VSGT: Variational Spatial and Gaussian Temporal Graph Models For EEG-based Emotion Recognition

Chenyu Liu[1†], **Xinliang Zhou**[1†], **Jiaping Xiao**[2], **Zhengri Zhu**[3], **Liming Zhai**[4*], **Ziyu Jia**[5*] and **Yang Liu**[1]
[1]School of Computer Science and Engineering, Nanyang Technological University
[2]School of Mechanical and Aerospace Engineering, Nanyang Technological University
[3]School of Electronic and Information Engineering, Beijing Jiaotong University
[4]School of Computer Science, Central China Normal University
[5]Institute of Automation Chinese Academy of Sciences
{chenyu003, xinliang001, jiaping001}@e.ntu.edu.sg, zhengrizhu@bjtu.edu.cn, limingzhai@ccnu.edu.cn, jia.ziyu@outlook.com, yangliu@ntu.edu.sg

Abstract

Electroencephalogram (EEG), which directly reflects the emotional activity of the brain, has been increasingly utilized for emotion recognition. Most works exploit the spatial and temporal dependencies in EEG to learn emotional feature representations, but they still have two limitations to reach their full potential. First, prior knowledge is rarely used to capture the spatial dependency of brain regions. Second, the cross temporal dependency between consecutive time slices for different brain regions is ignored. To address these limitations, in this paper, we propose Variational Spatial and Gaussian Temporal (VSGT) graph models to investigate the spatial and temporal dependencies for EEG-based emotion recognition. The VSGT has two key components: Variational Spatial Encoder (VSE) and Gaussian Temporal Encoder (GTE). The VSE identifies the dynamic spatial dependency based on prior knowledge using the variational Bayesian method. Besides, the GTE exploits the conditional Gaussian graph transform that computes comprehensive temporal dependency between consecutive time slices. Finally, the VSGT utilizes a recurrent structure to calculate the spatial and temporal dependencies for all time slices. Extensive experiments show the superiority of VSGT over state-of-the-art methods on multiple EEG datasets.

1 Introduction

Emotion recognition aims to identify the emotional state of an individual as a way to study the effect of emotions on human cognition. Compared with the behavioral clues of emotions, such as body language [Coulson, 2004], voice [Zeng *et al.*, 2007], and expression [Zhu *et al.*, 2017], the electroencephalogram (EEG) is thought to be the outcome of synchronized neuronal activity and signal transmission that can

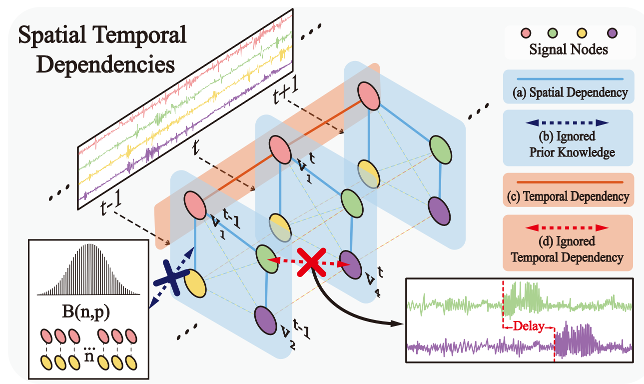

Figure 1: Spatial and temporal dependencies of emotional EEG. (a) Spatial dependency represents spatial correlations between different brain regions within the same time slice, *e.g.*, nodes v_1^{t-1} and v_2^{t-1} within the time slice $t-1$. (b) Existing methods do not consider prior knowledge, *e.g.*, the edge distribution between nodes. Temporal dependency includes two parts: (c) Temporal correlations of corresponding brain regions at different time slices, *e.g.*, nodes v_1^{t-1} and v_1^t at time slices $t-1$ and t; and (d) Temporal correlations of different brain regions across time slices, *e.g.*, nodes v_2^{t-1} and v_4^t. Such temporal correlations indicate the delay in the temporal dimension of the interactions of brain regions but are often ignored by existing methods.

directly reflect genuine emotion, which makes EEG widely used in emotion recognition. The EEG signals are collected from multiple electrodes placed over different brain regions and exhibit various structural characteristics, including spatial correlations among brain regions within a time slice and temporal correlations between time slices, as shown in Figure 1. These spatial and temporal dependencies are specifically critical for recognizing emotional states, and thus, exploiting such dependencies in EEG-based emotion recognition has recently become mainstream [Zhou *et al.*, 2023; Liu *et al.*, 2024].

Despite much progress made by existing emotion recognition methods [Jia *et al.*, 2021; Li *et al.*, 2018], there are still two deficiencies in terms of spatial and temporal dependencies that limit the performance of emotion recognition. First, prior knowledge is rarely used to capture the spatial dependency of brain regions. The prior knowledge refers to the established physiological paradigms that can depict the dynamics and stability of fundamental connectivity patterns of brain regions [Gonzalez-Castillo *et al.*, 2015; Braun *et al.*, 2015]. It contributes to the deep exploitation of emotional states. Unfortunately, very few emotion recognition methods take prior knowledge into account to capture the spatial dependency.

Second, the cross temporal dependency between consecutive time slices for different brain regions is ignored. Studies have shown that emotion persistence involves the asynchronous activity in brain regions, which represents the temporal delay in the response of other brain regions to the current brain region activity [Lewis, 2005] (*e.g.*, the correlation between v_2^{T-1} and v_4^T in Figure 1). Neglecting this cross temporal dependency may fail to describe the critical delayed response in brain regions for emotional states. However, for the temporal dependency between consecutive time slices, existing works [Feng *et al.*, 2022; Li *et al.*, 2019; Li *et al.*, 2021b] only consider the correlations in the same brain region, ignoring those between different brain regions.

To solve the above limitations, we draw inspiration from neuroscience to better guide the utilization of spatial and temporal dependencies and propose variational spatial and Gaussian temporal (VSGT) graph models for emotion recognition. We first design a spatial and temporal EEG-to-graph (S&T-E2G) block to transform the EEG data into a time-serial graph structure. To leverage the prior knowledge to facilitate the learning of emotion recognition models, we introduce the neural field theory to model the prior distribution of spatial dependency and present a variational spatial encoder (VSE) to compute a derivable posterior distribution to approximate the prior distribution, through which the spatial dependency can be captured with the variational Bayesian method. To capture the cross temporal dependency between consecutive time slices for different brain regions, we propose a Gaussian temporal encoder (GTE), in which a conditional Gaussian graph transform is used to calculate comprehensive temporal dependency between consecutive time slices. Finally, we use a graph convolutional network (GCN) to recurrently learn effective feature representations from the spatial and temporal dependencies for emotion recognition.

