Improving the Interpretability through Maximizing Mutual Information for EEG Emotion Recognition

Hua Yang, C. L. Philip Chen, *Fellow, IEEE,* Bianna Chen, Tong Zhang*, *Member, IEEE,*

***Abstract*—EEG-based emotion recognition has attracted significant attention. Recently, Graph Neural Networks (GNNs) have achieved remarkable performance by utilizing a dynamic matrix to model these spatial connectivities. However, those GNNs need more interpretability, meaning that they are untrustworthy. A trustworthy GNN should have good emotion classification performance with clear reasoning. Thus, the essential factors that limit GNN's interpretability include spurious spatial connectivities and unclear rationales of predictions. This paper aims to improve GNN's interpretability by solving the above two issues. An Adjacency-Explainable Graph Neural Network (AEG) is proposed to remove spurious connections. Inspired by Graph Information Bottleneck, AEG extracts an informative representation by optimizing the Mutual Information (MI) to capture genuine relationships. Extensive experiments are conducted using AEG on three databases: SEED, SEED-IV, and DREAMER. In addition, this paper also proposes a post-hoc explainer called Channel-wise Adaptive Class Activation Mapping Explainer (CACA) to provide rationales. CACA incorporates the spatial information of GNNs to identify underlying channels that maximize the MI effectively. A quantitative comparison regarding the faithfulness of post-hoc explainers is presented, demonstrating the superiority of CACA. Furthermore, this paper provides two possible applications to illustrate the potential of the proposed methods.**

***Index Terms*—EEG Emotion Recognition, Interpretability, Graph Neural Network, Explainable Artificial Intelligence**

I. INTRODUCTION

EMOTION recognition plays a vital role in comprehending human behavior in intelligent human-computer interaction [1]. Emotion recognition is widely used in health care [2], therapy of Autism Spectrum Disorders [3], fatigue driving [4], and other domains. Compared to other signals, the electroencephalogram (EEG) signal has gained significant attention in emotion recognition research [5] because of its non-invasive recording methods, ability to capture authentic emotions, and device portability. Multiple electrodes placed along the scalp record the EEG signal, forming irregular non-grid data. Previous neuroscience studies [6]–[8] have revealed that the functional connections between distinct brain regions are highly related to the generation of emotions, with each emotion being accompanied by unique activation patterns in specific brain regions. As a result, the graph-based deep learning method, particularly Graph Neural Networks (GNNs), has delivered remarkable results.

Recent studies [9]–[16] have proposed utilizing dynamic GNNs for extracting emotion-discriminative representations from EEG data. These models employ a dynamic adjacency matrix to capture the functional dependencies between brain regions (i.e., the connections between different EEG channels). However, their dynamic adjacency matrixes are learned directly in a data-driven manner, which increases the risk of capturing spurious connections (emotion-irrelevant connections) caused by data noise [12]. Besides, those GNNs are deployed without interpretability, meaning the underlying mechanisms behind their predictions are unclear. GNNs that lack interpretability may introduce biases, rendering them unsuitable for deployment in decision-sensitive areas, particularly in the medical field.

Previously Liu et al. [11] proposed a fine-grained post-hoc explainer called Concat-aided Grad-CAM, inspired by the Gradcam method, which aims to capture the underlying channels of pre-trained GNNs. Concat-aided Grad-CAM provided a means to verify whether the trained GNN is biased. However, since post-hoc explainers do not alter the trained GNNs, spurious connections are likely to persist. Miao et al. [17] pointed out that the trained GNNs often have a suboptimal training result, implying that more than relying solely on post-hoc explainers is required. On the other hand, GNN model is very sensitive to noise EEG data. This problem yields the potential safety concerns to deploy GNNs in the practical systems, such as health care. Thus, it is crucial to train GNNs to remove spurious connections among EEG channels. This paper focuses on improving the interpretability of GNNs in EEG-based emotion recognition tasks through two main aspects: removing the spurious connections during training and uncovering the underlying channels during inferencing.

For inherent interpretability, inspired by the previous research [17]–[19], the Graph Information Bottleneck (GIB) is introduced to eliminate the spurious connections among EEG channels. This paper proposes a practical and straightforward model called Adjacency-Explainable Graph Neural Network (AEG), which aims to maximize the GIB between the learned connections and the ground truth emotions. AEG learns the vital connections by capturing an optimal representation of the input data in the hidden space. AEG extracts more discriminative features with genuine emotion-relevant connections, which is empirically validated in emotion recognition experiments on three widely used datasets, i.e., SEED [20], SEED-IV [21], and DREAMER [22].

This paper also proposes a novel post-hoc explainer,

H. Yang, C. L. Philip Chen, B. Chen, and T. Zhang are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China. (*Corresponding authors: T. Zhang, E-mail: tony@scut.edu.cn.)

C. L. Philip and T. Zhang are also with the Pazhou Lab, Guangzhou 510335, China, and Engineering Research Center of the Ministry of Education on Health Intelligent Perception and Paralleled Digital-Human, Guangzhou, China.

T. Zhang is also with the Guangdong Provincial Key Laboratory of Computational Intelligence and Cyberspace Information.

Channel-wise Adaptive Class Activation Mapping Explainer (CACA), to identify the underlying channels of the model's decisions. CACA utilizes GNN spatial information as guidance to find a focused search space and ultimately identify the channels that maximizes Mutual Information (MI) as the explanation. The performance of several baselines and CACA is fairly compared using standard evaluation metrics for post-hoc explainers [23]–[25]. The results demonstrate that the proposed CACA obtains state-of-the-art performance. Additionally, this paper provides two potential applications, i.e., debugging trained GNN and investigating activation regions for different emotions, on improving the interpretability of GNNs in the field of EEG-based emotion recognition.

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

- This paper proposes an Adjacency-Explainable Graph Neural Network (AEG) to remove spurious connections among EEG channels. AEG optimizes two variational approximation bounds to extract more discriminative features with genuine emotion-relevant connections preserved.
- A novel post-hoc explainer, CACA, is also proposed to identify the underlying channels of trained GNNs. The CACA incorporates spatial information to guide the explanation process and searches for the channels that maximize the Mutual Information.
- Extensive experiments are conducted on three widely used datasets, namely SEED, SEED-IV, and DREAMER. Emotion recognition results in subject-dependent and subject-independent settings demonstrate the effectiveness of AEG. In addition, CACA is empirically shown to have a significant improvement over other post-hoc explainers.

The remainder of this paper is organized as follows: Section II concisely reviews relevant works. Section III presents two proposed methods. Section III-B specifies the AEG and the corresponding emotion recognition experiments are reported in Section III-C. Section III-D introduces the post-hoc explainer, CACA. The evaluation of CACA's faithfulness is discussed in Section III-E. In Section IV, two examples of analyzing emotion-related EEG channels are provided. Finally, Section V summarizes this paper.

II. RELATED WORKS

A. EEG Emotion Recognition

Recently, there has been a growing interest in the EEG-based emotion recognition task, owing to its potential for detecting genuine emotions. Early research primarily relied on machine learning methods based on handcrafted features. Examples of such methods include support vector machines (SVM) [26], k-nearest neighbors [27], and decision tree [28]. However, these methods struggle to extract deep features of EEG data, limiting further improvement in recognition performance. Consequently, recent studies have focused on using deep learning methods for emotion recognition. GNNs have gained popularity for EEG emotion recognition, as they can explore the connectivity relationships between different EEG channels. Song et al. [9] proposed using a dynamic graph convolutional neural network (DGCNN) to extract features and perform classification. Their approach incorporates a learnable matrix to model the connections between EEG channels. Wang et al. [15] introduced the broad learning system (BLS) [29] based on DGCNN and proposed a graph convolutional generalized network (GCB-Net) that can extract deeper information. To extract deeper EEG features and avoid overfitting, Li et al. [16] incorporated residual networks into GCB-Net. Additionally, Zhang et al. [30] incorporated sparsity constraints into the dynamic graph of DGCNN, inspired by the neuroscience observation that brain region functional dependencies tend to be highly localized and sparse. Song et al. [31] observed significant differences in the interdependence of EEG between individuals and proposed an instance-adaptive graph model (IAG) to capture individual-specific dependencies. Building upon this, they further proposed a variational instance-adaptive graph (V-IAG) [12] to account for uncertain information in EEG data and reduce the impact of noise.

Although recent research has focused extensively on capturing the interdependence among EEG electrode channels, they often lack interpretability. Consequently, the identified interdependence connections often include spurious connections, making it challenging to deploy these models in decision-sensitive areas such as health care. To address this issue, Liu et al. [11] attempted to use post-hoc interpretability methods to identify channels that contribute significantly to emotion recognition. However, the pre-trained model may be biased [17]. In other words, although specific channels may be considered essential for decision-making, they are unrelated to generating emotions.

