li, W. Zheng, Y. Zong, Z. Cui, T. Zhang, and X. Zhou, "A bi-hemisphere domain adversarial neural network model for EEG emotion recognition," *IEEE Transactions on Affective Computing*, 2018.
- [18] P. Erdős and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci.*, vol. 5, no. 1, pp. 17-60, 1960.
- [19] A. Fornito, A. Zalesky, and M. Breakspear, "Graph analysis of the human connectome: promise, progress, and pitfalls," *Neuroimage*, vol. 80, pp. 426-444, 2013.
- [20] E. Bullmore and O. Sporns, "The economy of brain network organization," *Nature Reviews Neuroscience*, vol. 13, no. 5, p. 336, 2012.
- [21] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger, "Simplifying graph convolutional networks," in *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 2019, pp. 6861-6871.
- [22] R. C. Craddock, S. Jbabdi, C.-G. Yan, J. T. Vogelstein, F. X. Castellanos, A. Di Martino, C. Kelly, K. Heberlein, S. Colcombe, and M. P. Milham, "Imaging human connectomes at the macroscale," *Nature methods*, vol. 10, no. 6, p. 524, 2013.
- [23] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096-2030, 2016.
- [24] X. Zhu and X. Wu, "Class noise vs. attribute noise: A quantitative study," *Artificial intelligence review*, vol. 22, no. 3, pp. 177-210, 2004.
- [25] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE Transactions on Cybernetics*, no. 99, pp. 1-13, 2018.
- [26] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 327-339, 2014.
- [27] K. Takahashi and A. Tsukaguchi, "Remarks on emotion recognition from multi-modal bio-potential signals," in *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and*

- Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)*, vol. 2. IEEE, 2003, pp. 1654–1659.
- [28] C. Tang, D. Wang, A.-H. Tan, and C. Miao, “EEG-based emotion recognition via fast and robust feature smoothing,” in *International Conference on Brain Informatics*. Springer, 2017, pp. 83–92.
- [29] Y. Liu and O. Sourina, “Real-time fractal-based valence level recognition from EEG,” in *Transactions on Computational Science XVIII*. Springer, 2013, pp. 101–120.
- [30] 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*, vol. 57, no. 7, pp. 1798–1806, 2010.
- [31] M. Akin, “Comparison of wavelet transform and FFT methods in the analysis of EEG signals,” *Journal of Medical Systems*, vol. 26, no. 3, pp. 241–247, 2002.
- [32] X. Wu, W.-L. Zheng, and B.-L. Lu, “Identifying functional brain connectivity patterns for EEG-based emotion recognition,” in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2019, pp. 235–238.
- [33] P. Li, H. Liu, Y. Si, C. Li, F. Li, X. Zhu, X. Huang, Y. Zeng, D. Yao, and Y. Zhang, “EEG based emotion recognition by combining functional connectivity network and local activations,” *IEEE Transactions on Biomedical Engineering*, 2019.
- [34] W.-L. Zheng, J.-Y. Zhu, Y. Peng, and B.-L. Lu, “EEG-based emotion classification using deep belief networks,” in *2014 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2014, pp. 1–6.
- [35] B. H. Kim and S. Jo, “Deep physiological affect network for the recognition of human emotions,” *IEEE Transactions on Affective Computing*, 2018.
- [36] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, “Learning representations from EEG with deep recurrent-convolutional neural networks,” *arXiv preprint arXiv:1511.06448*, 2015.
- [37] X. Li, D. Song, P. Zhang, G. Yu, Y. Hou, and B. Hu, “Emotion recognition from multi-channel EEG data through convolutional recurrent neural network,” in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2016, pp. 352–359.
- [38] J. Li, Z. Zhang, and H. He, “Hierarchical convolutional neural networks for EEG-based emotion recognition,” *Cognitive Computation*, pp. 1–13, 2018.
- [39] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, “A comprehensive survey on graph neural networks,” *arXiv preprint arXiv:1901.00596*, 2019.
- [40] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, “The graph neural network model,” *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2008.
- [41] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, “Spectral networks and locally connected networks on graphs,” *arXiv preprint arXiv:1312.6203*, 2013.
- [42] F. R. Chung and F. C. Graham, *Spectral Graph Theory*. American Mathematical Soc., 1997, no. 92.
- [43] M. Defferrard, X. Bresson, and P. Vandergheynst, “Convolutional neural networks on graphs with fast localized spectral filtering,” in *Advances in Neural Information Processing Systems*, 2016, pp. 3844–3852.
- [44] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *International Conference on Learning Representations (ICLR)*, 2017.
- [45] P. Velickovi, G. Cucurull, A. Casanova, A. Romero, P. Li, and Y. Bengio, “Graph attention networks,” in *International Conference on Learning Representations (ICLR)*, 2018.
- [46] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, “Gated graph sequence neural networks,” *arXiv preprint arXiv:1511.05493*, 2015.