The contributions of the paper are summarized as follows.

- We introduce the prior knowledge *i.e.*, neural field theory, to the spatial dependency and propose a variational spatial encoder (VSE) to capture the spatial dependency.
- We analyze the importance of cross temporal dependency for emotion recognition and propose a Gaussian temporal encoder (GTE) to capture the cross temporal dependency.
- We designed a recurrent spatial and temporal module (RST) to learn the spatial-temporal feature representations for time-serial graph structure.

2 Related Work

Currently, more and more researchers are devoted to analyzing EEG-based emotion recognition based on spatial or spatio-temporal dependency in graph modeling. For **spatial dependency modeling**, existing methods are divided into predetermined and dynamic dependencies. Predetermined dependency indicates that spatial dependency is built on certain physiological paradigms. Based on 'the decay of brain region correlations with physical distance' [Salvador *et al.*, 2005], [Zhong *et al.*, 2020; Xu *et al.*, 2023] apply spatial coordinates of electrodes to calculate spatial dependency related to physical distance; Based on 'the International 10-20 EEG Electrode Standard', [Deng *et al.*, 2021] set spatial dependency based on fixed 2D electrode connection graph; Based on 'the asymmetry of neural activity in the left and right hemispheres during emotional states' [Schmidt and Trainor, 2001], [Du *et al.*, 2022; Ding *et al.*, 2024] select specific cross-hemisphere electrode pairs to set spatial dependency. Dynamic dependency indicates that the spatial dependency can be updated during training. [Song *et al.*, 2018; Zhang *et al.*, 2019] utilize a parameter matrix to represent spatial dependency, which can be updated through the loss function; [Li *et al.*, 2021a; Ye *et al.*, 2022] calculates the phase difference between electrode signals as the spatial dependency; [Zeng *et al.*, 2022; Jin *et al.*, 2021] apply an attention mechanism that computes the adjacency matrix with model parameters.

Although predetermined dependency conforms to prior physiological paradigms, implementing prior constraints results in their inability to be continually updated, thus limiting the model's adaptability to different data. Although dynamic dependency can be continuously updated, it lacks physiological plausibility since it is not constrained by prior knowledge.

For **temporal dependency modeling**, graph-based methods in EEG-based emotion recognition typically combine GNN and temporal encoders. [Feng *et al.*, 2022; Li *et al.*, 2019; Li *et al.*, 2021b] learn the spatial dependency within each time slice independently using GNN and then concatenated the feature representations of all time slices together to learn temporal dependencies between time slices via LSTM. Similarly, [Peng *et al.*, 2022; Ding and Guan, 2023] utilize this structure, with the difference being the use of Transformer encoders as temporal encoders. In addition, [Jia *et al.*, 2021; Liu *et al.*,] forms a series of graphs from spatial dependencies of all time slices and uses the Gate Recurrent Unit (GRU) to infer the temporal dependencies between time slices. Also, [Li *et al.*, 2021a] averages the spatial dependencies of all time slices as the final spatial dependency of the emotional EEG.

For the aforementioned methods, despite differences in the choice of the temporal encoder, all apply a two-step process that captures spatial dependency within time slices independently and then utilizes a temporal encoder to capture temporal dependency across time slices. However, this temporal dependency ignores the correlations between different brain regions across time slices, which is a significant part of the temporal dependency of emotional EEG. As a result, the temporal dependency that fails to describe the critical delayed re-

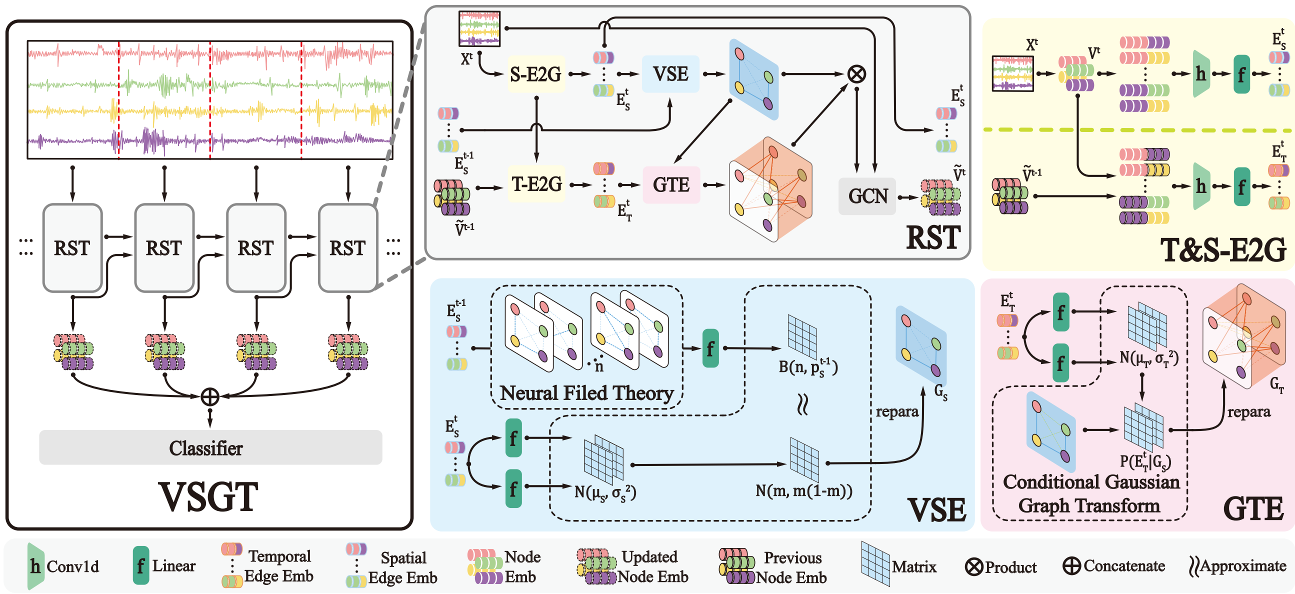

Figure 2: Overall structure of our VSGT graph model.

sponse of brain regions is incomplete, which in turn restricts the model performance.

3 Methodology

3.1 Overall Structure

The proposed model processes emotional EEG signals in time series and outputs a predicted emotion label. Given emotional EEG signal $X \in \mathbb{R}^{C \times N}$ with C the EEG channels and N the number of samples, the emotion recognition task is defined as learning a model $F(\cdot)$, which maps X into the corresponding emotion label $Y = F(X)$.

