MVGT: A Multi-view Graph Transformer Based on Spatial Relations for EEG Emotion Recognition

Yanjie Cui, **Xiaohong Liu***, **Jing Liang**, **Yamin Fu**

School of Computer Science (National Pilot Software Engineering School)
Beijing University of Posts and Telecommunications, Beijing, 100876, China
{yanjiecui, xiaohongliu, liangjing18, fuyamin}@bupt.edu.cn

Abstract—Electroencephalography (EEG), a medical imaging technique that captures scalp electrical activity of brain structures via electrodes, has been widely used in affective computing. The spatial domain of EEG data is rich in affective information. However, few of the existing studies have simultaneously analyzed EEG signals from the perspectives of geometric and anatomical structures in spatial domain. In this paper, we propose a multi-view graph transformer (MVGT) based on spatial relations, which integrates information from the temporal, frequency and spatial domains, including geometric and anatomical structures, so as to enhance the expressive power of the model comprehensively. We encode the spatial information of EEG channels into the model, thereby improving its ability to comprehend the spatial structure of the channels. Experimental results from publicly available datasets demonstrate that our proposed model outperforms state-of-the-art baseline methods. Furthermore, the results also show that the MVGT could extract information from multiple domains and capture inter-channel relationships in EEG-based emotion recognition tasks effectively.

Index Terms—EEG, emotion recognition, graph transformer, structure encoding

I. INTRODUCTION

Affective computing is commonly employed for the analysis of emotional states through Human-Computer Interaction (HCI) systems, which collect multimodal data from subjects, including voice signals, self-reports, body gestures and physiological signals. Compared to other modalities, physiological signals have certain advantages. These signals are directly captured from the subjects' mental states, thus prevent subjects from disguising or hiding. The physiological signals commonly used to measure emotions are electroencephalography (EEG), electrocardiography (ECG), electromyography (EMG), and galvanic skin response (GSR), etc., among which EEG is often utilized to analyze the cognitive functions of human brain. Electrical signals from brain neurons are collected through dry and noninvasive electrodes placed on the scalp [1]. Nowadays, due to its high temporal resolution, portability, and affordability, this method is widely employed to study brain changes in response to emotional stimuli [2].

Traditional EEG features are mainly divided into three categories, i.e., time domain, frequency domain, and time-frequency domain features. Given the low signal-to-noise

ratio and substantial fluctuations inherent in EEG signals, frequency domain features are often used for EEG-based emotion recognition tasks. The typical approach involves decomposing the raw signals into five frequency bands: δ, θ, α, β, γ. Frequency domain features, such as power spectral density (PSD) [3], differential entropy (DE) [4], [5], differential asymmetry (DASM) [6] and rational asymmetry (RASM) [7], are subsequently extracted from each frequency band respectively.

The spatial structure of the brain also contains rich emotional information. Emotional states may involve distributed circuits rather than only a single brain region [8]. Asymmetry between the left and right hemispheres can reflect changes in valence and arousal [9]. Recent studies have highlighted the importance of utilizing spatial domain information. Li et al. [10] introduced recurrent neural networks to learn the asymmetric differences between the left and right hemispheres. Li et al. [11] also utilized hierarchical neural networks to learn both regional and global information of spatial-temporal EEG features. Graph neural networks (GNNs) are emerging as a powerful tool for analyzing spatial information in EEG emotion recognition. Song et al. [12] dynamically learned relationships between EEG channels using a graph convolutional network (GCN). Zhong et al. [13] incorporated asymmetry of the hemispheres into the adjacency matrix to model graph structure. Li et al. [14] also utilized an multi-domain adaptive graph convolutional network (MD-AGCN) to learn relationships between channels. Ding et al. [15] incorporated lobe information as prior knowledge into the GNN. Jiang et al. [16] proposed an elastic graph transformer (EmoGT) to extract emotional information. Although these methods have achieved excellent performance in emotion recognition, they have a common issue: they all rely on GNNs based on neighborhood aggregation schemes which may pose potential risks such as over-smoothing [17]–[19], under-reaching [20], and over-squashing [21]. Additionally, these methods do not take the geometric and anatomical structure information of the brain into account.

The main contributions of this paper lie in three aspects: (1) We propose a multi-view graph transformer (MVGT) based on spatial relations, fusing information from multiple perspectives including temporal, frequency, and spatial domains. (2) Our method, based on graph transformer, mitigates the

---

*Corresponding author

potential risks of over-smoothing, under-reaching and over-squashing occurring in traditional GNNs. Additionally, it enhances the model's expressive power in emotion recognition by introducing spatial encodings based on geometric and brain lobe information. (3) Extensive experiments conducted on public datasets SEED and SEED-IV show our model achieves superior performance over the baseline models in emotion classification tasks.

II. RELATED WORK

In this section, we review the related work in terms of EEG-based emotion recognition and graph transformer.