B. Graph Neural Network

GNNs have demonstrated remarkable performance in handling irregular and non-Euclidean data tasks. Among GNN variants, Spectral GNNs utilize graph signal filters designed in the spectral domain. Spectral GNNs rely on graph theory [32]. Shuman et al. [33] achieved graph signal filtering through spectral decomposition of the graph Laplacian operator. Defferrard et al. [34] proposed using Chebyshev polynomials as an efficient alternative to Laplacian decomposition, significantly reducing the computational complexity. He et al. [35] introduced Bern-Net, a network that represents filtering operations using Bernstein polynomials. Bianchi et al. [36] proposed ARMA, which utilizes rational filters. Among these methods, ChebyNet using Chebyshev polynomials forms a complete basis in the polynomial space, resulting in the highest expressive capacity [37].

Given the crucial role of dependencies between EEG channels in emotion recognition, spectral graph neural networks have been widely used in this field [9], [12], [15], [16], [30], [31]. Most previous research has adopted dynamic graph modeling approaches to capture these dependencies and achieved promising results. However, it is worth noting that such straightforward approaches may capture spurious connections. For example, previous research [12] reports that learned dependencies among EEG channels are easily affected by the noises, such as noise in EEG signals.

C. Interpretability for Graph Neural Network

The expressive power of GNNs is built upon highly nonlinear interconnections of features in irregular graphs [17]. However, without reasoning the underlying mechanisms behind the predictions, deploying GNNs in decision-crucial domains is challenging. Additionally, non-interpretable GNNs are often influenced by redundant information in real-world graph data, limiting their performance. Therefore, interpretable GNNs have become a recent research focus. Interpretability in GNNs can be achieved through two distinct approaches: inherently interpretable methods and post-hoc explainers.

Research on inherently interpretable method is essential as it can modify the target model and extracts more task-related features [18]. Previous study [17] suggested that models trained on real-world graph data may risk capturing spurious correlations between input data and labels, leading to serious generalization issues. Therefore, models should be trained to align with the authentic connection between learned features and labels. Yu et al. [18] introduced the GIB for inherently interpretable method and proposed a bi-level optimization scheme to maximize the GIB. Miao et al. [17] proposed the implementation of GIB primarily through the injection of stochasticity, resulting in improved accuracy. However, the parameters of the interpretable method may not necessarily reflect the factual channels for model decisions.

Post-hoc explainer often proposes various combinatorial search methods to identify parts of input data that significantly impact model predictions [25]. Yuan et al. [24] explores subgraph-level explanations using Monte Carlo Tree Search (MCTS). Zhang et al. [23] suggested the utilization of Hamiache and Navarro (HN) value to incorporate structural information, significantly reducing computation time. While there have been numerous post-hoc explainers for GNNs, [38]–[42], the mainstream post-hoc explainer [11] for EEG-based emotion recognition is activity maps [11], [43], heat maps [14], [44], and degree centrality of learned adjacency matrix [12], [13]. However, there has been relatively less research on inherently interpretable methods. Hence, it is imperative to employ both inherently interpretable methods and post-hoc explainers to enhance the interpretability of GNNs in EEG-based emotion recognition tasks.

III. PROPOSED METHODS

This section introduces the details of the proposed Adjacency-Explainable Graph Neural Network (AEG) and Channel-wise Adaptive Class Activation Mapping Explainer (CACA). Corresponding experiments are conducted for each method to demonstrate their effectiveness empirically. The framework of the AEG and CACA is illustrated in Fig. 1 to provide a clear visualization of the proposed methods.

A. Notation for EEG Signals

The raw EEG signals are commonly decomposed into five different frequency bands: δ (1-4 Hz), θ (4-8 Hz), α (8-14 Hz), β (14-30 Hz), and γ (30-50 Hz). Here, n represents the number of nodes (EEG channels), d represents the number of frequency bands, and the transformed EEG signals are represented by $X \in \mathbb{R}^{n\times d}$. A directed graph is represented as $G = \{V, E, X, A\}$, where $V = \{v_i\}_{i=1}^{n}$ denotes the set of channels (i.e., nodes), E denotes the set of connections (i.e., edges) and $A \in \mathbb{R}^{n\times n}$ denotes the adjacency matrix. If $A_{i,j} = 0$, source channel v_i and destination channel v_j are not connected, otherwise $A_{i,j} \neq 0$.

B. Adjacency-Explainable Graph Neural Network

Adjacency-Explainable Graph Neural Network (AEG), inspired by GIB, captures genuine connections between EEG channels and removes spurious ones. The AEG aims to learn an informative representation of irregular input graphs by optimizing mutual information (MI).

Graph Information Bottleneck: Let $I(a;b) \triangleq \sum_{a,b}\mathbb{P}(a,b)log\frac{\mathbb{P}(a,b)}{\mathbb{P}(a)\mathbb{P}(b)}$ denotes the MI between two random variables a and b. A large MI indicates a strong correlation between two random variables. Using the notation above, the formulation of the GIB is as follows:

$$\max_{Z} I(Y;Z) s.t. I(G;Z) \leq I_c, \tag{1}$$

where G denotes the input graph, Y denotes the corresponding label, Z denotes the most informative yet compressed representation, and I_c denotes the information constraint between G and Z. Generally, to uncover the underlying channels, denoted as A_s, the following objective to optimize the MI between A_s and label Y is proposed, i.e., solving

$$\max_{A_s\in\mathbb{A}_{sub}} I(Y;A_s) - \epsilon I(A;A_s), \tag{2}$$

where $\mathbb{A}_{sub}$ denotes the set of all possible subset of A and $\epsilon{>}0$ is a Lagrange multiplier. Maximizing the objective in Equation (2) removes spurious correlations in the training set. Suppose the A_s^* is the optimal channels that contains all the information related to the label Y, namely A_s^* makes $I(A;Y|A_s^*) = 0$ and $I(A;A_s^*|Y) = 0$. Consider the following derivation: $I(Y;A_s) - \epsilon I(A;A_s) = (1-\epsilon)(I(Y;A) - I(A;Y|A_s) - \epsilon I(A;A_s|Y))$, thus the A_s^* maximizes the GIB, meaning that the spurious connections are removed from the original A. Detail about this derivation is given in [17]. However, optimizing the objective in Equation (2) is notoriously challenging due to the intractability of MI [17]–[19].

Optimizing Framework: A novel framework on how to optimize such an objective is introduced, which consists of two parts. The first term, $I(Y;A_s)$, measures the relevance between A_s and Y. In AEG, an extractor f_ϕ with parameter ϕ is trained to extract $A_s \in \mathbb{A}_{sub}$ and block the emotion-irrelevant connections between channels. Building upon previous research [17], [45], AEG derives a lower bound for the first term as follows:

$$I(Y;A_s) \geq \mathbb{E}_{A_s,Y}[log\ p(Y|A_s)]. \tag{3}$$

By substituting the true posterior $p(Y|A_s)$ with a variational approximation $f_\psi(y|f_\phi(A_s|A))$, a tractable lower bound of the first term:

$$I(Y;A_s) \quad \geq \quad \mathbb{E}_{A,Y}[f_\psi(y|f_\phi(A_s|A))], \tag{4}$$

where the $f_\psi(y|f_\phi(A_s|A))$ essentially works as the feature extractor in AEG. For the second term $I(A;A_s)$, an upper