EEG signals are derived from different brain regions and exhibit dynamic changes in connectivity patterns between brain regions over time. According to such characteristics, we first model the EEG data as time-serial graphs, then capture the spatial and temporal dependencies in EEG graphs, and finally adopt a graph convolutional network (GCN) to extract feature representations for emotion recognition.

Figure 2 shows the overall structure of our variational spatial and Gaussian temporal graph models (VSGT), which consists of recurrent spatial and temporal modules (RSTs) and a classifier. Each RST contains a spatial&temporal EEG-to-graph (S&T-E2G) block, a variational spatial encoder (VSE), and a Gaussian temporal encoder (GTE).

S&T-E2G To process the EEG data in a graph space, we devise an S-E2G block and a T-E2G block to transform the EEG signals into a time-serial graph structure, in which the nodes represent the brain regions (*i.e.*, EEG channels) and the edges represent the connections between different brain regions. The resulting EEG graphs will be further processed by VSE and GTE to characterize dependencies. See Section 3.2 for the details of S&T-E2G.

VSE The VSE aims to capture the spatial dependency of an EEG graph in the current time slice by leveraging the prior knowledge of neural field theory. The reason we use neural field theory is as follows. The neural field theory defines the brain activity produced by emotional stimuli as a superposition of traveling waves generated by neurons. The traveling wave represents the propagation of the electrical signal from one neuron to the whole neural system, and such a propagation route measures the spatial dependency between that neuron and all the neurons connected to it. So, it is reasonable to integrate neural field theory into the modeling of dependency relations between graph nodes. The details of applying neural field theory to spatial dependency are in Section 3.3.

GTE The GTE tries to capture the temporal dependency between consecutive time slices. The temporal dependency not only contains the traditional temporal dependency in the same brain region but also contains the cross temporal dependency ignored by existing works. The cross temporal dependency is analyzed based on the response delay phenomenon in brain regions [Lewis, 2005]. See Section 3.4 for capturing the cross temporal dependency.

RST The spatial dependency and temporal dependency produced by VSE and GTE are merged and fed into a GCN to extract node embeddings. The VSE and GTE only process two consecutive time slices. To tackle the whole EEG data with T time slices, we use a recurrent structure, *i.e.*, RST, which updates all time slices sequentially and retains the information from the previous time slice to infer the temporal and spatial dependencies of the current time slice. This ensures that all spatial dependencies can follow prior knowledge and also allows modeling the correlations of different brain regions at different time slices. The RST and the final classifier constitute the VSGT graph model, whose optimization objective is detailed in Section 3.5.

3.2 Spatial & Temporal Node-to-Graph (S&T-E2G)

The S&T-E2G block can be considered a preprocessing for follow-up dependency analysis and feature extracting. The EEG data X are first cut into T time slices $\{X^t|t \in [1,T]\}$ (32 samples for each time slice). For the t^{th} time slice X^t, we employ the same setup as in [Jia *et al.*, 2021] to extract the node embedding $v_i^t \in \mathbb{R}^d$, where i is the node index and d is the embedding dimension. All the v_i^t in a graph constitute a node set V^t. Then we concatenate every two nodes and use Conv1d $\boldsymbol{h}(\cdot)$ and Linear layer $\boldsymbol{f}(\cdot)$ to generate edge embeddings as follows:

$$E_S^t = \sum_{i,j \in [1,C]} \boldsymbol{f}(\boldsymbol{h}(v_i^t, v_j^t)) \tag{1}$$

$$E_T^t = \sum_{i,j \in [1,C]} \boldsymbol{f}(\boldsymbol{h}(v_i^t, v_j^{t-1})) \tag{2}$$

where $\boldsymbol{h}(,\cdot,)$ denotes a Conv1d function, $\boldsymbol{f}$ denotes a linear layer, C is the node number (*i.e.*, channel number), and E_S^t and E_T^t are the sets of spatial edge embeddings and temporal edge embeddings for the t^{th} time slice. Note that the Conv1d and Linear layers in T-E2G and S-E2G are different.

3.3 Variational Spatial Encoder (VSE)

To better leverage the prior knowledge, the regularities of prior knowledge first need to be characterized. As discussed in Section 3.1, emotional brain activity is a superposition of traveling waves generated by innumerable neurons. According to Thermodynamic Limit, when the number of neurons tends to infinity, the coupling of an infinite number of traveling waves provides an infinite approximation to the actual spatial dependency of the brain. We define this approximation as a Binomial distribution:

$$\mathcal{P}(e_{i,j}) \sim \lim_{\substack{n \to \infty \\ p_{i,j} \to 0}} \mathcal{B}(n, p_{i,j}) \tag{3}$$

where $\mathcal{P}(e_{i,j})$ represents the overall edge distribution between two nodes v_i and v_j, $p_{i,j}$ denotes the probability of edge existence within each traveling wave, and $\mathcal{B}(\cdot)$ is the Binomial distribution. $n \to \infty$ corresponds to infinite neurons, and thus, the number of edges tends to infinity, which is utilized to simulate the infinite potential neuronal connections in the brain. $p_{i,j} \to 0$ indicates that single neural activity is physiologically slight and unobservable.

We assume the prior distribution as the distribution of the spatial dependency in the $t-1^{th}$ time slice:

$$\mathcal{P}(E_S^{t-1}) \sim \mathcal{B}(n, p_S^{t-1}) \tag{4}$$

where $p_S^{t-1} \in \mathbb{R}^{N_e}$ ($N_e = C(C-1)$ is the number of edges) is computed from E_S^{t-1} by the linear layer $\boldsymbol{f}$.

We also assume the distribution of spatial dependency in the current time slice matches such a prior distribution. Inspired by VRNN [Chung *et al.*, 2015], an RNN encoder is applied to estimate the parameters of the approximate posterior distribution. However, the posterior distribution of spatial dependency cannot be computed like [Chung *et al.*, 2015] due to the presence of infinite n. According to the De Moivre-Laplace Theorem, as n tends to infinity, the binomial distribution $\mathcal{B}(n, p_{i,j})$ can be approximated by a Gaussian distribution $\mathcal{N}(np_{i,j}, np_{i,j}(1-p_{i,j}))$. Based on this, we calculate a Gaussian distribution proxy to void the direct parameterization of both the infinite parameter n and the near-zero parameter $p_{i,j}$.