- [47] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. J. Smola, “Correcting sample selection bias by unlabeled data,” in *Advances in Neural Information Processing Systems*, 2007, pp. 601–608.
- [48] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, “A kernel two-sample test,” *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723–773, 2012.
- [49] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, “Deep reconstruction-classification networks for unsupervised domain adaptation,” in *European Conference on Computer Vision*. Springer, 2016, pp. 597–613.
- [50] S. Benaïm and L. Wolf, “One-sided unsupervised domain mapping,” in *Advances in Neural Information Processing Systems*, 2017, pp. 752–762.
- [51] F. M. Carriucci, L. Porzi, B. Caputo, E. Ricci, and S. R. Bulò, “Autodial: Automatic domain alignment layers,” in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 5077–5085.
- [52] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu, “Making deep neural networks robust to label noise: A loss correction approach,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1944–1952.
- [53] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus, “Training convolutional networks with noisy labels,” *arXiv preprint arXiv:1406.2080*, 2014.
- [54] B. Van Rooyen, A. Menon, and R. C. Williamson, “Learning with symmetric label noise: The importance of being unhinged,” in *Advances in Neural Information Processing Systems*, 2015, pp. 10–18.
- [55] J. P. Brooks, “Support vector machines with the ramp loss and the hard margin loss,” *Operations Research*, vol. 59, no. 2, pp. 467–479, 2011.
- [56] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, “Training deep neural networks on noisy labels with bootstrapping,” *arXiv preprint arXiv:1412.6596*, 2014.
- [57] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa, “Joint optimization framework for learning with noisy labels,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5552–5560.
- [58] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, “Deep label distribution learning with label ambiguity,” *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2825–2838, 2017.
- [59] R. Salvador, J. Suckling, M. R. Coleman, J. D. Pickard, D. Menon, and E. Bullmore, “Neurophysiological architecture of functional magnetic resonance images of human brain,” *Cerebral cortex*, vol. 15, no. 9, pp. 1332–1342, 2005.
- [60] S. J. Dimond, L. Farrington, and P. Johnson, “Differing emotional response from right and left hemispheres,” *Nature*, vol. 261, no. 5562, p. 690, 1976.
- [61] G. Zhao, Y. Zhang, and Y. Ge, “Frontal EEG asymmetry and middle line power difference in discrete emotions,” *Frontiers in Behavioral Neuroscience*, vol. 12, 2018.
- [62] S. Achard and E. Bullmore, “Efficiency and cost of economical brain functional networks,” *PLoS Computational Biology*, vol. 3, no. 2, p. e17, 2007.
- [63] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, “How powerful are graph neural networks?” in *International Conference on Learning Representations (ICLR)*, 2019.
- [64] S. Kullback and R. A. Leibler, “On information and sufficiency,” *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.
- [65] L.-C. Shi and B.-L. Lu, “Off-line and on-line vigilance estimation based on linear dynamical system and manifold learning,” in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2010, pp. 6587–6590.
- [66] W. Zheng, “Multichannel EEG-based emotion recognition via group sparse canonical correlation analysis,” *IEEE Transactions on Cognitive and Developmental Systems*, vol. 9, no. 3, pp. 281–290, 2016.
- [67] T. Song, W. Zheng, C. Lu, Y. Zong, X. Zhang, and Z. Cui, “MPED: A multi-modal physiological emotion database for discrete emotion recognition,” *IEEE Access*, vol. 7, pp. 12 177–12 191, 2019.
- [68] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, “Domain adaptation via transfer component analysis,” *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2010.
- [69] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, “Unsupervised visual domain adaptation using subspace alignment,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2960–2967.
- [70] R. Collobert, F. Sinz, J. Weston, and L. Bottou, “Large scale transductive svms,” *Journal of Machine Learning Research*, vol. 7, pp. 1687–1712, 2006.
- [71] E. Sangineto, G. Zen, E. Ricci, and N. Sebe, “We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer,” in *Proceedings of the 22nd ACM International Conference on Multimedia*. ACM, 2014, pp. 357–366.
- [72] H. Li, Y.-M. Jin, W.-L. Zheng, and B.-L. Lu, “Cross-subject emotion recognition using deep adaptation networks,” in *Proceedings of the International Conference on Neural Information Processing*. Springer, 2018, pp. 403–413.
- [73] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

- [74] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [75] W. J. Ray and H. W. Cole, "EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes," *Science*, vol. 228, no. 4700, pp. 750–752, 1985.
- [76] T. Costa, E. Rognoni, and D. Galati, "EEG phase synchronization during emotional response to positive and negative film stimuli," *Neuroscience Letters*, vol. 406, no. 3, pp. 159–164, 2006.
- [77] G. Mattavelli, M. Rosanova, A. G. Casali, C. Papagno, and L. J. R. Lauro, "Timing of emotion representation in right and left occipital region: evidence from combined tms-EEG," *Brain and Cognition*, vol. 106, pp. 13–22, 2016.