A. EEG-based Emotion Recognition

EEG signals are inherently noisy and susceptible to channel crosstalk [22]. Due to the complexity of EEG signals, it is challenging to isolate clean and independent signals. Therefore, it is crucial to select what form of data to analyze under conditions of high noise. Effective features of EEG signals can reduce noise and facilitate the recognition of cognitive patterns in specific tasks. Experimental evidence suggests that frequency domain features are often associated with behavioral patterns [23], hence they are commonly used in EEG analysis.

Along with the development of deep learning, increasingly complex models with rich expressive abilities have emerged and have been extensively utilized in EEG signal analysis. Zheng et al. [5] employed a deep belief network to analyze important frequency domain components and effective channels based on the learned parameters. Song et al. [12] used a GCN method based on Chebyshev polynomials [24] to dynamically learn the representations of EEG signals. Zhong et al. [13] innovatively incorporated the inter-channel asymmetry of the hemispheres as prior knowledge into the adjacency matrix in 3D space. The reasonable combination of multi-domain information improves the accuracy in the emotion recognition tasks. Li et al. [14] proposed the MD-AGCN that integrates the temporal domain, frequency domain, and functional connectivity. Ding et al. [15], inspired by neuroscience research, combined intra-region convolution and inter-region convolution based on brain lobe information to learn cognitive patterns. Jiang et al. [16] utilized the advantages of GCN in the spatial domain and Transformer in the temporal domain to improve the accuracy of emotion classification.

B. Graph Transformer

The GNNs used in the methods above are based on neighborhood aggregation schemes. However, classical GNNs based on message passing (MPGNNs) may lead to over-smoothing [17]–[19], under-reaching [20], and may also fail to fit long-range signals due to over-squashing [21], which limit the expressive power of the model. Graph transformers (GTs) alleviate such effects as they have a global receptive field [25]. Nevertheless, without sufficiently expressive structural and positional encodings, GTs cannot capture effective graph structures [26]. Dwivedi et al. [27] utilized eigenvectors of graph Laplacian as position encodings in a fully connected

graph transformer and integrated edge features into the attention mechanism. Building on this, SAN [28] used a full Laplacian spectrum to learn the positional encodings for each node. Graphormer [29], [30] employed node centrality and node distance metric to implement structural and relative positional encodings, achieving state-of-the-art performance on challenging graph datasets. For EEG emotion recognition, Li et al. [31] innovatively combined a masked autoencoder based on self-supervised learning with a CNN-Transformer hybrid structure, effectively improving classification accuracy. However, this method only used the sine-cosine positional encoding, limiting the Transformer's ability to learn spatial information.

III. PRELIMINARY

A. Graph Neural Network (GNN)

Let $G = (V, E)$ define a graph, where $V = \{v_1, v_2, \cdots, v_n\}$ represents the nodes in the graph, and $E = \{e_1, e_2, \cdots, e_m\}$ is the edges between the nodes. The representation of node v_i is denoted as $x_i \in \mathbb{R}^d$. Most existing GNNs [17], [32]–[35] adopt neighborhood aggregation schemes, iteratively aggregating representations of its first or higher-order neighbors, followed by using backpropagation (BP) to learn data-driven feature representations. We define the representation of node v_i at the l-th iteration as h_i^l and define $h_i^0 = x_i$. The l-th iteration can be represented as:

$$a_i^l = \text{AGGREGATE}^l \left(\left\{ \varphi_\theta(h_j^{l-1}, e_{ji}) : j \in \mathcal{N}(v_i) \right\} \right), \quad (1)$$

$$h_i^l = \text{UPDATE}^l \left(h_i^{l-1}, a_i^l \right), \quad (2)$$

where φ_θ represents a differentiable function used for feature transformation of node and edge information. The set $\mathcal{N}(v_i)$ is the neighbors of v_i. The AGGREGATE function is used to aggregate the transformed representations using a differentiable, permutation invariant function (such as mean, sum, max, etc.). The goal of UPDATE function is to integrate the information from neighbors into the node representation. For graph classification, the READOUT operation is typically used to obtain a representation of the entire graph, which is then fed into a classifier to determine the graph label.