Given a Gaussian distribution $\mathcal{N}(\mu_S, {\sigma_S}^2)$, where $\mu_S \in \mathbb{R}^{N_e}$ and $\sigma_S \in \mathbb{R}^{N_e}$ are computed from E_S^t and letting $\mu_S < \frac{1}{2}$, we define the posterior Gaussian distribution as

$$\mathcal{P}(E_S^t) \sim \mathcal{N}(m, m(1-m)), \tag{5}$$

where m denotes the mean of the Gaussian approximation of the binomial distribution and is defined as

$$m = \frac{1 + 2\lambda {\sigma_S}^2 - \sqrt{1 + 4\lambda^2 {\sigma_S}^4}}{2}, \tag{6}$$

where λ is an approximation used to prevent explosions in the calculation, and it is defined as

$$\lambda = \frac{1}{1 - 2\mu_S} \sim \zeta(\mu_S) + \varepsilon, \tag{7}$$

where ε is a very small hyperparameter (Set to $1\mathbf{e}-2$) and $\zeta(\cdot)$ is a softplus function.

The $\mathcal{P}(E_S^t)$ is a Gaussian proxy to the Binomial distribution. Therefore, the $\mathcal{P}(E_S^t)$ approximates the prior distribution while also allowing the use of the reparameterization trick [Kingma and Welling, 2013], which allows this distribution to be derived. We draw samples from such a distribution as the spatial dependency of the current time slice:

$$\mathcal{G}_S = \sqrt{m(1-m)} \times \epsilon + m \tag{8}$$

where ϵ is a random variable sampled from Standard Normal distribution. $\mathcal{G}_S \in \mathbb{R}^{N_e}$ indicates the spatial dependency of current time slice.

3.4 Gaussian Temporal Encoder (GTE)

GTE infers the temporal dependency of the current time slice. It takes temporal edge embedding E_T^t and spatial dependency $\mathcal{G}_S$ of the current time slice as inputs and outputs the temporal dependency $\mathcal{G}_T$ of the current time slice. The spatial and temporal dependencies of the current time slice could be regarded as weighting each edge of spatial dependency with a Gaussian variable. We define such Gaussian variables as the temporal dependency between time slices t and $t-1$. We further assume such Gaussian variables to be conditioned on spatial dependency as:

$$\mathcal{P}(E_T^t|\mathcal{G}_S) = \mathcal{N}(\mathcal{G}_S \cdot \mu_T, \mathcal{G}_S \cdot {\sigma_T}^2) \tag{9}$$

We refer to the above operation as the conditional Gaussian graph transform. G_S is the spatial dependency of time slice t from Eq. (8). $\mathcal{N}(\mu_T, {\sigma_T}^2)$ is a Gaussian distribution generated from E_T^t. The purpose of conditional probability is that the delayed response between brain regions is based on the existence of interactions between brain regions. Therefore, we conditioned spatial dependency samples to indicate the presence of interactions between brain regions that are likely to form a delayed response. The temporal dependency of time slice t is defined as samples from such a conditional distribution:

$$\mathcal{G}_T = \sqrt{\mathcal{G}_S} \cdot \sigma_T \times \epsilon + \mathcal{G}_S \cdot \mu_T \tag{10}$$

where ϵ is a random variable sampled from Standard Normal distribution. $\mathcal{G}_T \in \mathbb{R}^{N_e}$ indicates the temporal dependency between time slices t and $t-1$.

3.5 Learning of VSGT

The loss function of VSGT contains two parts: the difference between the prediction and the ground truth labels and the Kullback-Leibler Divergence (KLD) between the prior and posterior distributions in VSE. The significance of the second part is to constrain the spatial dependency distribution of the current time slice by the spatial dependency distribution of the previous time slice that conforms to prior knowledge, which further ensures the stability of all spatial dependencies, *i.e.*, they all conform to the fundamental connectivity patterns. The loss function $\mathcal{L}$ is as follows:

$$\mathcal{L} = \mathrm{BCE}\{Y, \tilde{Y}\} + \boldsymbol{w} \cdot \sum_{t=1}^{T} \mathrm{KLD}\{\mathcal{P}(E_S^{t-1}) || \mathcal{P}(E_S^t)\} \tag{11}$$

where $\boldsymbol{w}$ is loss weight (Set to 0.1). While BCE$\{\cdot\}$ is easy to compute, the KLD part of $\mathcal{L}$ cannot be computed directly because the prior probability $\mathcal{P}(E_S^{t-1})$ is a Binomial distribution $\mathcal{B}(n, p_S^{t-1})$ with an infinite parameter n. Since minimizing the KLD part can also be regarded as minimizing the upper bound, the second part of $\mathcal{L}$ is defined as:

$$\begin{aligned} &\mathrm{KLD}\{\mathcal{P}(E_S^{t-1}) || \mathcal{P}(E_S^t)\} \\ =&\mathrm{KLD}\{\mathcal{B}(n, p_S^{t-1}) || \mathcal{N}(m, m(1-m))\} \\ <&m \log \frac{m+\varepsilon}{p_S^{t-1} + \epsilon} + (1-m) \log \frac{1-m+m^2/2+\varepsilon}{1-p_S^{t-1} + {p_S^{t-1}}^2/2+\varepsilon} \end{aligned} \tag{12}$$

4 Experiment

Datasets We examine our method on two different datasets: DEAP [Koelstra *et al.*, 2011] and DREAMER [Soleymani *et al.*, 2011]. They are both EEG signal datasets tailored for human emotion analysis. Subjects will watch videos and record their emotional responses characterized by valence (high/low) and arousal (high/low). They evaluate the valence and arousal on a scale from 1 to 9. We adopt the general EEG-based emotion recognition paradigm in which valence and arousal scores are divided into two categories bounded by 5. Separate experiments were conducted in each subject and averaged as the final result of the dataset. We present the result on value and arousal separately in our experiment results.

Training Paradigm We adopt the k-fold cross-validation paradigm to divide the training and testing sets. For each subject, the data is divided into 10 folds, where one fold is the testing set, and the remaining are used to train. The data used for training is divided into $8:1$, where one fold is the validation set, and the remaining is the training set.

Cropping Strategy To expand the data volume, we used a cropped strategy ([Schirrmeister *et al.*, 2017]), which crops the data of each trail into $4s$ non-overlapping segments. The cropping is done strictly after splitting the training, validation, and testing sets, which prevents data leakage. We make all baselines follow the same cropping strategy to prevent abnormally high accuracy and F1 scores.