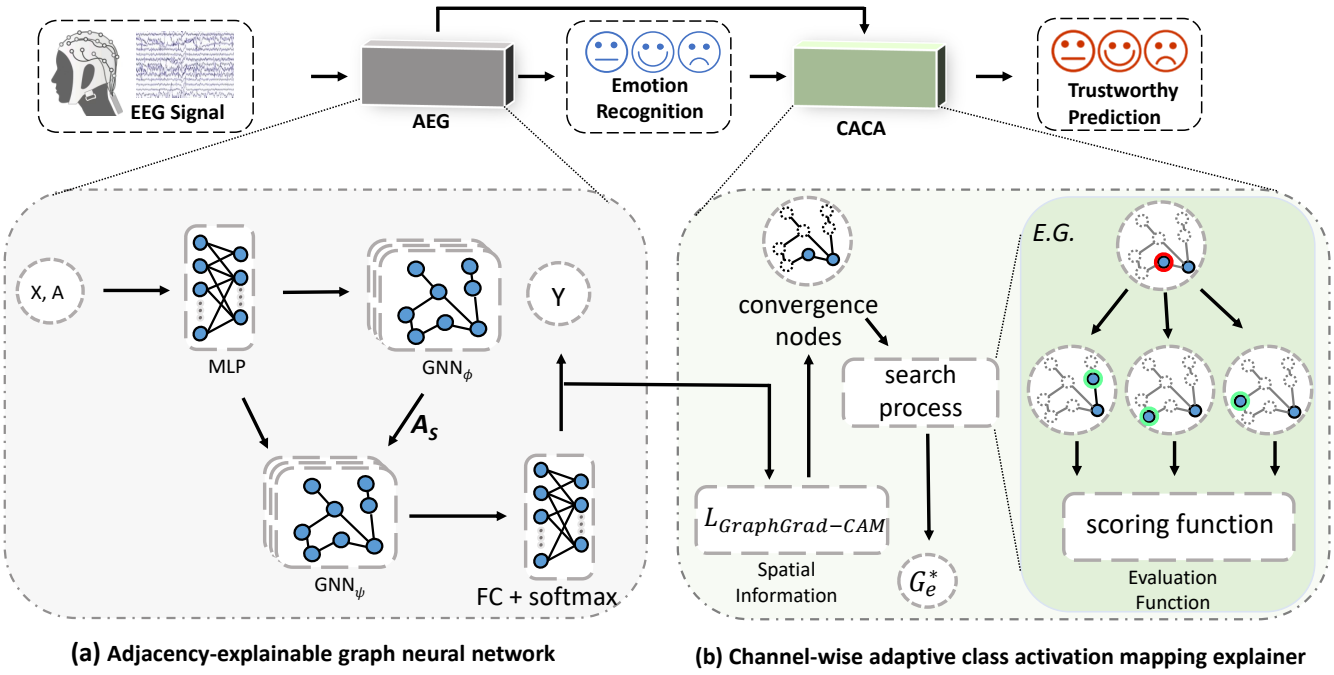

Fig. 1. The framework of proposed Adjacency-Explainable Graph Neural Network (AEG) and Channel-wise Adaptive Class Activation Mapping Explainer (CACA).

bound is obtained by introducing a variational approximation $\mathbb{Q}(A_s)$:

$$I(A; A_s) \leq \mathbb{E}_A[KL(f_\phi(A_s|A)||\mathbb{Q}(A_s))]. \tag{5}$$

By plugging in the above two inequalities, a tractable objective is given:

$$\begin{aligned} min - \mathbb{E}_{A,Y}[f_\psi(y|f_\phi(A_s|A))]+ \\ \epsilon\mathbb{E}_A[KL(f_\phi(A_s|A)||\mathbb{Q}(A_s))]. \end{aligned} \tag{6}$$

The Algorithm of AEG: Since GNNs have the ability to capture critical information from irregular and non-Euclidean data, both the emotion-relevant connections extractor f_ϕ and the feature extractor f_ψ are implemented as spectral GNNs model using Chebyshev polynomials [34]. The graph Laplacian matrix $\mathbf{L}$ of graph structure data x is used to build the basis of the Graph Fourier Transform. Specifically, by taking eigenvectors matrix $\mathbf{U}$ of $\mathbf{L}$ as the basis for the Fourier transform, the graph structure data x can be converted by $\hat{x} = \mathbf{U}^\top x$ [46]. The convolution operations can be performed in the spectral domain:

$$g_\eta(\mathbf{L})x = g_\eta(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top)x = \mathbf{U}g_\eta(\mathbf{\Lambda})\mathbf{U}^\top x, \tag{7}$$

where $\mathbf{\Lambda}$ is the diagonal matrix corresponding to $\mathbf{U}$. However, directly calculating the expression of $g_\eta(\mathbf{\Lambda})$ is challenging. The Chebyshev polynomials simplify this calculation:

$$g_\eta(\mathbf{\Lambda}) = \sum_{k=0}^{K-1} \theta_k T_k(\tilde{\mathbf{\Lambda}}), \tag{8}$$

where θ_k is the coefficient of Chebyshev polynomials, $\tilde{\mathbf{\Lambda}}$ is the normalized $\mathbf{\Lambda}$, K is the order, and $T_k(x)$ can be recursively calculated using the following recursive expressions:

$$\begin{cases} T_0(x) = 1,\ T_1(x) = x, \\ T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x),\ k \geq 2. \end{cases} \tag{9}$$

After extracting discriminative representation by GNN$_\psi$), a full connection layer and a softmax layer are applied to output the predicted labels. The overall framework of AEG is presented in Fig. 1. Besides, AEG uses a multi-layer perceptron layer (mlp) to transform the input EEG data to a representation X_{trans}, which provides a flexible way to adjust the input EEG data. The bound Equation (5) is always true for any $\mathbb{Q}(A_s)$. While the connection between any two EEG channels is possibly related to emotion generation, the $\mathbb{Q}(A_s)$ is defined as an all-ones matrix. For every two node (channel) pair (v_i, v_j) in G, AEG define the $A_{i,j} = 1$. The upper bound in Equation (5) becomes:

$$\begin{aligned} \mathbb{E}_A[KL(f_\phi(A_s|A)||\mathbb{Q}(A_s))] &= \mathbb{E}_A[KL(f_\phi(A_s|A)||1)] \\ &= \mathbb{E}_A[f_\phi(A_s|A)ln(f_\phi(A_s|A))]. \end{aligned} \tag{10}$$

Thus the KL-divergence term becomes a sparse constraint on the learned A_s. And AEG using a L_2-*norm* to approximately optimize KL divergence and reduce computational cost. The details of the AEG algorithm are provided in Algorithm 1.

C. Emotion Recognition Experiment with AEG

In order to empirically prove the spurious connections are removed from the training dataset, a fair comparison between AEG with other baseline methods in emotion recognition

experiments is provided. The intuition behind this is that the learned feature from the input EEG signal will be more discriminative when the connections are solely emotion-relevant. The emotion recognition experiment is conducted on three widely used EEG emotion datasets, i.e., SJTU emotion EEG dataset (SEED) [20], SEED-IV [21], and DREAMER [22]. For the fairness of the comparison, the experiment protocol is the same as previous research [9], [12]–[16], [31], [43]. Two types of experiments, subject-dependent and subject-independent experiments, are implemented to evaluate the proposed method fully.

Algorithm 1 Procedure of the AEG

```
Initialize model parameters
for each i ∈ [0, epoch] do do
    Calculate X_trans with mlp layer;
    Calculate A_s with f_φ, A;
    Calculate H_rep with A_s, f_ψ, X_trans:
    H_rep = f_ψ(A_s, X_trans)
    Calculate Y_pred with the FC layer and the softmax layer;
    Calculate the loss:
    L = L_CrossEntropy(Y_pred, Y) + ||A_s||_2;
    Update parameters in the AEG with gradient descent;
end for
```

1) Implementation Details: For the AEG model, the Chebyshv filter sizes K of f_ϕ and f_ψ are set to 1 and 4, respectively. Besides, the hidden layers of mlp are set to 60, the number of EEG channels n is 62, the number of frequency bands d is set to 5, and the learning rate varies for different datasets. In particular, AEG is implemented by pytorch on a Nvidia 2080 GPU.

2) Emotional EEG Datasets: The SEED consists of EEG signals from 15 subjects (8 females) using the ESI NeuroScan System with 62 channels. Each subject participated in three sessions, and each session comprised 15 trials of EEG signals elicited by emotion-inducing movies (negative, neutral, and positive emotions).

SEED-IV dataset, similar to SEED, comprises EEG signals from 15 subjects, with three sessions for each subject. However, in each session of SEED-IV, there are four types of emotions (happy, neutral, sad, and fear), and each emotion is associated with six distinct film clips. Consequently, there are 24 trials for each subject in each session.

DREAMER is a multi-modal emotion dataset consisting of EEG and ECG signals. The EEG signals were recorded from 23 subjects using 14 electrodes at a sampling rate of 256 Hz. Each subject watched 18 film clips (18 trials) to elicit nine emotions. After watching a film clip, every subject provided subjective assessments of valence, arousal, and dominance using the self-assessment manikins (SAM). All the EEG signals are labeled with binary states (low/high valance, low/high arousal, low/high dominance). Following the same setting as the previous research [9], [22], PSD features were extracted from each trail. More details about this preprocessing are provided in [9].

3) Experiment on SEED: In the subject-dependent setting on SEED, the training set for each subject consists of the first nine trials, while the testing set comprises the remaining six trials. The model performance is evaluated by calculating the accuracy averaged across all subjects over two sessions of SEED. The leave-one-subject-out (LOSO) cross-validation strategy is used for the subject-independent setting on SEED. One session is utilized for each subject to calculate the mean accuracy. The energy feature used in AEG is the differential entropy (DE) provided in [54].