B. Graphormer

The Transformer [36] is undeniably one of the most popular deep neural network architectures today, driving significant advancements in natural language processing and computer vision. From the perspective of GNNs, Transformer can be interpreted as a GNN acting on a fully connected graph. Therefore, it is feasible to use the Transformer to address tasks on graph data. The ability to properly incorporate the structural information of graphs into the model is the key for leveraging its expressive power. Graphormer [29], [30] can go beyond classical MPGNNs in expressive power and achieves state-of-the-art performance on large molecular benchmarks. Graphormer incorporates centrality encoding into the graph data and integrates spatial encoding, edge encoding into the attention mechanism, which can be expressed as:
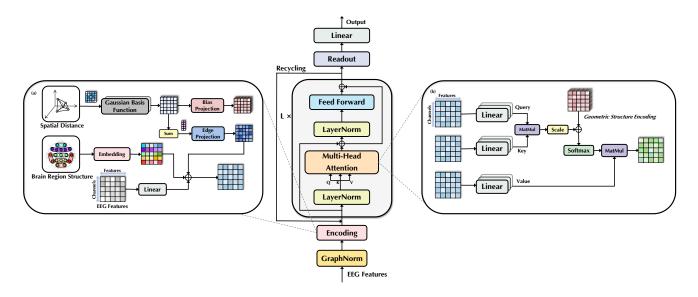
Figure 1. Overall structure of MVGT. (a) represents the encoding process of brain region structure and geometric structure. (b) depicts the process of calculating inter-channel correlations based on the attention mechanism and geometric structure encoding.

$$\boldsymbol{A}_{ij} = \frac{(\boldsymbol{h}_i \boldsymbol{W}_Q)(\boldsymbol{h}_j \boldsymbol{W}_K)^\top}{\sqrt{d}} + b_{\phi(v_i, v_j)} + c_{ij}, \qquad (3)$$

where c_{ij} represents the edge encoding on the shortest path and the bias term $b_{\phi(v_i, v_j)}$ can adaptively adjust the correlations between v_i and v_j.

IV. METHODS

In this section, we introduce the methods employed in the EEG emotion recognition task. First, we elaborate on the embedding of temporal information. Second, leveraging the spatial geometry and physiological anatomy of the brain, we propose two novel and simple designs of encoding that enable the model to adaptively learn the inter-channel correlations. Finally, we present the detailed implementations of MVGT.

A. Problem Definition

EEG signals can be represented as a two-dimensional matrix with respect to channels and time. Given that channels exhibit spatial structure, they can be structured into a fully connected graph $G = (V, E)$, where V denotes the nodes, representing EEG channels, and E denotes the edges, representing the connections between channels. The features of the nodes at time t are denoted by $\boldsymbol{X}_t = \left[\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \cdots, \boldsymbol{x}_n^\top\right]^\top \in \mathbb{R}^{n \times d}$, where $n = |V|$ represents the number of nodes and d represents the feature dimension.
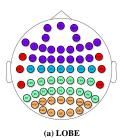
B. Temporal Embedding

EEG signals have high temporal resolution and contain rich temporal information. Because of the multi-electrode acquisition method, EEG signals can be regarded as multivariate time series. When processing time series, the embedding of temporal information are crucial. EmoGT treated the features of different channels at the same time points as tokens
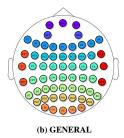
and employed an attention mechanism to extract temporal correlations between them. Due to the anisotropic volume conduction characteristics [37] in human brain tissues, there may be temporal delays between different channels, which in turn leads to time-unaligned events at a single moment thus causing performance degradation. MD-AGCN utilized the convolutional operation to extract temporal information along the time axis from continuous EEG segments, with the receptive field limited by the size of the convolution kernel. Inspired by [38], we broaden the receptive field by considering the entire time series as an embedded token rather than a single time point. First, following the methods of MD-AGCN and EmoGT, we use overlapping sliding windows of size T to split EEG signals along the time axis and use these segments as samples, which are then fed into the attention module in the form of continuous segments. After processing with sliding windows, we obtain $\tilde{\boldsymbol{X}} = \left(\tilde{\boldsymbol{X}}_1, \tilde{\boldsymbol{X}}_2, \cdots, \tilde{\boldsymbol{X}}_S\right)$, where the s-th sample is $\tilde{\boldsymbol{X}}_s \in \mathbb{R}^{n \times Tf}$, S is the number of continuous EEG segments, n is the number of channels, and f is the dimension of frequency domain features.

According to the universal approximation theorem [39], the feed-forward neural network (FFN), as the basic module of the Transformer encoder, can learn the intrinsic properties to describe a time series and is a superior predictive representation learner compared to self-attention [38]. Therefore, using continuous time segments as the input to the FFN may be more effective in extracting the temporal information of each channel independently.

C. Spatial Encoding

The special structure of the brain encompasses rich spatial information. Fully exploiting structural information is beneficial for the recognition and analysis of cognitive patterns in the brain. Therefore, to better recognize emotional patterns

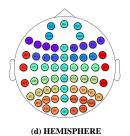**(a) LOBE**   **(b) GENERAL**   **(c) FRONTAL**   **(d) HEMISPHERE**

Figure 2. The brain region division scheme is illustrated. (a) LOBE scheme shows a coarse partitioning based on lobe structures. (b) GENERAL scheme represents a fine-grained partitioning of the brain lobes. (c) FRONTAL scheme introduces symmetry of the left and right frontal regions based on the GENERAL scheme. (d) HEMISPHERE scheme further enhances the channel symmetry in the partitioning scheme. Channels of the same color belong to the same brain region.

in emotion classification tasks, we employ two simple but effective methods of spatial encoding: brain region encoding and geometric structure encoding.