Baseline and Hyperparameters We compared VSGT with three types of baselines. One category is the non-graph methods; another is the graph-based methods; and the last one is the methods combining GNN and temporal encoders. By comparing with these baselines, we hope to demonstrate that the VSGT can make full use of spatio-temporal dependencies to improve the accuracy of EEG-based emotion recognition. We set up a shared environment for all baselines to ensure consistency except for hyperparameters.

4.1 Comparison with Prior Art

The accuracy and F1 scores of the test results are shown in Table 1. Compared to non-graph methods, VSGT achieves a significant improvement in performance on two datasets. MM-ResLSTM [Ma *et al.*, 2019] and ACRNN [Tao *et al.*, 2020] utilize a temporal encoder to capture the temporal dependency of emotional EEG. TSception [Ding *et al.*, 2022] and SST-EmotionNet [Jia *et al.*, 2020] both apply convolution in the spatial and temporal dimensions of the signal to learn spatial and temporal dependencies. The combination of spatio-temporal dependencies allowed TSception and SST-EmotionNet to achieve better results. But essentially, the above baselines' temporal dependency inference process is still performed independently, which ignores the correlations between different brain regions across time slices and thus performs worse than VSGT.

Compared to graph-based methods, VSGT still demonstrated superior performance. RGNN [Zhong *et al.*, 2020] and LGGNet [Ding *et al.*, 2023] both introduce specific prior knowledge as the spatial dependency. The prior knowledge used by RGNN to define spatial dependency is that correlations between brain regions show regular decay with physical distance. Therefore, RGNN calculates spatial distances from the 3D coordinates of the electrodes and uses this as the spatial dependency. LGGNet relies on prior knowledge of existing correlations between functional brain regions and emotions. Connections between electrodes were predefined based on specific brain region divisions. However, since the prior knowledge is fixed, the spatial dependency cannot be adjusted during training. Instead of using prior knowledge, both DGCNN [Song *et al.*, 2018] and MTGNN [Wu *et al.*, 2020] parameterize the spatial dependency as part of the model, which is updated by loss. All of the above methods transform the time series signals into frequency-domain features and, therefore, do not include temporal dependency. Also, not implementing the use of prior knowledge to capture spatial dependency further limits their performance.

Compared to methods combining GNN and temporal encoders, VSGT benefits from better results due to more comprehensive spatial and temporal dependencies. HetEmotionNet [Jia *et al.*, 2021], ASTG-LSTM [Li *et al.*, 2021b], ST-GCLSTM [Feng *et al.*, 2022] and EmoGT [Jiang *et al.*, 2023]

Baseline	DEAP				DREAMER			
	Arousal		Valence		Arousal		Valence	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
MM-ResLSTM [Ma *et al.*, 2019]	70.84 ± 9.48	68.07 ± 10.81	70.89 ± 7.01	64.21 ± 7.12	68.43 ± 7.25	63.77 ± 7.31	71.34 ± 6.72	70.63 ± 5.39
ACRNN [Tao *et al.*, 2020]	69.71 ± 8.41	65.48 ± 8.04	68.82± 6.23	69.33 ± 7.09	70.24 ± 9.26	65.78 ± 9.54	65.23 ± 5.11	60.27 ± 6.99
TSception [Ding *et al.*, 2022]	74.38 ± 8.18	76.21 ± 9.84	68.94 ± 7.25	72.04 ± 6.94	69.77 ± 6.56	69.27 ± 9.43	68.16 ± 7.12	65.06 ± 6.71
SST-EmotionNet [Jia *et al.*, 2020]	73.49 ± 6.61	76.84 ± 10.72	70.94 ± 6.21	69.87 ± 6.99	71.54 ± 8.43	69.83 ± 8.71	70.39 ± 7.03	68.1 ± 5.27
DGCNN [Song *et al.*, 2018]	64.89 ± 8.73	66.17 ± 10.16	68.14 ± 6.72	72.89 ± 6.44	68.56 ± 8.51	67.85 ± 9.55	68 ± 6.51	63.95 ± 7.11
RGNN [Zhong *et al.*, 2020]	60.55 ± 5.31	62.49 ± 7.14	61.37 ± 6.93	60.75 ± 6.16	62.48 ± 5.03	61.07 ± 6.47	61.98 ± 5.87	61.11 ± 5.37
MTGNN [Wu *et al.*, 2020]	70.98 ± 7.37	68.45 ± 9.98	69.14 ± 6.12	68.97 ± 6.19	71.24 ± 7.26	69.87 ± 9.53	67.07 ± 5.93	68.07 ± 8.31
LGGNet [Ding *et al.*, 2023]	72.31 ± 8.38	73.28 ± 8.78	68.53 ± 7.79	69.15 ± 7.59	70.49 ± 8.84	68.94 ± 8.47	68.69 ± 6.08	66.12 ± 8.14
HetEmotionNet [Jia *et al.*, 2021]	74.36 ± 7.36	75.71 ± 10.82	72.08 ± 5.66	71.15 ± 6.1	70.58 ± 7.63	69.24 ± 4.92	70.03 ± 6.64	67.49 ± 6.38
ASTG-LSTM [Li *et al.*, 2021b]	72.85 ± 6.48	73.18 ± 11.01	71.96 ± 6.48	70.8 ± 7.26	70.23 ± 8.43	69.14 ± 5.34	68.76 ± 6.76	68.94 ± 5.49
ST-GCLSTM [Feng *et al.*, 2022]	76.68 ± 7.37	71.34 ± 9.85	71.81 ± 6.72	70.36 ± 9.14	69.29 ± 6.21	69.43 ± 5.93	70.49 ± 5.98	70.23 ± 5.84
EmoGT [Jiang *et al.*, 2023]	70.82 ± 6.94	70.14 ± 9.47	68.43 ± 7.16	70.63 ± 7.97	68.54 ± 7.32	67.39 ± 6.21	69.03 ± 5.86	68.71 ± 6.27
Our VSGT	**78.95 ± 6.89**	**78.65 ± 8.24**	**74.15 ± 6.41**	**74.05 ± 6.43**	**75.19 ± 6.87**	**72.98 ± 7.54**	**73.41 ± 6.36**	**72.73 ± 5.71**

Table 1: Performance comparison of all baselines in terms of accuracy and f1 score (in %) for classification tasks (the greater, the better) on DEAP and DREAMER. The best results are bolded.