TABLE I
COMPARISONS OF AEG AND STRONG BASELINES ON SEED DATASET FOR SUBJECT-DEPENDENT EXPERIMENT.

Method	ACC / STD (%)
SVM [20]	83.99 / 09.72
GCNN [34]	87.40 / 09.20
DGCNN [9]	90.40 / 08.49
R2G-STNN [47]	93.38 / 05.96
GCB-net+BLS [15]	94.24 / 06.70
Residual GCB-Net+BLS [16]	94.56 / 06.61
V-IAG [12]	95.64 / 05.08
HD-GCN [13]	96.40 / 04.54
AEG (ours)	**97.91 / 03.85**

TABLE II
COMPARISONS OF AEG AND STRONG BASELINES ON SEED DATASET FOR SUBJECT-INDEPENDENT EXPERIMENT.

Method	ACC / STD (%)
SVM [20]	56.73 / 16.29
KPCA [48]	61.28 / 14.62
TCA [49]	63.64 / 14.88
TPT [50]	76.31 / 15.89
DANN [51]	75.08 / 11.18
DGCNN [9]	79.95 / 09.02
BiDANN [52]	83.28 / 09.60
BiDANN-S [53]	84.14 / 06.87
R2G-STNN [47]	84.16 / 07.63
V-IAG [12]	88.38 / 04.80
HD-GCN [13]	89.23 / 04.68
AEG (ours)	**90.53 / 05.11**

Results on Subject-dependent Experiments: The baseline methods include SVM, graph convolutional neural network (GCNN), DGCNN, the spatial and temporal neural networks with regional to global hierarchical feature learning process (R2G-STNN), graph convolutional broad network (GCB-net+BLS), residual graph convolutional broad network (Residual GCB-Net+BLS), variational instance-adaptive graph (V-IAG) and hierarchical dynamic GCN (HD-GCN). The average accuracies and standard deviations are reported as shown in TABLE I. The proposed AEG achieve the best performance among these methods. AEG outperforms the traditional method SVM prominently. While graph-based methods, such as GCNN, DGCNN, GCB-net+BLS, Residual GCB-Net+BLS,

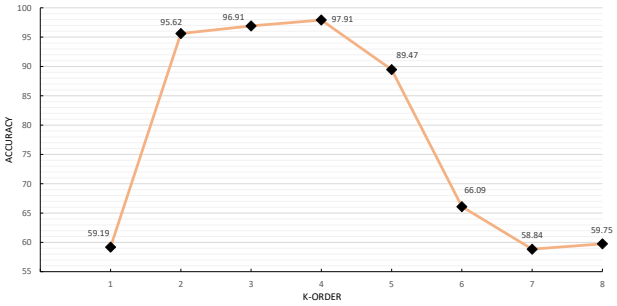

Fig. 2. Comparisons of the accuracies (%) of subject-independent experiment on SEED dataset using different Chebyshev filter sizes.

V-IAG, and HD-GCN, achieved remarkable results, AEG outperforms all of them, benefiting from the removal of spurious connections through maximizing the GIB objective.

Parameter Analysis on Chebyshv Filter Size: The Chebyshv filter size K is an essential hyper-parameter that significantly impacts the performance of a graph-based model in EEG emotion recognition. The representation of a node is extracted from its K-order neighbor nodes [37]. Since the f_ϕ aims to capture the emotion-relevant connections, the K of f_ϕ is set to be 1 to directly model the connections between any two channels (nodes) in the EEG data. However, the K of f_ψ is unclear. An additional experiment is provided to analyze the results of different Chebyshv filter size K of f_ψ on the SEED dataset in the subject-dependent setting. The results with K= 1,2,$\cdots$,8 are shown in Fig. 2. The AEG achieves the best performance when K = 4. When K is greater than 5, the performance of AEG has a relatively noticeable decline, which may attribute to the influence of over-smoothing.

Results on Subject-dependent Experiments: TABLE II reports the subject-independent results on SEED obtained using AEG and various baseline methods, including SVM, kernel principal component analysis (KPCA), transfer component analysis (TCA), transductive parameter transfer (TPT), DGCNN, domain adversarial neural networks (DANN), BiDANN, BiDANN-S, R2G-STNN, V-IAG, and HD-GCN. Graph-based methods achieved significantly better results than some transfer learning methods. AEG achieves the best results among all of them. The superior performance of AEG can be attributed to the nature of the GIB objective. The GIB objective prevents overfitting and improves robustness by encouraging the representation to be maximally informative and compact.

4) Experiment on SEED-IV: In the subject-dependent setting on SEED-IV, the training set comprises the first 16 trials of EEG. The remaining eight trials, including two trials per emotion class, are considered the testing set. The LOSO cross-validation strategy is also applied to the subject-independent setting on SEED-IV. The mean accuracy is calculated using all three sessions for both settings above.

Results on SEED-IV: To further evaluate the advantages of AEG, TABLE III reports a fair comparison with the state-of-the-art methods results including SVM, BiDANN, DGCNN, attention long short-term memory network (ALSTM), BiHDM w/o DA, RGNN w/o DA, and graph-based multi-task self-supervised model (GMSS) on SEED-IV dataset. AEG performs significantly better than other methods for subject-dependent (left) settings. The SEED-IV dataset contains four categories of emotion labels and is much more complex than SEED to extract emotional features. Thus the excellent performance of AEG indicates that the GIB objective is beneficial for multi-emotion recognition. AEG is a simple yet powerful implementation of GIB without access to the target data, i.e., AEG is trained without utilizing the unlabeled testing data. AEG outperforms other domain adaptation-free methods, namely BiHDM w/o DA and RGNN w/o DA, with a 4.69% and 0.51% improvement in accuracy, respectively. This result indicates that AEG can extract more general data representations for different subjects. AEG achieves competitive results compared to GMSS. It is worth noting that the performance of GMSS is attributed to its multi-task framework and self-supervised learning tasks, which are compatible with AEG. As a graph-based model, AEG has proved its ability to construct a robust brain dependency with higher accuracy than most graph-based methods.

5) Experiment on DREAMER: For DREAMER, the subject-dependent leave-one-trial-out cross-validation strategy is adopted. Specifically, one of the 18 trials is chosen for each subject as the testing set, while the remaining trials serve as the

TABLE III
COMPARISONS OF AEG AND STRONG BASELINES ON SEED-IV DATASET.

Method	Dependent ACC / STD (%)	Independent ACC / STD (%)
SVM [20]	56.61 / 20.05	37.99 / 12.52
BiDANN [52]	63.07 / 12.66	47.59 / 10.01
DGCNN [9]	69.88 / 16.29	52.82 / 09.23
A-LSTM [55]	69.50 / 15.65	55.03 / 09.28
BiHDM w/o DA [43]	72.22 / 14.69	67.47 / 08.22
RGNN w/o DA [14]	-	71.65 / 09.34
GMSS [10]	86.52 / 06.22	**73.48 / 07.41**
AEG (ours)	**91.11 / 11.21**	72.16 / 08.04

TABLE IV
COMPARISONS OF AEG AND STRONG BASELINES ON DREAMER DATASET.

Method	Valence ACC / STD	Arousal ACC / STD	Dominance ACC / STD
SVM [20]	60.14 / 33.34	68.84 / 24.94	75.84 / 20.76
GraphSLDA [56]	57.70 / 13.89	68.12 / 17.53	73.90 / 15.85
GSCCA [43]	56.65 / 21.50	70.30 / 18.66	77.31 / 15.44
DGCNN [9]	86.23 / 12.29	84.54 / 10.18	85.02 / 10.25
GCB-net [15]	86.99 / 6.21	89.32 / 5.01	89.20 / 04.33
Residual GCB-net [16]	87.43 / 14.89	91.55 / 14.78	89.37 / 16.78
VIAG [12]	92.82 / -	93.09 / -	- / -
HD-GCN [13]	93.95 / -	94.64 / -	- / -
AEG (ours)	**98.49 / 05.56**	**97.26 / 11.51**	**97.25 / 11.36**

Algorithm 2 Procedure of the CACA

Input: GNN model $f(\cdot)$ with order K; the prediction Y^c; input graph $G = (V, E)$; the t-th activation of the last graph convolutional layer $\mathcal{A}^t$; a sparsity constraint λ

Output: explanation $G_e = G[S^*]$

```
Compute L^c_{GraphGrad-CAM} following Equation (19) and (20)
Find the convergence channels: V_crg following Equation (16), (17), and (18)
Initialize the explanation: S = V_crg;
Compute scores: s_now = [(f(G[S])]_c − [f(G[V \ S])]_c
for channel τ ∈ V_crg do
    Find τ's K-hop neighbors set: N^K(v)
    for channel ξ ∈ N^K(v) ∩ ξ ∉ S do
        replace τ ∈ V_crg with ξ: S' = S ⋃{ξ} \ {τ}
        Compute new score:
        s_new = [(f(G[S'])]_c − [f(G[V \ S'])]_c
        if s_now ≤ s_new then
            s_now = s_new
            S = S'
        end if
    end for
end for
return S* = S
```

training set. This process is repeated until each trial has been used as the testing set once. The overall classification accuracy for each subject is then obtained by averaging the recognition accuracies of all 18 experiments. Finally, the average emotion classification accuracy is calculated by averaging the results across all 23 subjects.