1) Brain Region Encoding: Neuroscience research demonstrated that the activation of a specific brain region often leads to the concurrent activation of related brain regions responsible for the same high-level cognition [40]. In EEG emotion recognition, incorporating relevant neuroscience findings can typically enhance recognition accuracy. Zhong et al. [13] integrated the asymmetry of neural activity between the left and right hemispheres as prior knowledge into the adjacency matrix, effectively enhancing recognition accuracy. Li et al. [10] improved emotion pattern recognition performance by capturing the differences between the left and right hemispheres. Ding et al. [15] divided EEG channels into different regions and combined local intra-region convolution with global inter-region convolution, achieving good results on the DEAP [41] dataset. With reference to the three divisions of [15], we adopt four brain region divisions, which divide the EEG channels into different regions based on a prior knowledge, aiming to incorporate the brain region information into the model. The division schemes are as follows:

- We divide the regions based on the anatomical structure of the brain and implement LOBE scheme.
- To further investigate the expressive power of brain region encoding, we conduct a detailed division of brain lobes according to the 10-20 system based on electrode positions, employing the GENERAL scheme.
- Asymmetric EEG activity in the frontal lobe can be utilized for discriminating valence changes [9]. The left frontal lobe exhibits a stronger correlation with joy and happiness, while the right frontal lobe is more strongly correlated with fear and sadness. Thus we further divide the frontal lobe region into two symmetrical regions to obtain the FRONTAL scheme.
- According to the symmetry of brain structure [42], we make a finer division of the brain lobe regions, defining the HEMISPHERE scheme.

The four schemes mentioned above are shown in Fig. 2. In terms of specific implementation, we assign a brain region tag to each electrode, then map the tags into an embedding space using a learnable projection function, and simply add the embeddings to the node features. The encoding of node i is represented as follows:

$$r_i = \text{Embedding}(\text{Tag}(\boldsymbol{x}_i)), \ \boldsymbol{r}_i \in \mathbb{R}^d, \quad (4)$$

$$\boldsymbol{h}_i^0 = \boldsymbol{x}_i \boldsymbol{W}_{\mathcal{X}} + \boldsymbol{r}_i, \quad (5)$$

where $\boldsymbol{W}_{\mathcal{X}} \in \mathbb{R}^{Tf \times d}$ is a learnable projection function, and d represents the dimension of the embedding. Through the above encoding method, we integrate the brain's anatomy information into the model.

2) Geometric Structure Encoding: In the real world, the human reasoning process considers not only the semantic relationships between objects but also their spatial relations. EEG channels have a 3D structure, and the functional connectivity between these channels lacks precise definitions. Therefore, we represent the relationships between EEG channels as a fully connected directed graph. Firstly, let $\phi(i, j)$ represent the Euclidean distance between node i and node j, and encode $\phi(i, j)$ using a set of Gaussian basis functions [30], [43]. Let $\boldsymbol{b}_k \in \mathbb{R}^{n \times n}$ denote one of the Gaussian basis functions. The element (i, j) of this function can be expressed as:

$$\boldsymbol{b}_k(i, j) = \mathcal{G}_k \left(\alpha_{ij} \phi(i, j) + \beta_{ij} - \mu_k, \sigma_k \right), \quad (6)$$

where α_{ij}, β_{ij}, μ_k, and σ_k are learnable parameters, and i and j denote the index of the source and target node, respectively. The result of the basis functions can be represented as $\boldsymbol{B} = \|_{k=1}^K \boldsymbol{b}_k$, with $\boldsymbol{B} \in \mathbb{R}^{n \times n \times K}$, where $\|$ denotes the concatenation. All geometric encodings of each node are then summed up along the second dimension and then transformed linearly.

$$\boldsymbol{h}_i^0 = \boldsymbol{x}_i \boldsymbol{W}_{\mathcal{X}} + \boldsymbol{z}_i \boldsymbol{W}_{\mathcal{Z}} + \boldsymbol{r}_i, \ \boldsymbol{z}_{i,k} = \sum_{j=1}^n \boldsymbol{B}_{i,j,k}, \quad (7)$$

$$\boldsymbol{B}' = \text{Projection}(\boldsymbol{B}), \quad (8)$$

where i is the node index, k is the index of basis function and $\boldsymbol{W}_{\mathcal{Z}} \in \mathbb{R}^{K \times d}$ is a learnable projection matrix. Projection : $\mathbb{R}^{n \times n \times K} \mapsto \mathbb{R}^{n \times n \times M}$ is a nonlinear transformation, where M is the number of attention heads. We incorporate this encoding as a bias term into the softmax attention.