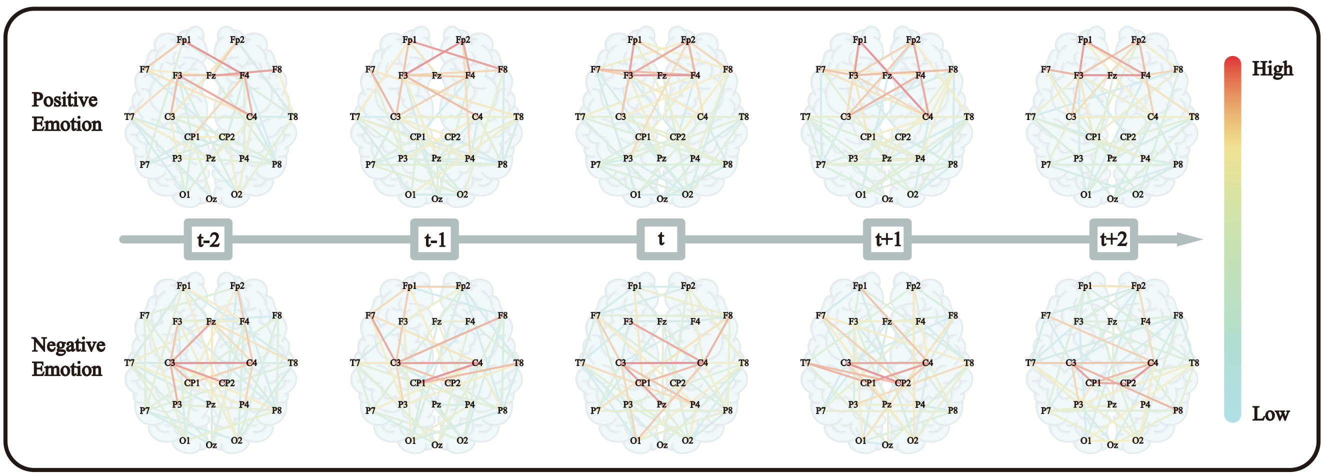

Figure 3: Visualization of spatial dependency. We visualized five time slices. For readability considerations, we only showed 50 edges between 21 electrodes. For positive emotion, high correlations were concentrated at electrodes $F3$, $F4$, $Fp1$, and $Fp2$. For negative emotion, high correlations were concentrated at electrodes $C3$, $C4$.

utilize the same spatial and temporal dependency inference process: model the spatial dependency within each time slice using GNNs, and apply a temporal encoder to infer the temporal dependency of all time slices. The difference is that HetEmotionNet uses Gated Recurrent Units, ASTG-LSTM and ST-GCLSTM use LSTM, and EmoGT uses a transformer encoder. Due to the introduction of temporal dependency and the construction of more comprehensive spatial dependency via GNNs, the above approach achieves better results than the other two types of approaches. However, the spatial and temporal dependencies inference approach they used was unable to establish correlations between different brain regions across time slices, hence the better results achieved by VSGT.

4.2 Interpreting Spatial Dependency

We further verified the interpretability of the spatial dependency inferred by VSGT on DEAP in Figure 3. For readability reasons, we only show part of the dependencies between 21 electrodes. The spatial dependency can be summarized as follows: For positive emotions, electrodes $F3$, $F4$, $Fp1$, and $Fp2$ showed high correlations at different time slices; for negative emotions, electrodes $C3$ and $C4$ showed high correlations at different time slices. This phenomenon is consistent with [Min *et al.*, 2022]'s conclusion that the brain's frontal lobe ($Fp1$, $Fp2$) and frontal cortex ($F3$, $F4$) show a high correlation under positive emotion; under negative emotion, the central sulcus ($C3$, $C4$) showed a high correlation. Therefore, the spatial dependency maintains stability (i.e., the high correlation region is the same for different time stamps) and also maintains dynamics (the high correlation shows a certain degree of fluctuation).

4.3 Interpreting Temporal Dependency

We visualized several sets of signals from different channels with high correlation in continuous time slices in Figure 4. In Figure 4 (a), Signal 2 has an amplitude decrease at the time

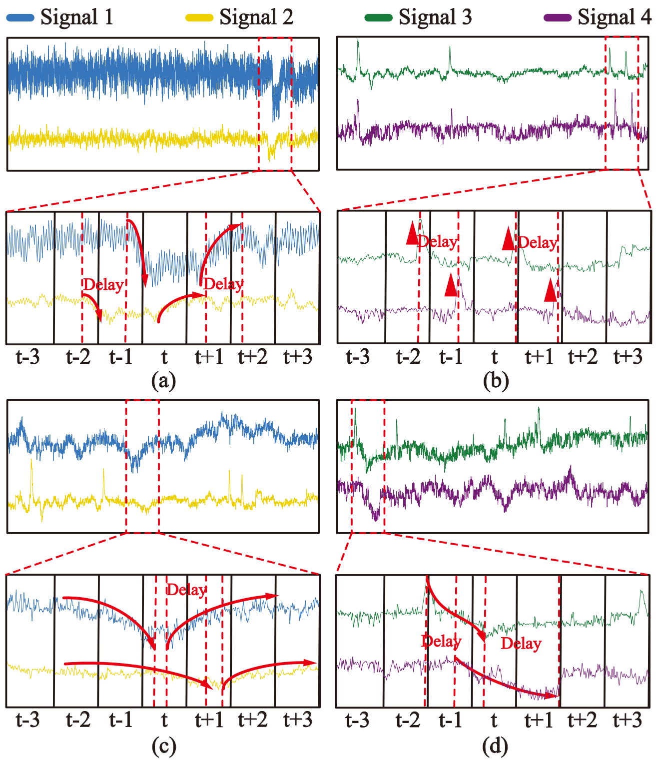

Figure 4: Temporal dependencies of emotional EEG.