Results on Subject-dependent Experiments: AEG is further compared with several baseline methods on DREAMER using the same PSD features as provided in [15]. The baseline methods on DREAMER include SVM, GraphSLDA, GSCCA, DGCNN, GCB-net, Residual GCB-net, VIAG, and HD-GCN. The mean accuracies for each emotion dimension (i.e., valence, arousal, and dominance) are reported in TABLE IV. AEG achieves the highest accuracies among those methods in all three dimensions, which validates the effectiveness of the proposed method.

D. Channel-wise Adaptive Class Activation Mapping Explainer

Apart from removing spurious spatial connectivities, another challenge is capturing the rationale behind the prediction. (i.e., which channel contributes most to the emotion prediction.) This paper offers a solution by introducing a novel post-hoc explainer named Channel-wise Adaptive Class Activation Mapping Explainer (CACA). Illustrated in Fig 1, CACA approaches the GNN explanation task (specifically, identifying the most decision-relevant EEG channels in this paper) as a subgraph scoring problem, where subgraphs are assessed to discover the optimal channel-induced subgraph that maximizes the MI.

Objective Function: Following the idea in [23], [24], [57], [58], CACA formalizes the explanation problem of GNNs through channels importance scoring. The target pre-trained GNN model is denoted as f_ρ. $0 < \lambda < 1$ denotes a sparsity constraint to enforce a concise number of channels. The goal of finding the most decision-relevant channels (nodes) is to find a channel-induced subgraph that maximizes a given evaluation metric $\mathrm{EVAL}(\cdot,\cdot,\cdot)$. This metric quantifies the faithfulness of the explanation G_e, which consists of the prediction-related channels S^*:

$$S^* = \underset{S\subseteq V, |S|\leq \lambda|V|}{argmax} \mathrm{EVAL}(f_\rho, G, G[S]), \tag{11}$$

where $G[S]$ denotes the induced subgraph of channel subset S in G. Drawing upon the established approach in prior research [38], [39], CACA considers the MI between underlying channels and the model's prediction as the $\mathrm{EVAL}(\cdot,\cdot,\cdot)$ function:

$$\begin{aligned} \mathrm{EVAL}(f_\rho, G, G[S]) &= I(f_\rho(G); G[S]) && (12)\\ &= H(Y) - H(Y|G = G[S]). && (13) \end{aligned}$$

MI quantifies the change in the probability of GNN's prediction when the input graph is limited to explanation $G[S]$ [38]. Consequently, optimizing Equation (13) is equivalent to maximizing the probability of GNN's prediction when the input graph is limited to $G[S]$, which can be expressed as follows:

$$G_e = G[S^*] \triangleq \underset{G[S]}{argmax} [f(G[S])]_c. \tag{14}$$

However, there may be multiple succinct channels that can reasonably explain the exact prediction [59]. A constraint is added to ensure the explanation is necessary. The adjusted objective function is:

$$G_e = G[S^*] \triangleq \underset{G[S]}{argmax} [(f(G[S])]_c - [f(G[V \setminus S])]_c, \tag{15}$$

the $V \setminus S$ represents the channel-subset obtained by eliminating possible channels from the input graph's channel set. The term $[f(G[V \setminus S])]_c$ evaluates if the complement channels of the explanation is adequate for the prediction. Thus, maximizing Equation (15) guarantees that the explanations are both necessary and sufficient for the target prediction.

Spatial Information of GNNs: The explainer should not depend on predefined rules from the domain knowledge about label Y. The underlying channels S^* can be any subset of V. The most direct approach to finding S^* is to evaluate every possible subset of G using Equation (15). However, this simple way is still time-consuming, and the time complexity is $\mathcal{O}(n^2)$. CACA provides a faster solution to this problem by leveraging the spatial information of GNNs, including activation and prediction probabilities. The spatial information of GNNs is employed to identify the Convergence channel. The convergence channel, $v_{\pi i} \in V_{crg}$, is the top channel ranked by the importance distribution provided by the spatial information. We formally define the set of convergence channels, V_{crg}, for GNNs as follows:

$$V_{crg} = \{v_{\pi 1}, v_{\pi 2}, \cdots, v_{\pi k}\}, \tag{16}$$

$$k = \lfloor \lambda |V| \rfloor, \tag{17}$$

$$v_{\pi i} = \underset{\substack{v_{\pi i}\in V \\ v_{\pi i}\notin\{v_{\pi 1}, v_{\pi 2}, \cdots, v_{\pi(i-1)}\}}}{argmax} \mathcal{L}^c_{\text{GraphGrad-CAM}}(v_{\pi i}), \tag{18}$$

where the $\mathcal{L}^c_{\text{GraphGrad-CAM}}$ represents importance distribution of each channel in G, as defined in Equation (19). The importance of channel $v_{\pi i}$ after the graph convolution operation is denoted by $\mathcal{L}^c_{\text{GraphGrad-CAM}}(v_{\pi i})$. The CACA uses the top $\lfloor \lambda |V| \rfloor$ channels as the convergence channels after sorting channels in descending order according to their importance. Then, CACA modifies the Grad-CAM approach by utilizing the gradient information of the last graph convolutional layer:

$$\mathcal{L}^c_{\text{GraphGrad-CAM}} = Relu(\sum_{t=0}^{k} \alpha_t^c \mathcal{A}^t), \tag{19}$$

$$\alpha_t^c = \frac{1}{|V|} \sum_{i=0}^{|V|} \frac{\partial Y^c}{\partial \mathcal{A}_i^t}, \tag{20}$$

where $\mathcal{L}^c_{\text{GraphGrad-CAM}}$ denotes the importance distribution for prediction Y^c and c is the target class. Additionally, let $\mathcal{A}^t$ represent the t-th activation map of the final graph convolutional layer, and $\mathcal{A}_i$ denote the value of channel i in $\mathcal{A}$. $\alpha^c t$ quantifies the 'importance' of the feature map $\mathcal{A}^t$ with respect to a specific target class c. In spectral Graph Neural Networks (GNNs), the learned channel representation is derived from its K-order neighboring channels, where K signifies the order of polynomials. Consequently, the channels in G_e should be any subset of V_{crg} and their K-hop neighboring channels. Upon identifying the top $\lfloor \lambda |V| \rfloor$ channels in $\mathcal{L}^c_{\text{GraphGrad-CAM}}$ as the convergence channels, CACA proceeds to explore explanation-possible subgraphs by substituting each channel in the convergence channels with one of its K-hop neighbor channels.

The algorithm of CACA: The channels that carry important predictive information are located in the K-hop neighborhood of the convergence channels. A natural approach would be to test every induced subgraph using a subset of convergence channels and their K-hop neighboring channels. However, this approach would have high time complexity, making it impractical. Therefore, CACA adopts a more stringent set of subgraphs to find the explanations. Since each convergence channel gathers significant predictive information from its K-hop neighboring channels, CACA replaces this channel with one of its K-hop neighboring channels to evaluate the explanation's extent. The objective is to determine if the subgraph induced from this replacement increases the value of Equation (15). CACA selects the channels that maximizes Equation (15) as an explanation. Due to the complexity of GNNs, the explanation found by CACA is usually a local optimum; however, experimental results show that this simple procedure often leads to high-quality explanations. The details of the CACA algorithm are provided in Algorithm 2. Next, a fair comparison with convincing metrics is presented to demonstrate the superiority of the proposed CACA.