Our proposed spatial encoding matrix is directed, which is inconsistent with the assumption of a symmetric adjacency matrix [13], [16]. Using directed connections provides the model with greater expressive power because the correlation

between node pairs (i, j) and (j, i) may differ. Since we assume nodes are fully connected, we avoid specific assumptions about inter-channel correlations and learn the functional correlations between nodes through the encodings. Letting l denote the model depth, and i denote the index of multi-head attention, the functional brain connectivity of the s-th sample can be represented as:

$$\boldsymbol{A}_s^{l,i} = \text{Softmax}\left(\frac{\left(\boldsymbol{H}_s^l \boldsymbol{W}_{\mathcal{Q}}^{l,i}\right)\left(\boldsymbol{H}_s^l \boldsymbol{W}_{\mathcal{K}}^{l,i}\right)^\top}{\sqrt{d_h^{l,i}}} + \boldsymbol{B}'^i\right), \quad (9)$$

where the projections are learnable parameters $\boldsymbol{W}_{\mathcal{Q}}^{l,i} \in \mathbb{R}^{d \times d_h}$ and $\boldsymbol{W}_{\mathcal{K}}^{l,i} \in \mathbb{R}^{d \times d_h}$. The scalar $d_h^{l,i}$ is the second dimension of $\boldsymbol{W}_{\mathcal{K}}^{l,i}$. This encoding method integrates temporal, frequency, and spatial domain features into the model, enhancing its expressive power. We compute the attention scores between nodes using embedded vectors, representing the semantic correlations between different nodes from multiple perspectives. Finally, the attention scores are added to the spatial geometric encoding to obtain the correlations between channels.

D. Implementation Details of MVGT

In this section, we describe the overall architecture of the model, including spatial encodings and the Transformer encoder, as illustrated in Fig. 1. For better optimization, we first apply GraphNorm [44] to normalize the input features to a range between 0 and 1. Subsequently, we perform geometric and regional structure encodings to obtain multi-domain embeddings. The encodings could be characterized as below:

$$\boldsymbol{X}_s' = \text{GraphNorm}(\boldsymbol{X}_s) \quad (10)$$
$$\boldsymbol{H}_s^0 = \text{SpatialEncoding} + \text{Proj}(\boldsymbol{X}_s') \quad (11)$$

We employ a Pre-LN Transformer structure, applying layer normalization (LN) before the multi-head attention (MHA) and the FFN. A recent study suggests that the Pre-LN structure yields more stable gradients and is more favorable for optimizer, enabling faster convergence [45] compared to Post-LN. Additionally, we utilize dropout to mitigate overfitting. This process is represented as follows:

$$\boldsymbol{H}_s'^l = \text{MHA}(\text{LN}(\boldsymbol{H}_s^{l-1})) + \boldsymbol{H}_s^{l-1} \quad (12)$$
$$\boldsymbol{H}_s^l = \text{FFN}(\text{LN}(\boldsymbol{H}_s'^l)) + \boldsymbol{H}_s'^l \quad (13)$$

Inspired by [30], [46], we feed the outputs recursively into the same modules, denoted as "recycling" in Fig. 1. The iterative refinement progressively refines the model's ability to discriminate encoded information and understand emotional patterns, thereby helping the model capture more effective details.

V. EXPERIMENTS

A. Datasets

For our experiments, we select the SEED [5] and SEED-IV [47] datasets to evaluate the effectiveness of our model.

These datasets consist of EEG signals recorded from subjects watching emotion-eliciting videos.

SEED dataset comprises data from 15 subjects who participated in three sessions, each separated by at least one week. Each session consists of 15 trials capturing emotional labels, with the emotion labels being positive, negative, and neutral.

SEED-IV dataset is constituted by EEG signals from 15 subjects across three separate sessions conducted at different times, using the same device as the SEED dataset. This dataset encompasses four emotion labels: neutral, sad, fear, and happy. In each session, each subject underwent 24 trials.

B. Settings

To prevent potential data leakage that could arise from segment-wise shuffling, we split the training and test sets at the trial level. Following the settings of previous studies [5], [10], [12]–[14], [16], [31], [47], we use pre-computed differential entropy (DE) features for the recognition task. For the SEED dataset, we use the first 9 trials of each subject as the training set and the last 6 trials as the test set, as done in previous research. The DE features are computed using five frequency bands extracted from 1s nonoverlapping windows. The model performance is evaluated based on the average accuracy and standard deviation across all subjects over two sessions of EEG data. Similarly, for the SEED-IV dataset, we use the first 16 trials as the training set and the last 8 trials as the test set. The DE features for SEED-IV are calculated using 4s windows. The performance of our model is assessed using data from all

Table I
THE CLASSIFICATION ACCURACIES (MEAN/STD) ON SEED AND SEED-IV. MVGT-L, MVGT-G, MVGT-H, MVGT-F: MVGT USING LOBE, GENERAL, HEMISPHERE AND FRONTAL SCHEMES.

