Attention-Based Temporal Graph Representation Learning for EEG-Based Emotion Recognition

Chao Li, *Member, IEEE*, Feng Wang, Ziping Zhao, *Member, IEEE*, Haishuai Wang, and Björn W. Schuller, *Fellow, IEEE*

***Abstract*—Due to the objectivity of emotional expression in the central nervous system, EEG-based emotion recognition can effectively reflect humans' internal emotional states. In recent years, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have made significant strides in extracting local features and temporal dependencies from EEG signals. However, CNNs ignore spatial distribution information from EEG electrodes; moreover, RNNs may encounter issues such as exploding/vanishing gradients and high time consumption. To address these limitations, we propose an attention-based temporal graph representation network (ATGRNet) for EEG-based emotion recognition. Firstly, a hierarchical attention mechanism is introduced to integrate feature representations from both frequency bands and channels ordered by priority in EEG signals. Second, a graph convolutional neural network with top-k operation is utilized to capture internal relationships between EEG electrodes under different emotion patterns. Next, a residual-based graph readout mechanism is applied to accumulate the EEG feature node-level representations into graph-level representations. Finally, the obtained graph-level representations are fed into a temporal convolutional network (TCN) to extract the temporal dependencies between EEG frames. We evaluated our proposed ATGRNet on the SEED, DEAP and FACED datasets. The experimental findings show that the proposed ATGRNet surpasses the state-of-the-art graph-based mehtods for EEG-based emotion recognition.**

***Index Terms*—Affective computing, attention mechanism, EEG, emotion recognition, graph convolution network.**

This work was supported in part by the National Natural Sci- ence Foundation of China under Grant 62071330, Grant 61702370, Grant 61902282, Grant 62322120, and Grant U21B2010 and in part by Tianjin Postgraduate Scientific Research Innovation Project under Grant 2022SKYZ267. *(Corresponding author: Ziping Zhao.)*

Chao Li, Feng Wang, and Ziping Zhao are with the College of Computer and Information Engineering, Tianjin Normal University, Tianjin 300387, China (e-mail: superlee@tjnu.edu.cn; waverlywang@qq.com; ztianjin@126.com).

Haishuai Wang is with the Department of Computer Science, Zhejiang University College, Hangzhou 310058, China (e-mail: haishuai.wang@gmail.co).

Björn W. Schuller is with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany, and also with GLAM, Imperial College London, SW7 2AZ London, U.K. (e-mail: schuller@ieee.org).

Digital Object Identifier 10.1109/JBHI.2024.3395622

I. Introduction

EMOTION, as a psychological and physiological state experienced by human beings, drives human cognition and behaviors [1]. Emotions are critical for interpersonal communication and mental-physical health, and they also substantially influence decision-making capabilities [2]. As a result, precise recognition of emotions is imperative for diverse applications, encompassing medical diagnosis, brain-computer interfaces, and education [3].

The original studies focus on emotion recognition through facial expressions, gestures, and speech. However, due to the influence of subjective consciousness, a person's inner emotional experience may not be accurately reflected in an assessment of their emotional state made based on external behaviors [4], which could confound the outcomes of affective computing. In comparison, physiological signals including electroencephalograms (EEG) [5], electrooculograms (EOG), and electromyograms (EMG), which originate from unconscious recordings in the central nervous system, can provide a more veridical representation of one's emotional state [6]. These physiological signals, being devoid of subjective modulation, serve as objective markers of human affect and provide a promising mehtod for emotion recognition.