E. Quantitative Studies of Faithfulness on CACA

A compelling post-hoc explainer should provide a faithful explanation of the model. Consistent with prior research [23]–[25], two quantitative metrics are employed, namely, Fidelity and Inverse Fidelity (Inv-Fidelity), to evaluate the faithfulness of the explanations. Fidelity and Inv-Fidelity assess whether the explanation is consistent with the model's prediction, either by removing important structures or retaining only the important ones. The formulas for these metrics are given below:

$$\text{Fidelity}(\text{G}, \text{G}_\text{e}) = [f(G)]_c - [f(G \setminus G_e)]_c, \tag{21}$$

$$\text{Inv-Fidelity}(\text{G}, \text{G}_\text{e}) = [f(G)]_c - [f(G_e)]_c. \tag{22}$$

Fidelity and Inv-Fidelity are complementary and are both critical for a good explanation. Since they are analogous to precision and recall, previous studies [23], [59] propose to use a single-scalar-metric "harmonic fidelity" (H-Fidelity), where Fidelity and Inv-Fidelity are normalized by sparsity and take their harmonic mean. Sparsity measures the extent of sparsity in the explanations by calculating the fraction of explanations, which is the same for each method. The better explanation should have a higher H-Fidelity. Formally, H-Fidelity is calculated by:

$$m1 = ([f(G)]_c - [f(G \setminus G_e)]_c) \times (1 - \frac{|G_e|}{G}), \tag{23}$$

$$m2 = ([f(G)]_c - [f(G_e)]_c) \times (1 - \frac{|G_e|}{G}), \tag{24}$$

$$\text{H-Fidelity}(\text{G}, \text{G}_\text{e}) = \frac{(1 + m1) \times (1 - m2)}{(2 + m1 - m2)}. \tag{25}$$

Datasets and Baselines: To evaluate the effectiveness of CACA, the SEED and SEED-IV are used in the experiment, known for their complex yet widely used EEG channels system. The baseline methods include activity maps [11], [43], heat maps [14], [44], and degree centrality of learned adjacency matrix [12], [13]. All of them are commonly applied in the EEG-based emotion recognition task. For activity maps of EEG electrodes, GraphGrad-CAM, shown in Equation (19), is utilized to identify important channels, as it is model-agnostic. For heat maps, the diagonal values of the learned adjacency matrix are used to determine the channel importance. The degree centrality $\mathcal{C}$ [60] assesses the connectivity of a channel with other channels. The degree centrality of the i-th channel, i.e., the i-th EEG channel, is calculated as follows:

$$\mathcal{C}_i = \sum_{j=1}^{n} A_{i,j} + \sum_{m=1}^{n} A_{m,i} - 2A_{i,i} (i = 1, \cdots, n), \tag{26}$$

where the n is the number of channels in the EEG dataset, which is 62 for both SEED and SEED-IV. Firstly, the AEG model is trained to converge on both SEED and SEED-IV datasets under the subject-dependent setting, following the same procedure described in Section III-C. Next, the GraphGrad-CAM, diagonal values (DV), degree centrality (DC), and CACA methods are utilized to identify the most prediction-important channels, which correspond to channels with higher values of channel importance. Subsequently, all three evaluation metrics based on the explainer's explanations are computed. Notably, only the test sets of SEED and SEED-IV datasets are used for studying the explanations. To enable comprehensive comparisons, the number of prediction-important channels is set to 15, 31, and 46.

TABLE V
COMPARISONS OF CACA AND STRONG BASELINES ON SEED AND SEED-IV DATASETS.

Dataset	Method	15 channels			31 channels			46 channels		
		Fidelity	Inv-Fidelity	H-Fidelity	Fidelity	Inv-Fidelity	H-Fidelity	Fidelity	Inv-Fidelity	H-Fidelity
SEED	DV [14], [44]	36.04	53.40	50.81	40.63	47.30	45.51	46.43	39.20	41.30
	DC [12], [13]	40.95	51.22	51.55	47.87	44.01	46.58	52.07	37.49	41.91
	GraphGrad-CAM [11], [43]	31.65	64.11	49.43	29.92	60.84	42.30	28.83	57.38	35.20
	CACA (ours)	**94.26**	**-04.93**	**63.54**	**94.35**	**-05.19**	**60.33**	**94.38**	**-05.31**	**56.48**
SEED-IV	DV [14], [44]	31.32	42.26	50.87	35.39	38.87	46.37	39.16	35.83	42.01
	DC [12], [13]	35.24	37.00	51.72	37.90	33.69	47.17	39.38	32.98	42.25
	GraphGrad-CAM [11], [43]	27.75	46.30	50.20	27.04	44.58	44.84	24.54	41.38	39.80
	CACA (ours)	**65.76**	**-17.81**	**60.89**	**66.42**	**-18.52**	**59.44**	**66.23**	**-18.56**	**57.19**

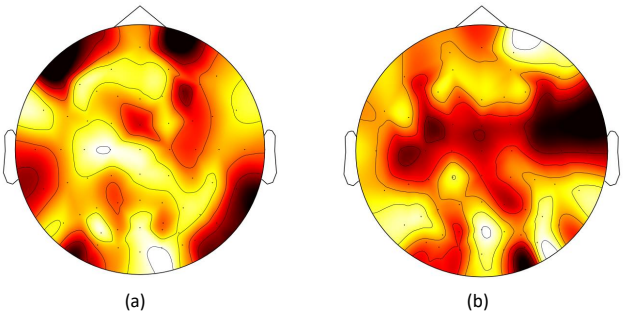

Fig. 3. The visualization of explanation provided by the CACA for incorrect predictions on SEED and SEED-IV datasets with subject-dependent settings, separately. The explanation for (a) the incorrect predictions on SEED; (b) the incorrect predictions on SEED-IV.

Evaluation Results: Table V shows the experimental results of the CACA and other baselines. From Table V, there are two major observations:

- The proposed CACA method outperforms other methods in three evaluation metrics for 15, 31, and 46 channels. While the DC method, as the most widely used post-hoc explainer for EEG-based emotion recognition in recent research, achieves the second-best performance, it fails to consider the effect of multi-channel combinations. Unlike the DC method, CACA searches for the optimal combination of channels that maximizes MI, leading to its outstanding performance.
- The proposed CACA method identifies a better set of channels that contribute more to the prediction. In evaluating the Inverse-Fidelity (Inv-Fidelity), the results of CACA are less than 0, indicating that the channels not in the explanation likely contribute nothing but noise to the emotion recognition. As a possible direction for future work, achieving better EEG data collection with fewer channels using CACA is suggested.

IV. APPLICATION

This section provides two possible applications of the AEG and CACA. Notice that the number of essential channels is 15 for all examples.

A. Model Debugging for AEG

This section demonstrates how to utilize the explanations generated by CACA for model analysis. The CACA excels at identifying the channels that contribute the most to the emotion prediction, even when the classification performance is poor. As a result, the CACA can aid in diagnosing why the trained GNN makes an incorrect prediction. Namely, for an incorrect prediction, if the explanation of trained GNN does not match the domain knowledge, the GNN may unexpectedly process the data.

Two empirical examples are provided in Fig. 3, where the trained GNNs used are the AEG with subject-dependent settings on SEED and SEED-IV datasets, respectively. Previous research [11], [61], [62] reports that the prefrontal, the frontal, the left temporal, the right temporal, and the midline area of the central lobes are the brain regions that related to the emotion generation. For Fig. 3 (a), it illustrates the channel importance distribution of the incorrect predictions in the SEED dataset. As shown in Fig. 3 (a), the poor performance of AEG might be attribute to the following reason: the trained AEG model fails to focus on the midline area of the central lobe as the critical region or can not extract discriminative representation from left temporal lobe. Furthemore, A similar analysis applies to Fig. 3 (b). The CACA explains the incorrect prediction on SEED-IV by pointing out that EEG data in the left temporal, prefrontal, and frontal lobes are ignored, suggesting a possible way to improve the classification performance by changing the model's focus.

B. Activation Regions for Differet Emotions

Considering the remarkable emotion classification performance of AEG and the ability of CACA to capture the essential channels, investigating the activation regions holds promise. For example, when considering neutral emotion, the possible activation region is located in the midline area of the central and frontal lobes. The channels of other emotions are provided in Table VI. It should be noted that the results in Table VI are mainly limited by the number of subjects and the quantity of EEG signals, which is left for future work.