Model	SEED	SEED-IV
DGCNN [12]	90.40/08.49	69.88/16.29
BiHDM [10]	93.12/06.06	74.35/14.09
R2G-STNN [11]	93.34/05.96	-
RGNN [13]	94.24/05.95	79.37/10.54
MD-AGCN [14]	94.81/04.52	87.63/05.77
EmoGT [16]	95.02/05.99	91.20/09.60
MV-SSTMA [31]	95.32/3.05	92.82/5.03
MVGT-L	95.36/05.37	91.51/09.03
MVGT-G	94.43/05.35	**93.57**/08.60
MVGT-H	95.19/05.48	90.19/10.42
MVGT-F	**96.45**/04.40	91.62/09.05

three sessions.

For the input data, we use overlapping sliding windows of size T along the time axis to extract sample fragments, with T being set to 5. During experiments, the hidden dimension is set to 64 and the number of Gaussian basis functions is 32. The number of MHA layers is 4 and the number of attention heads is 2. The iterative refinement process is performed three times. We set the batch size to 32 and the learning rate within the range of 3e-5 to 3e-3. Cross-entropy is used as the loss function, and AdamW [48] is employed as the optimizer with a weight decay rate of 0.1.

C. Results Analysis

We compare the classification results based on the SEED and SEED-IV datasets with recent state-of-the-art models, as shown

in Table I. It is evident that our proposed model significantly outperforms the baseline models under the same experimental settings. For the SEED dataset, the model adopting the FRONTAL scheme achieves the best performance, with a classification accuracy of 96.45%. The LOBE scheme also achieves a slightly superior accuracy of 95.36%, compared to other models. For the SEED-IV dataset, the classification accuracy under the GENERAL scheme is 93.57%, achieving the best performance compared to baseline models. The MVGT model also demonstrates strong performance under other division schemes. Overall, our model achieves the best classification accuracy compared to the baselines. The results also suggest that selecting the specific division scheme relevant to the emotion task could enhance the expressive power of MVGT.
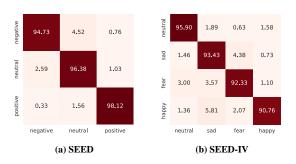

(a) SEED   **(b) SEED-IV**

Figure 3. Confusion matrices of MVGT. (a) Confusion matrix of MVGT-F on the SEED dataset. (b) Confusion matrix of MVGT-G on the SEED-IV dataset. Each row of the matrix represents the true labels while each column serves as the predicted labels.

Table II
ABLATION STUDY FOR THE CLASSIFICATION ACCURACIES (MEAN/STD) ON THE SEED AND SEED-IV DATASETS. SYMBOL "✓" INDICATES THE COMPONENT IS EMPLOYED.

Geometric Structure	Brain Region	SEED	SEED-IV
-	-	93.79/07.15	89.49/10.40
✓	-	94.52/06.04	90.00/09.62
-	✓	94.11/05.77	89.87/10.41
✓	✓	**96.45/04.40**	**93.57/08.60**

Fig. 3 illustrates the confusion matrices of MVGT-F on the SEED and MVGT-G on the SEED-IV, respectively. The values represent the classification accuracy of the model for different emotion classes. For the SEED dataset, our model achieves the highest accuracy in recognizing positive emotion (98.12%), followed by neutral emotion (96.38%), with negative emotion being slightly lower (94.73%). Only 0.33% of positive emotion samples are misclassified as negative, while only 0.76% of negative emotion samples are recognized as positive, indicating the model's effectiveness in distinguishing valence changes. For the SEED-IV dataset, our model performs best in recognizing neutral emotion, with an accuracy of 95.90%, while its performance on happy emotion is slightly lower than the other three emotions, with an accuracy of 90.76%. This could be attributed to the GENERAL scheme setting, making the model more sensitive to the balanced emotion.

Our model achieves state-of-the-art performance on both SEED and SEED-IV, primarily due to our comprehensive consideration of frequency, temporal, and spatial geometric information, combined with prior knowledge from neuroscience.

D. Ablation Study

To validate the effectiveness of spatial encodings, we conduct ablation experiments on the SEED and SEED-IV datasets, as presented in Table II. By removing both types of spatial encodings, we repeat the aforementioned experiments under the same experimental settings. On the SEED dataset, the model achieves an accuracy of 93.79% with a standard deviation of 7.15%. Compared to MVGT-F, the accuracy decreases by 2.66% and the standard deviation increases by 2.75% after removing spatial encodings. For the SEED-IV dataset, the accuracy drops by 4.08%, resulting in 89.49%, with the standard deviation rising by 1.80% to 10.40%, when compared to MVGT-G. The experiments demonstrate that incorporating spatial structure information benefits the model performance in emotion recognition tasks. Under experimental settings that consider only geometric structure or brain region structure, the model's classification accuracy improves over the plain model without any spatial encoding. Evidently, when considering both types of spatial structures simultaneously, the model performance significantly outperforms that of the plain model and models only using single spatial information. This indicates the effectiveness of our proposed spatial encodings and confirms that the expressive power of the graph transformer relies on the structural and positional encodings.