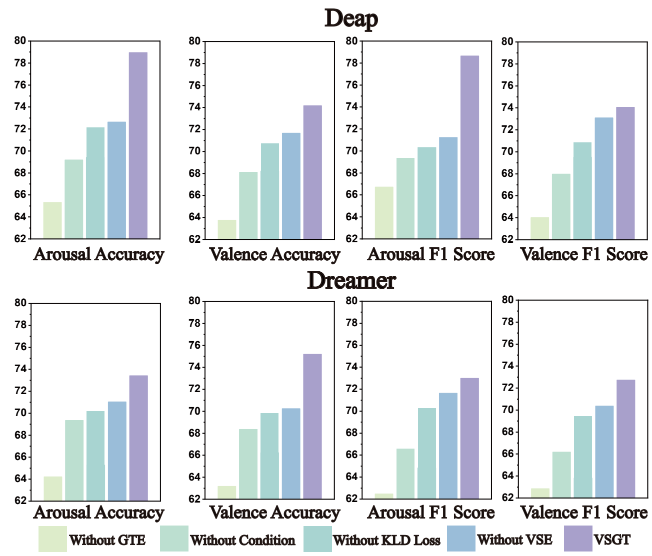

Figure 5: Ablation Study results.

slice $t-2$ and an amplitude increase at the time slice t, while signal 1 has an amplitude decrease only at the time slice $t-1$ and an amplitude increase at the time slice $t+1$. This suggests that the VSGT noticed that there was a significant delay in the response time of Signal 1, even though both Signal 1 and Signal 2 appeared similarly trough. In Figure 4 (b), Signal 4 appears to have successive peaks at the time slices $t-1$ and $t+1$, whereas similar successive peaks for signal 3 occur at the time slices t and $t-2$. This similarly suggests that there is a response delay in the brain regions corresponding to these two signals. In Figure 4 (c), the similar troughs of Signal 1 and Signal 2 begin at the same time slice while the signal 2 reaches its lowest point one time slice later than the signal 1. In Figure 4 (4), Signal 3 and Signal 4 all show a significant amplitude decrease, but the decrease of Signal 4 starts later and lasts longer. We suggest that these correlations between different brain regions across time slices contain important temporal dependency information. For example, from the physiological perspective, this delay responds to asynchronous activity between brain regions; from the signal processing perspective, the channel in which the first response occurs may be closer to the center of activity in the brain. Thus, the time dependency captured by the VSGT is physiologically plausible.

4.4 Ablation Study

We performed ablation studies to assess the impact of VSE and GTE and their core components. Figure 5 illustrates the results for the four cases. 1) Without VSE: Removing the VSE brings the least impact on the model because the temporal dependency inferred by the GTE is built on two neighboring time slices, which also contain a certain degree of spatial dependency information. 2) Without GTE: Removing GTE has the greatest impact on the model because when there is no GTE, the KLD part of Loss makes the spatial dependency of each time slice converge to the same, and this same spatial dependency is directly used to update the node embedding. 3) Without Condition: Not using the temporal dependency distribution as the conditional probability distribution of the spatial dependency results in the temporal dependency distribution behaving randomly when some samples of the spatial dependency converge to zero. 4) Without KLD: Removing the KLD part loss makes the VSE a simple linear module, and the spatial dependency of the current time slice can neither be correlated with the previous time slice nor be constrained by prior knowledge, so the result is relatively close to the direct removal of the VSE.

5 Conclusion

We propose the VSGT for EEG-based emotion recognition. Through VSE, VSGT realizes spatial dependency that is based on prior knowledge and can be updated. With GTE, VSGT captures the correlation of different brain regions across time slices in the temporal dependency of emotional EEG, which has been neglected in the past. RSTs integrate spatial-temporal dependencies with a recurrent structure and update the representation of each time slice. The experimental results show that the comprehensive spatial-temporal dependencies of EEG can be learned and utilized to improve the accuracy of EEG-based emotion recognition. In addition, interpretability experiments showed that the spatial and temporal dependencies inferred by the VSGT are consistent with existing physiological phenomena, which further explains the underlying working mechanism of the model.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62372334.

Contribution Statement

†Co-first Authors, *Corresponding Authors

References

[Braun *et al.*, 2015] Urs Braun, Axel Schäfer, Henrik Walter, Susanne Erk, Nina Romanczuk-Seiferth, Leila Haddad, Janina I Schweiger, Oliver Grimm, Andreas Heinz, et al. Dynamic reconfiguration of frontal brain networks during executive cognition in humans. *Proceedings of the National Academy of Sciences*, 112(37):11678–11683, 2015.

[Chung *et al.*, 2015] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. *Advances in neural information processing systems*, 28, 2015.

[Coulson, 2004] Mark Coulson. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of nonverbal behavior*, 28:117–139, 2004.

[Deng *et al.*, 2021] Xiangwen Deng, Junlin Zhu, and Shangming Yang. SFE-Net: EEG-based emotion recognition with symmetrical spatial feature extraction. In *Proceedings of the 29th ACM international conference on multimedia*, pages 2391–2400, 2021.

[Ding and Guan, 2023] Yi Ding and Cuntai Guan. GIGN: Learning graph-in-graph representations of EEG signals for continuous emotion recognition. In *45th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 1–5, 2023.

[Ding *et al.*, 2022] Yi Ding, Neethu Robinson, Su Zhang, Qiuhao Zeng, and Cuntai Guan. Tsception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. *IEEE Transactions on Affective Computing*, 2022.

[Ding *et al.*, 2023] Yi Ding, Neethu Robinson, Chengxuan Tong, Qiuhao Zeng, and Cuntai Guan. LGGNet: Learning from local-global-graph representations for brain–computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Ding *et al.*, 2024] Yi Ding, Su Zhang, Chuangao Tang, and Cuntai Guan. MASA-TCN: Multi-anchor space-aware temporal convolutional neural networks for continuous and discrete EEG emotion recognition. *IEEE Journal of Biomedical and Health Informatics*, pages 1–12, 2024.

[Du *et al.*, 2022] Guanglong Du, Jinshao Su, Linlin Zhang, Kang Su, Xueqian Wang, Shaohua Teng, and Peter Xiaoping Liu. A multi-dimensional graph convolution network for EEG emotion recognition. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11, 2022.

[Feng *et al.*, 2022] Lin Feng, Cheng Cheng, Mingyan Zhao, Huiyuan Deng, and Yong Zhang. EEG-based emotion recognition using spatial-temporal graph convolutional lstm with attention mechanism. *IEEE Journal of Biomedical and Health Informatics*, 26(11):5406–5417, 2022.

[Gonzalez-Castillo *et al.*, 2015] Javier Gonzalez-Castillo, Colin W Hoy, Daniel A Handwerker, Meghan E Robinson, Laura C Buchanan, Ziad S Saad, and Peter A Bandettini. Tracking ongoing cognition in individuals using brief, whole-brain functional connectivity patterns. *Proceedings of the National Academy of Sciences*, 112(28):8762–8767, 2015.

[Jia *et al.*, 2020] Ziyu Jia, Youfang Lin, Xiyang Cai, Haobin Chen, Haijun Gou, and Jing Wang. SST-EmotionNet: Spatial-spectral-temporal based attention 3D dense network for EEG emotion recognition. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2909–2917, 2020.