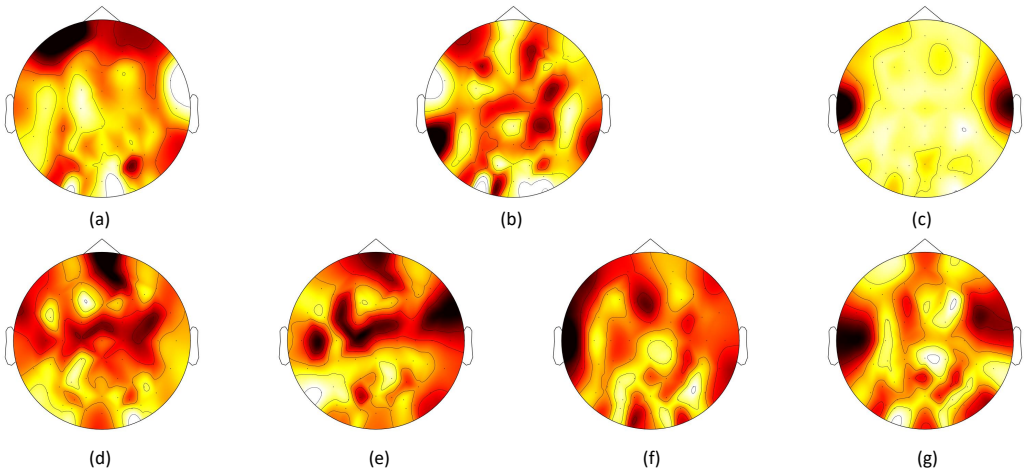

Fig. 4. The visualization of explanation provided by the CACA for different emotions on SEED and SEED-IV datasets with subject-dependent settings, separately. The explanation for (a) the neutral emotion on SEED; (b) the negative emotion on SEED; (c) the positive emotion on SEED; (d) the neutral emotion on SEED-IV; (e) the sad emotion on SEED-IV; (f) the fear emotion on SEED-IV; (g) the happy emotion on SEED-IV.

TABLE VI
THE 15 IMPORTANT CHANNELS FOR DIFFERENT EMOTION ON SEED AND SEED-IV DATASETS.

Dataset	Emotion	Important Channels
SEED	Neutral	Fp1,F7,PO6,Fp2,F5,Fpz,AF4,F8, PO7,TP8,PO5,F6,CB2,FC6,P8
	Negative	TP7,FC4,AF4,CP2,F3,CP4,O1,F2, C1,F7,Fp2,Cz,TP8,FC6,PO6
	Positive	T7,T8,FT8,FT7,POz,C6,C5,F7, CB1,FC5,FC6,AF4,Fp1,F2,TP8
SEED-IV	Neutral	Fpz,AF4,FC6,FCz,C4,C1,FC4,C5, Fp2,C2,FC2,FT7,CP4,CP6,T7
	Sad	FT8,C5,C1,FC6,FC3,FCz,Fpz,FC2, F3,AF4,PO3,T8,Cz,Fp1,FC4
	Fear	T7,Fz,FT7,F7,F1,O1,TP7,FC4, CP4,TP8,Fp1,AF3,F8,FCz,T8
	Happy	T7,FC6,C5,FT8,T8,FT7,P4,F1, Pz,C6,TP7,PO4,FC1,PO7,F6

V. CONCLUSION

This paper focuses on the interpretability of GNNs applied in EEG-based emotion recognition and proposes two novel methods to enhance the inherent interpretability and capture the underlying channels of predictions separately. In contrast to previous research that uses dynamic matrices for channel connections, the proposed AEG addresses the emotion-relevant connection recognition problem to infer a maximally informative yet compressed edge matrix achieved through the GIB objective. Furthermore, a MI estimator for EEG data and a sparsity loss are employed to implement the learning process. The emotion recognition results on three widely used EEG emotion datasets validate the superior properties of AEG. Additionally, this paper proposes CACA, a framework for generating compact and faithful explanations for GNNs. CACA utilizes the MI between the explanation and the model's prediction as the weight for each combination of various channels, resulting in a more reasonable channel importance distribution. Compared to many post-hoc explainers previously used in this area, the explanations from CACA are significantly more accurate, as empirically proven by extensive evaluation metrics. Besides, two practical applications are provided, including model debugging and investigating activation regions, demonstrating the enormous value of the proposed methods.

ACKNOWLEDGMENTS

This work was funded in part by the National Key Research and Development Program of China under number 2019YFA0706200, in part by the National Natural Science Foundation of China grant under number 62222603, 62076102, and 92267203, in part by the Guangdong Natural Science Funds for Distinguished Young Scholar under number 2020B1515020041, and in part by the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2019ZT08X214).

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[2] M. Ali, A. H. Mosa, F. Al Machot, and K. Kyamakya, "Eeg-based emotion recognition approach for e-healthcare applications," in *2016 eighth international conference on ubiquitous and future networks (ICUFN)*. IEEE, 2016, pp. 946–950.

[3] M. Leo, M. Del Coco, P. Carcagni, C. Distante, M. Bernava, G. Pioggia, and G. Palestra, "Automatic emotion recognition in robot-children interaction for asd treatment," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 145–153.

[4] J. S. K. Ooi, S. A. Ahmad, Y. Z. Chong, S. H. M. Ali, G. Ai, and H. Wagatsuma, "Driver emotion recognition framework based on electrodermal activity measurements during simulated driving conditions," in *2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. IEEE, 2016, pp. 365–369.

[5] X. Li, D. Song, P. Zhang, Y. Zhang, Y. Hou, and B. Hu, "Exploring eeg features in cross-subject emotion recognition," *Frontiers in neuroscience*, vol. 12, p. 162, 2018.

[6] H. Kober, L. F. Barrett, J. Joseph, E. Bliss-Moreau, K. Lindquist, and T. D. Wager, "Functional grouping and cortical–subcortical interactions in emotion: a meta-analysis of neuroimaging studies," *Neuroimage*, vol. 42, no. 2, pp. 998–1031, 2008.

[7] M. J. Kim, R. A. Loucks, A. L. Palmer, A. C. Brown, K. M. Solomon, A. N. Marchante, and P. J. Whalen, "The structural and functional connectivity of the amygdala: from normal emotion to pathological anxiety," *Behavioural brain research*, vol. 223, no. 2, pp. 403–410, 2011.

[8] K. A. Lindquist, T. D. Wager, H. Kober, E. Bliss-Moreau, and L. F. Barrett, "The brain basis of emotion: a meta-analytic review," *Behavioral and brain sciences*, vol. 35, no. 3, pp. 121–143, 2012.

[9] T. Song, W. 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.

[10] Y. Li, J. Chen, F. Li, B. Fu, H. Wu, Y. Ji, Y. Zhou, Y. Niu, G. Shi, and W. Zheng, "Gmss: Graph-based multi-task self-supervised learning for eeg emotion recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2022.

[11] B. Liu, J. Guo, C. L. P. Chen, X. Wu, and T. Zhang, "Fine-grained interpretability for eeg emotion recognition: Concat-aided grad-cam and systematic brain functional network," *IEEE Transactions on Affective Computing*, pp. 1–14, 2023.

[12] T. Song, S. Liu, W. Zheng, Y. Zong, Z. Cui, Y. Li, and X. Zhou, "Variational instance-adaptive graph for eeg emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 343–356, 2023.

[13] M. Ye, C. L. P. Chen, and T. Zhang, "Hierarchical dynamic graph convolutional network with interpretability for eeg-based emotion recognition," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2022.

[14] P. Zhong, D. Wang, and C. Miao, "Eeg-based emotion recognition using regularized graph neural networks," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1290–1301, 2022.

[15] T. Zhang, X. Wang, X. Xu, and C. L. P. Chen, "Gcb-net: Graph convolutional broad network and its application in emotion recognition," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 379–388, 2022.

[16] Q. Li, T. Zhang, C. L. P. Chen, K. Yi, and L. Chen, "Residual gcb-net: Residual graph convolutional broad network on emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems*, pp. 1–1, 2022.

[17] S. Miao, M. Liu, and P. Li, "Interpretable and generalizable graph learning via stochastic attention mechanism," in *International Conference on Machine Learning*. PMLR, 2022, pp. 15 524–15 543.

[18] J. Yu, T. Xu, Y. Rong, Y. Bian, J. Huang, and R. He, "Graph information bottleneck for subgraph recognition," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=bM4Iqfg8M2k

[19] T. Wu, H. Ren, P. Li, and J. Leskovec, "Graph information bottleneck," *Advances in Neural Information Processing Systems*, vol. 33, pp. 20 437–20 448, 2020.

[20] 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.

[21] 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*, vol. 49, no. 3, pp. 1110–1122, 2019.

[22] S. Katsigiannis and N. Ramzan, "Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 98–107, 2018.

[23] S. Zhang, Y. Liu, N. Shah, and Y. Sun, "Gstarx: Explaining graph neural networks with structure-aware cooperative games," *Advances in Neural Information Processing Systems*, vol. 35, pp. 19 810–19 823, 2022.