E. Visualization of Inter-channel relations

To better illustrate the correlations between channels, we visualize the inter-channel relations of MVGT-F on the SEED and MVGT-G on the SEED-IV. Given that the inter-channel relations might vary among different subjects, we calculate the average weights across all subjects. We focus on the last iteration of iterative refinement and select the 10 strongest connections of channel pairs. Fig. 4 shows the visualization results, where the rows represent the attention heads and the columns represent the layers of the MHA.

The parameters based on the SEED dataset indicate that emotion patterns are reflected in the activities of multiple brain regions. In the first layer of MVGT-F, the channels in the left frontal region have higher participation in the first attention head, while the channels in the right frontal region are more involved in the second head, potentially corresponding to positive and negative emotion patterns [9], respectively. In the second layer of the model, the parietal and occipital regions show higher involvement, which aligns with the findings on emotion patterns in [49]. As the model depth increases, the symmetrical connections in the lateral temporal regions of both hemispheres are enhanced, consistent with previous research by [5], [13], [16]. For the SEED-IV dataset, the connections in the frontal, parietal, and occipital regions are the most active, consistent with the findings of [13]. In the first attention
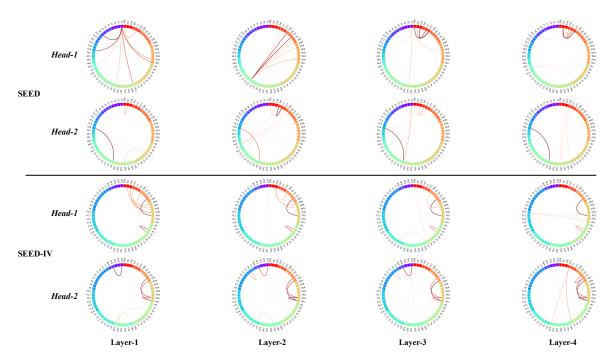
Figure 4. The learned inter-channel relationships from the SEED by the MVGT-F and from the SEED-IV by the MVGT-G are illustrated. The figures show the results of the last iteration in the iterative refinement, highlighting the top 10 channel pairs with the highest weights after softmax (darker colors indicate higher weights). Channels of the same brain region are represented in the same color. Rows correspond to attention heads, while columns represent the layers of the MHA mechanism.

head of MVGT-G, the strongest correlation is between O1 and PO3, followed by P4 and P2. Other connections are mainly distributed in the temporal and frontal regions. In the second head, the channel pairs (O1, PO5), (CB1, PO7), and (PO5, PO7) contribute the most to emotion recognition. Additionally, the connection between AF3 and FP1 provides important information for emotion processing, which aligns with the conclusions of [13], [16].

Overall, our model does not focus solely on the local information of a single brain region but instead considers both intra-regional and inter-regional information in depth. This confirms that emotional states result from interactions among widely distributed functional networks in the brain, as discussed by [50].

VI. CONCLUSIONS

In this paper, we propose a multi-view graph transformer based on spatial relations for EEG-based emotion recognition. Our model integrates information from multiple perspectives, including temporal, frequency and spatial domains. We incorporate spatial geometric encoding and brain region encoding to enhance the graph transformer's ability to perceive spatial structures. Additionally, the model adaptively learns interchannel relationships through the attention mechanism and the encoding of channel geometry. Extensive experiments on public emotion recognition datasets demonstrate that our proposed model outperforms other competitive baseline models.

Furthermore, the analysis of channel correlations indicates that emotional activities in the brain are not confined to a single local region but result from the coordinated action of multiple brain areas. Information from frontal, parietal, occipital, and lateral temporal lobes is valuable for emotion recognition to varing extents.

REFERENCES

[1] Babak A Taheri, Robert T Knight, and Rosemary L Smith. A dry electrode for eeg recording. *Electroencephalography and clinical neurophysiology*, 90(5):376–383, 1994.

[2] Christopher Niemic. Studies of emotion: a theoretical and empirical review of psychophysiological studies of emotion. *Journal of Undergraduate Research*, 2004.

[3] Lester I Goldfischer. Autocorrelation function and power spectral density of laser-produced speckle patterns. *Josa*, 55(3):247–253, 1965.

[4] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th international IEEE/EMBS conference on neural engineering (NER)*, pages 81–84. IEEE, 2013.

[5] Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on autonomous mental development*, 7(3):162–175, 2015.

[6] Yisi Liu and Olga Sourina. Real-time fractal-based valence level recognition from eeg. In *Transactions on computational science XVIII: special issue on Cyberworlds*, pages 101–120. Springer, 2013.