[Jia *et al.*, 2021] Ziyu Jia, Youfang Lin, Jing Wang, Zhiyang Feng, Xianghe Xie, and Caijie Chen. HetEmotionNet: two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1047–1056, 2021.

[Jiang *et al.*, 2023] Wei-Bang Jiang, Xu Yan, Wei-Long Zheng, and Bao-Liang Lu. Elastic graph transformer networks for EEG-based emotion recognition. In *ICASSP*, pages 1–5. IEEE, 2023.

[Jin *et al.*, 2021] Ming Jin, Hao Chen, Zhunan Li, and Jinpeng Li. EEG-based emotion recognition using graph convolutional network with learnable electrode relations. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5953–5957. IEEE, 2021.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Koelstra *et al.*, 2011] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

[Lewis, 2005] Marc D Lewis. Bridging emotion theory and neurobiology through dynamic systems modeling. *Behavioral and brain sciences*, 28(2):169–194, 2005.

[Li *et al.*, 2018] Yang Li, Wenming Zheng, Zhen Cui, Tong Zhang, and Yuan Zong. A novel neural network model based on cerebral hemispheric asymmetry for EEG emotion recognition. In *IJCAI*, pages 1561–1567, 2018.

[Li *et al.*, 2019] Yang Li, Wenming Zheng, Lei Wang, Yuan Zong, and Zhen Cui. From regional to global brain: A novel hierarchical spatial-temporal neural network model for EEG emotion recognition. *IEEE Transactions on Affective Computing*, 13(2):568–578, 2019.

[Li *et al.*, 2021a] Rui Li, Yiting Wang, and Bao-Liang Lu. A multi-domain adaptive graph convolutional network for EEG-based emotion recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5565–5573, 2021.

[Li *et al.*, 2021b] Xiaoxu Li, Wenming Zheng, Yuan Zong, Hongli Chang, and Cheng Lu. Attention-based spatio-temporal graphic LSTM for EEG emotion recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[Liu *et al.*,] Chenyu Liu, Xinliang Zhou, Zhengri Zhu, Liming Zhai, Ziyu Jia, and Yang Liu. VBH-GNN: Variational bayesian heterogeneous graph neural networks for cross-subject emotion recognition. In *The Twelfth International Conference on Learning Representations*.

[Liu *et al.*, 2024] Chenyu Liu, Xinliang Zhou, Yihao Wu, Ruizhi Yang, Liming Zhai, Ziyu Jia, and Yang Liu. Graph neural networks in EEG-based emotion recognition: A survey. *arXiv preprint arXiv:2402.01138*, 2024.

[Ma *et al.*, 2019] Jiaxin Ma, Hao Tang, Wei-Long Zheng, and Bao-Liang Lu. Emotion recognition using multimodal residual LSTM network. In *Proceedings of the 27th ACM international conference on multimedia*, pages 176–183, 2019.

[Min *et al.*, 2022] Jungwon Min, Kaoru Nashiro, Hyun Joo Yoo, Christine Cho, Padideh Nasseri, Shelby L Bachman, Shai Porat, Julian F Thayer, Catie Chang, Tae-Ho Lee, et al. Emotion downregulation targets interoceptive brain regions while emotion upregulation targets other affective brain regions. *Journal of Neuroscience*, 42(14):2973–2985, 2022.

[Peng *et al.*, 2022] Yong Peng, Fengzhe Jin, Wanzeng Kong, Feiping Nie, Bao-Liang Lu, and Andrzej Cichocki. OGSSL: A semi-supervised classification model coupled with optimal graph learning for EEG emotion recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:1288–1297, 2022.

[Salvador *et al.*, 2005] Raymond Salvador, John Suckling, Martin R Coleman, John D Pickard, David Menon, and ED Bullmore. Neurophysiological architecture of functional magnetic resonance images of human brain. *Cerebral cortex*, 15(9):1332–1342, 2005.

[Schirrmeister *et al.*, 2017] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.

[Schmidt and Trainor, 2001] Louis A Schmidt and Laurel J Trainor. Frontal brain electrical activity (EEG) distinguishes valence and intensity of musical emotions. *Cognition & Emotion*, 15(4):487–500, 2001.

[Soleymani *et al.*, 2011] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing*, 3(1):42–55, 2011.

[Song *et al.*, 2018] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2018.

[Tao *et al.*, 2020] Wei Tao, Chang Li, Rencheng Song, Juan Cheng, Yu Liu, Feng Wan, and Xun Chen. EEG-based emotion recognition via channel-wise attention and self attention. *IEEE Transactions on Affective Computing*, 2020.

[Wu *et al.*, 2020] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 753–763, 2020.

[Xu *et al.*, 2023] Tao Xu, Wang Dang, Jiabao Wang, and Yun Zhou. DAGAM: a domain adversarial graph attention model for subject-independent EEG-based emotion recognition. *Journal of Neural Engineering*, 20(1):016022, 2023.

[Ye *et al.*, 2022] Mengqing Ye, CL Philip Chen, and Tong Zhang. Hierarchical dynamic graph convolutional network with interpretability for EEG-based emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[Zeng *et al.*, 2007] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: audio, visual and spontaneous expressions. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 126–133, 2007.

[Zeng *et al.*, 2022] Hong Zeng, Qi Wu, Yanping Jin, Haohao Zheng, Mingming Li, Yue Zhao, Hua Hu, and Wanzeng Kong. Siam-GCAN: a siamese graph convolutional attention network for EEG emotion recognition. *IEEE Transactions on Instrumentation and Measurement*, 71:1–9, 2022.

[Zhang *et al.*, 2019] Tong Zhang, Xuehan Wang, Xiangmin Xu, and CL Philip Chen. GCB-Net: Graph convolutional broad network and its application in emotion recognition. *IEEE Transactions on Affective Computing*, 13(1):379–388, 2019.

[Zhong *et al.*, 2020] Peixiang Zhong, Di Wang, and Chunyan Miao. EEG-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing*, 13(3):1290–1301, 2020.

[Zhou *et al.*, 2023] Xinliang Zhou, Chenyu Liu, Liming Zhai, Ziyu Jia, Cuntai Guan, and Yang Liu. Interpretable and robust AI in EEG systems: A survey. *arXiv preprint arXiv:2304.10755*, 2023.

[Zhu *et al.*, 2017] Xinge Zhu, Liang Li, Weigang Zhang, Tianrong Rao, Min Xu, Qingming Huang, and Dong Xu. Dependency exploitation: A unified cnn-rnn approach for visual emotion recognition. In *IJCAI*, pages 3595–3601, 2017.