[24] H. Yuan, H. Yu, J. Wang, K. Li, and S. Ji, "On explainability of graph neural networks via subgraph explorations," in *International conference on machine learning*. PMLR, 2021, pp. 12 241–12 252.

[25] H. Yuan, H. Yu, S. Gui, and S. Ji, "Explainability in graph neural networks: A taxonomic survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 5, pp. 5782–5799, 2022.

[26] R. M. Mehmood and H. J. Lee, "Eeg based emotion recognition from human brain using hjorth parameters and svm," *International Journal of Bio-Science and Bio-Technology*, vol. 7, no. 3, pp. 23–32, 2015.

[27] M. Li, H. Xu, X. Liu, and S. Lu, "Emotion recognition from multichannel eeg signals using k-nearest neighbor classification," *Technology and health care*, vol. 26, no. S1, pp. 509–519, 2018.

[28] W. Jiang, G. Liu, X. Zhao, and F. Yang, "Cross-subject emotion recognition with a decision tree classifier based on sequential backward selection," in *2019 11th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, vol. 1. IEEE, 2019, pp. 309–313.

[29] C. L. P. Chen and Z. Liu, "Broad learning system: An effective and efficient incremental learning system without the need for deep architecture," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 1, pp. 10–24, 2018.

[30] G. Zhang, M. Yu, Y.-J. Liu, G. Zhao, D. Zhang, and W. Zheng, "Sparsedgcnn: Recognizing emotion from multichannel eeg signals," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 537–548, 2023.

[31] T. Song, S. Liu, W. Zheng, Y. Zong, and Z. Cui, "Instance-adaptive graph for eeg emotion recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 03, 2020, pp. 2701–2708.

[32] F. R. Chung, *Spectral graph theory*. American Mathematical Soc., 1997, vol. 92.

[33] 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*, vol. 30, no. 3, pp. 83–98, 2013.

[34] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," *Advances in neural information processing systems*, vol. 29, 2016.

[35] M. He, Z. Wei, H. Xu *et al.*, "Bernnet: Learning arbitrary graph spectral filters via bernstein approximation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 239–14 251, 2021.

[36] F. M. Bianchi, D. Grattarola, L. Livi, and C. Alippi, "Graph neural networks with convolutional arma filters," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3496–3507, 2021.

[37] X. Wang and M. Zhang, "How powerful are spectral graph neural networks," in *International Conference on Machine Learning*. PMLR, 2022, pp. 23 341–23 362.

[38] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/d80b7040b773199015de6d3b4293c8ff-Paper.pdf

[39] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang, "Parameterized explainer for graph neural network," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 19 620–19 631. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/e37b08dd3015330dcbb5d6663667b8b8-Paper.pdf

[40] W. Lin, H. Lan, and B. Li, "Generative causal explanations for graph neural networks," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 6666–6679. [Online]. Available: https://proceedings.mlr.press/v139/lin21d.html

[41] W. Lin, H. Lan, H. Wang, and B. Li, "Orphicx: A causality-inspired latent variable model for interpreting graph neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 13 729–13 738.

[42] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann, "Explainability methods for graph convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 772–10 781.

[43] Y. Li, L. Wang, W. Zheng, Y. Zong, L. Qi, Z. Cui, T. Zhang, and T. Song, "A novel bi-hemispheric discrepancy model for eeg emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 2, pp. 354–367, 2021.

[44] M. Jin, E. Zhu, C. Du, H. He, and J. Li, "Pgcn: Pyramidal graph convolutional network for eeg emotion recognition," *arXiv preprint arXiv:2302.02520*, 2023.

[45] Y. Wu, X. Wang, A. Zhang, X. He, and T.-S. Chua, "Discovering invariant rationales for graph neural networks," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=hGXij5rfiHw

[46] S. W. Smith *et al.*, "The scientist and engineer's guide to digital signal processing," 1997.
[47] Y. Li, W. Zheng, L. Wang, Y. Zong, and Z. Cui, "From regional to global brain: A novel hierarchical spatial-temporal neural network model for eeg emotion recognition," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 568–578, 2022.
[48] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation*, vol. 10, no. 5, pp. 1299–1319, 1998.
[49] 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.
[50] 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*, 2014, pp. 357–366.
[51] 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.
[52] Y. Li, W. Zheng, Z. Cui, T. Zhang, and Y. Zong, "A novel neural network model based on cerebral hemispheric asymmetry for eeg emotion recognition." in *IJCAI*, 2018, pp. 1561–1567.
[53] 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*, vol. 12, no. 2, pp. 494–504, 2018.
[54] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for eeg-based emotion classification," in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, 2013, pp. 81–84.
[55] 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.
[56] Y. Li, W. Zheng, Z. Cui, and X. Zhou, "A novel graph regularized sparse linear discriminant analysis model for eeg emotion recognition," in *International conference on neural information processing*. Springer, 2016, pp. 175–182.
[57] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/0060ef47b12160b9198302ebdb144dcf-Paper.pdf
[58] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
[59] K. Amara, R. Ying, Z. Zhang, Z. Han, Y. Shan, U. Brandes, S. Schemm, and C. Zhang, "Graphframex: Towards systematic evaluation of explainability methods for graph neural networks," *arXiv preprint arXiv:2206.09677*, 2022.
[60] X. Zhang, G. Cheng, and Y. Qu, "Ontology summarization based on rdf sentence graph," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 707–716.
[61] L. Pessoa, "A network model of the emotional brain," *Trends in cognitive sciences*, vol. 21, no. 5, pp. 357–371, 2017.
[62] N. T. Markov, M. Ercsey-Ravasz, D. C. Van Essen, K. Knoblauch, Z. Toroczkai, and H. Kennedy, "Cortical high-density counterstream architectures," *Science*, vol. 342, no. 6158, p. 1238406, 2013.

Hua Yang received the B.S. degree in electronic information science and technology from Central South University, ChangSha, China, in 2021. He is currently pursuing the M.S. degree in computer technology, South China University of Technology, Guangzhou, China. His current research interests include affective computing, graph neural network, and interpretable artificial intelligence.

C. L. Philip Chen (S'88-M'88-SM'94-F'07) received the M.S. degree from the University of Michigan at Ann Arbor, Ann Arbor, MI, USA, in 1985 and the Ph.D. degree from the Purdue University in 1988, all in electrical and computer science. He is the Chair Professor and Dean of the College of Computer Science and Engineering, South China University of Technology. He is the former Dean of the Faculty of Science and Technology. He is a Fellow of IEEE, AAAS, IAPR, CAA, and HKIE; a member of Academia Europaea (AE) and European Academy of Sciences and Arts (EASA). He received IEEE Norbert Wiener Award in 2018 for his contribution in systems and cybernetics, and machine learnings. He is also a highly cited researcher by Clarivate Analytics in 2018, 2019, 2020, and 2021. He was the Editor-in-Chief of the IEEE Transactions on Cybernetics (2020-2021) after he completed his term as the Editor-in-Chief of the IEEE Transactions on Systems, Man, and Cybernetics: Systems (2014-2019), followed by serving as the IEEE Systems, Man, and Cybernetics Society President from 2012 to 2013. Currently, he serves as an deputy director of CAAI Transactions on AI, an Associate Editor of the IEEE Transactions on AI, IEEE Trans on SMC: Systems, and IEEE Transactions on Fuzzy Systems, an Associate Editor of China Sciences: Information Sciences. He received Macau FDCT Natural Science Award three times and a First-rank Guangdong Province Scientific and Technology Advancement Award in 2019. His current research interests include cybernetics, computational intelligence, and systems.

Bianna Chen received the B.S. degree in information engineering from Guangdong University of Technology, Guangzhou, China, in 2017, the M.S degree in electronic and information engineering from South China University of Technology, Guangzhou, China, in 2020. She is currently pursuing the Ph.D. degree in computer science and engineering from South China University of Technology, Guangzhou, China. Her research interests include Graph neural network and affective computing.

Tong Zhang (S'12-M'16) received the B.S. degree in software engineering from Sun Yat-sen University, at Guangzhou, China, in 2009, and the M.S. degree in applied mathematics from University of Macau, at Macau, China, in 2011, and the Ph.D. degree in software engineering from the University of Macau, at Macau, China in 2016. Dr. Zhang currently is a professor with the School of Computer Science and Engineering, South China University of Technology, China. His research interests include affective computing, evolutionary computation, neural network, and other machine learning techniques and their applications. Prof. Zhang is the Associate Editor of the IEEE Transactions on Affective Computing, IEEE Transactions on Computational Social Systems, and Journal of Intelligent Manufacturing. He has been working in publication matters for many IEEE conferences.