[7] Yuan-Pin Lin, Chi-Hong Wang, Tzyy-Ping Jung, Tien-Lin Wu, Shyh-Kang Jeng, Jeng-Ren Duann, and Jyh-Horng Chen. Eeg-based emotion recognition in music listening. *IEEE Transactions on Biomedical Engineering*, 57(7):1798–1806, 2010.

[8] Iris B Mauss and Michael D Robinson. Measures of emotion: A reviews. *Cognition and emotion*, pages 109–137, 2010.

[9] 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.

[10] Yang Li, Lei Wang, Wenming Zheng, Yuan Zong, Lei Qi, Zhen Cui, Tong Zhang, and Tengfei Song. A novel bi-hemispheric discrepancy model for eeg emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 13(2):354–367, 2020.

[11] 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, 2022.

[12] 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.

[13] 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.

[14] 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.

[15] 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*, pages 1–14, 2023.

[16] Wei-Bang Jiang, Xu Yan, Wei-Long Zheng, and Bao-Liang Lu. Elastic graph transformer networks for eeg-based emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[17] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *International conference on machine learning*, pages 1725–1735. PMLR, 2020.

[18] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3538–3545, 2018.

[19] Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020.

[20] Pablo Barceló, Egor V. Kostylev, Mikael Monet, Jorge Pérez, Juan Reutter, and Juan Pablo Silva. The logical expressiveness of graph neural networks. In *International Conference on Learning Representations*, 2020.

[21] Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations*, 2021.

[22] Michal Teplan et al. Fundamentals of eeg measurement. *Measurement science review*, 2(2):1–11, 2002.

[23] Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven Siegelbaum, A James Hudspeth, Sarah Mack, et al. *Principles of neural science*, volume 4. McGraw-hill New York, 2000.

[24] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

[25] Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35:14501–14515, 2022.

[26] Luis Müller, Mikhail Galkin, Christopher Morris, and Ladislav Rampášek. Attending to graph transformers. *Transactions on Machine Learning Research*, 2024.

[27] Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *AAAI Workshop on Deep Learning on Graphs: Methods and Applications*, 2021.

[28] Devin Kreuzer, Dominique Beaini, Will Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking graph transformers with spectral attention. *Advances in Neural Information Processing Systems*, 34:21618–21629, 2021.

[29] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34:28877–28888, 2021.

[30] Yu Shi, Shuxin Zheng, Guolin Ke, Yifei Shen, Jiacheng You, Jiyan He, Shengjie Luo, Chang Liu, Di He, and Tie-Yan Liu. Benchmarking graphormer on large-scale molecular modeling datasets. *arXiv preprint arXiv:2203.04810*, 2022.

[31] Rui Li, Yiting Wang, Wei-Long Zheng, and Bao-Liang Lu. A multi-view spectral-spatial-temporal masked autoencoder for decoding emotions with self-supervised learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6–14, 2022.

[32] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *International conference on machine learning*, pages 6861–6871. PMLR, 2019.

[33] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[34] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[35] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[37] Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of neural engineering*, 16(3):031001, 2019.

[38] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024.

[39] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

[40] Hedy Kober, Lisa Feldman Barrett, Josh Joseph, Eliza Bliss-Moreau, Kristen Lindquist, and Tor D Wager. Functional grouping and cortical–subcortical interactions in emotion: a meta-analysis of neuroimaging studies. *Neuroimage*, 42(2):998–1031, 2008.

[41] 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.

[42] Roland H Grabner and Bert De Smedt. Oscillatory eeg correlates of arithmetic strategies: a training study. *Frontiers in psychology*, 3:35080, 2012.

[43] Muhammed Shuaibi, Adeesh Kolluru, Abhishek Das, Aditya Grover, Anuroop Sriram, Zachary Ulissi, and C Lawrence Zitnick. Rotation invariant graph neural networks using spin convolutions. *arXiv preprint arXiv:2106.09575*, 2021.

[44] Tianle Cai, Shengjie Luo, Keyulu Xu, Di He, Tie-yan Liu, and Liwei Wang. Graphnorm: A principled approach to accelerating graph neural network training. In *International Conference on Machine Learning*, pages 1204–1215. PMLR, 2021.

[45] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture, 2020.

[46] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[47] Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE transactions on cybernetics*, 49(3):1110–1122, 2018.

[48] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[49] Xiao-Wei Wang, Dan Nie, and Bao-Liang Lu. Emotional state classification from eeg data using machine learning approach. *Neurocomputing*, 129:94–106, 2014.

[50] Lisa Feldman Barrett and Ajay Bhaskar Satpute. Large-scale brain networks in affective and social neuroscience: towards an integrative functional architecture of the brain. *Current opinion in neurobiology*, 23(3):361–372, 2013.