As a type of neurophysiological signal, EEG reflects the electrical activity generated by the central nervous system and can be captured from the scalp [7]. Due to their high temporal resolution, low acquisition cost, and ability to reflect internal emotional states, researchers have increasingly turned to deep learning models to extract EEG features in order to recognize emotions in recent years [8]. For example, ACRNN is proposed as a method of extracting spatial and temporal features from raw EEG signals [9]. R2G-STNN are introduced as a means of learning discriminative spatio-temporal EEG features [10]. STRNN is designed to create a unified spatio-temporal dependency model by combining feature learning from the spatial and temporal information of signal sources [11]. Additionally, the TSception is designed to capture temporal dynamics and spatial asymmetry from EEG signals [12]. Furthermore, the 3DCANN model has been expressly designed to extract the dynamic interconnections between multi-channel EEG signals and their intrinsic spatial relations over continuous time periods [13]. To extract multivariate modulated oscillations for EEG emotion recognition, the proposed mehtod involves transforming EEG signals from the time-domain into time-frequency representation images, which are then fed into a pre-trained image classification network [14], [15]. The GLFANet enhanced the effectiveness of emotion recognition by focusing on both the topological and local features of EEG signals [16]. FLD3QN significantly improved the accuracy of valence and arousal recognition by simulating key neuroanatomical structures involved in reward learning [17]. Liu et al. employs coordinate attention and a pre-trained convolution capsule network to handle the complexity and diversity of emotion signals [18]. By incorporating a semantic-aware attention mechanism, a bidirectional recurrent prediction model has been proposed to explore the spatio-temporal and semantic relationships between attributes [19]. However, most of the above-mentioned methods only consider a single characteristic or a fusion of two attributes. In most cases, these methods disregard the complementarity of the spatial distribution data of electrodes, temporal dependencies, and frequency elements within EEG signals.

EEG signals inherently contain spatial information [20]. To acquire EEG signals, electrodes are positioned on the scalp to sense brain activity and transpose it into a signal. [21]. The spatial distribution of these electrodes determines the graphical structure of the signal, which may contain valuable information for emotion recognition [22]. Therefore, leveraging spatial information to improve EEG-based emotion recognition has emerged as an active research area. Recent studies have utilized deep learning methods based on graph neural networks (GNNs) to explore the spatial representation of EEG signals. DGCNN leverages graph modeling to capture features in multi-channel electroencephalographic signals and dynamically learn intrinsic relationships among different channels [23]. A regularized graph neural networks (RGNN) enhances the accuracy of emotion recognition by modeling the topological relationships between different brain regions and using two regularization methods to improve both the robustness and generalization performance [24]. The EEG-GCN method employs spatio-temporal adaptive graph convolutional networks, spatial attention mechanisms, and adaptive brain network adjacency matrices for single-view and multi-view EEG-based emotion recognition [25]. The graph structure of the self-organizing graph neural network (SOGNN) is dynamically constructed by self-organizing modules, and it has demonstrated excellent performance in cross-subject EEG-based emotion recognition [26]. A multi-domain fusion deep graph convolution neural network (MdGCNN) automatically extracts brain topology features by introducing graph convolution theory and functional connectivity [27]. CSGNN, an improved graph convolution method with dynamic channel selection, also effectively demonstrates emotion classification in complex dataset environments [28]. STGATE utilizes a transformer-encoder to learn time-frequency features and applies a spatial-temporal graph attention mechanism for emotion classification [29]. whereas the above-mentioned graph-based methods have demonstrated the ability to capture spatial representations of EEG signals, they may fail to fully consider the crucial spectral-temporal domain features. These features are critical in characterizing the dynamic patterns of brain activity over time; thus neglecting them can lead to incomplete or inaccurate analyses.

To address the existing challenges discussed above, we propose a hierarchical attention-based temporal graph representation network, named ATGRNet. In this work, we first utilize hierarchical attention mechanisms and graph neural networks to aggregate spatial and spectral features of EEG data. Furthermore, we employ TCN to extract temporal dependencies within EEG signals. The proposed method effectively captures spatial, spectral, and temporal features of EEG data, leading to significant improvements in emotion recognition performance. The primary innovations of this work can be summarized as follows:

1) We propose a hierarchical attention mechanism that dynamically assigns differential weights to integrate feature representations from both frequency bands and channels ordered by priority in EEG signals.
2) We further employ a graph convolutional neural network with top-k operation to capture internal relationships between EEG electrodes. Moreover, we introduce a residual-based graph readout mechanism to aggregate the EEG feature node-level representations into graph-level representations.
3) TCN is adopted to learn spatio-temporal features and extract the temporal dependencies between EEG frames efficiently.
4) The proposed ATGRNet fully capitalizes on the complementarity of spatial features, temporal dependencies, and frequency domain features to to enhance emotion recognition capabilities. The outcomes of experiments conducted on the SEED [30], [31] and DEAP [32] datasets exhibit the superiority and viability of the proposed mehtod.

The organization of the remainder of this paper is as follows. Section II reviews related works including those on feature extraction, graph neural networks, and temporal dependency. Section III first provides an overview of the proposed ATGRNet, followed by a delineation of the data preprocessing steps. Finally, we elaborate on our ATGRNet architecture. Section IV documents the outcomes of our experiments. Section V analyzes and discusses the proposed ATGRNet. To conclude, Section VI provides the final remarks and conclusions of this article.

II. Related Work

In this section, we review studies related to EEG-based emotion recognition, including works utilizing feature extraction, graph neural networks, and temporal dependency.

A. Feature Extraction

The time-domain waveform of raw EEG data is extremely complex and contains all information in both the temporal and frequency domains, which makes it difficult for models to extract and learn emotion-related features from this data. To address this challenge, researchers have explored frequency-domain [33] and time-frequency domain features [34]. These types of features can filter out redundant information and highlight emotion-related features, ultimately making modeling easier.

Data preprocessing, the first step in raw EEG signal feature extraction, includes noise removal [35], resampling, and baseline correction [9]. After basic data preprocessing, the EEG signal can be extracted into five frequency bands using short-time Fourier transform [36], [37], [38]: δ (1–4 Hz), θ (4– Hz), α (8–13 Hz), β (13–30 Hz), and γ (>30 Hz). Subsequently, common frequency–domain features can be extracted from these signals: examples include the differential entropy (DE) feature [31], [38], the power spectral density (PSD) feature [39], [40], the differential asymmetry (DASM) feature [41], the rational asymmetry (RASM) feature [42], and the differential caudality (DCAU) feature [30]. Research studies have demonstrated the effectiveness of these features for EEG emotion recognition.

B. Graph Neural Networks

Graph Neural Networks (GNNs) adapt deep learning to non-Euclidean graph data, extending convolutional concepts to graphs. Inspired by CNNs, Bruna et al. [43] linked spectral graph theory to neural networks, using spectral filters from the graph Laplacian to generalize convolutions to graphs. Recent research explores GNNs for EEG-based emotion recognition [23], [24], [25], [26], [27], utilizing the graph-like structure of EEG electrode distribution to model spatial dependencies. In this context, a graph $G = (V, E, A)$ represents EEG signals, where $V(|V| = n)$ are nodes (electrodes), $E(|E| = n^2)$ are edges, and $A \in \mathbb{R}^{n\times n}$ is the adjacency matrix. This mehtod adapts Fourier transform and convolution to graphs via the eigenvectors of the graph's Laplacian matrix.

The Laplacian matrix L of graph G can be calculated from $L = D - A$, where $D \in \mathbb{R}^{n\times n}$ is the diagonal matrix and each diagonal entry D_{ii} is is calculated as $D_{ii} = \sum_{j=1}^{n} A_{ij}$. By performing the singular value decomposition of the Laplacian matrix L, we can obtain the Fourier basis U of graph G as $L = U\Lambda U^T$, where $\Lambda \in \mathbb{R}^{n\times n}$ is the eigenvalue matrix of L. The graph Fourier transform $\hat{x}$ can be obtained by $\hat{x} = U^T x$ and its inverse is given by $x = U\hat{x}$. The convolution of two signals x_1 and x_2 on graph G, denoted as $G(x_1 * x_2)$, is defined as follows:

$$G(x_1 * x_2) = U\left[\left(U^T x_1\right) \odot \left(U^T x_2\right)\right], \tag{1}$$

where $\odot$ is the element-wise Hadamard product. This formula represents the convolution operation on a graph, where U is the matrix of eigenvectors of the graph Laplacian, and $U^T x_1$ and $U^T x_2$ are the signals x_1 and x_2 transformed into the spectral domain using the eigenvectors U. This means the filtering operation of signal x can be expressed as follows:

$$y = g(L)x = g\left(U\Lambda U^T\right)x = Ug(\Lambda)U^T x, \tag{2}$$

where $g(\Lambda) = diag([\theta_1, \theta_2, \ldots, \theta_n])$ is the defined filtering function, $g(\Lambda)$ is a function that applies the filter coefficients θ_i to the eigenvalues λ_i of L.

However, the filtering function g has high computational cost for large-scale graphs, scaling with complexity $\mathcal{O}(n^3)$. To tackle this problem, Defferrard et al. [44] developed a method for fast localized convolutions. To approximate filters, the researchers employed a recursive formulation of K-th-order Chebyshev polynomials, which enabled them to obtain a representation for each node by aggregating information from its K-th-order neighborhood. The K-th-order Chebyshev expansion can be expressed as follows:

$$g(\Lambda) = \sum_{k=0}^{K-1} \theta_k' \Lambda^k \simeq \sum_{k=0}^{K-1} \theta_k T_k(\widetilde{\Lambda}), \tag{3}$$

where θ_k is the coefficient of the Chebyshev polynomials and $\widetilde{\Lambda}$ denotes the normalized eigenvalues of the Laplacian matrix L. The recursive formula provided below can be used to calculate the Chebyshev polynomials $T_k(\widetilde{\Lambda})$:

$$\begin{cases} T_0(x) = 1, T_1(x) = x \\ T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x), \quad k \geq 2 \end{cases} \tag{4}$$

$$\widetilde{\Lambda} = 2\Lambda/\lambda_{\max} - I_n, \tag{5}$$

where $\lambda_{\max}$ denotes the maximum diagonal entry in Λ, and I_n is the $n \times n$ identity matrix. As mentioned above, the l th dynamic graph graph convolution is defined as follows:

$$X^l = \sigma\left(\sum_{k=0}^{K-1} \theta_k^l T_k(\hat{L}) X^{l-1}\right), \tag{6}$$

where θ_k^l is a trainable parameter, $\hat{L}$ is the normalized Laplacian matrix, and σ is an activation function. In addition to the network parameters, the adjacency matrix A is also optimized during the training process to learn the optimal matrix.

C. Temporal Dependency

EEG signals are time-sequential data, and it is essential to consider the temporal dependency in emotion recognition [45]. Recurrent neural networks (RNNs) and long short-term memory (LSTM) networks are two popular mehtods developed to handle the temporal dependency in time-series data [46]. RNNs utilize a recursive structure to model temporal dependencies, updating the hidden state at each time step using the prior hidden state and current input. However, RNNs suffer from the vanishing and exploding gradient problems when processing long sequences, which makes them less suitable for handling long-term dependencies [47]. LSTMs address vanishing gradients in traditional RNNs by incorporating a memory cell to retain long-term dependencies and gating units to modulate information flow. These gating mechanisms include input, output, and forget gates, which together enable LSTMs to learn longer-term dependencies effectively [48].

The Temporal Convolutional Network (TCN) [49] is a neural network architecture designed to process a long sequence of data. A TCN consists of dilated causal convolutions and residual connections, and has achieved a long effective history size. Recent research has shown that TCNs are more effective than both LSTMs [48] and gated recurrent units (GRUs) [50] when it comes to processing long sequence data. TCNs demonstrate superior capability to capture long-term dependencies between consecutive time frames when dealing with EEG signals, which

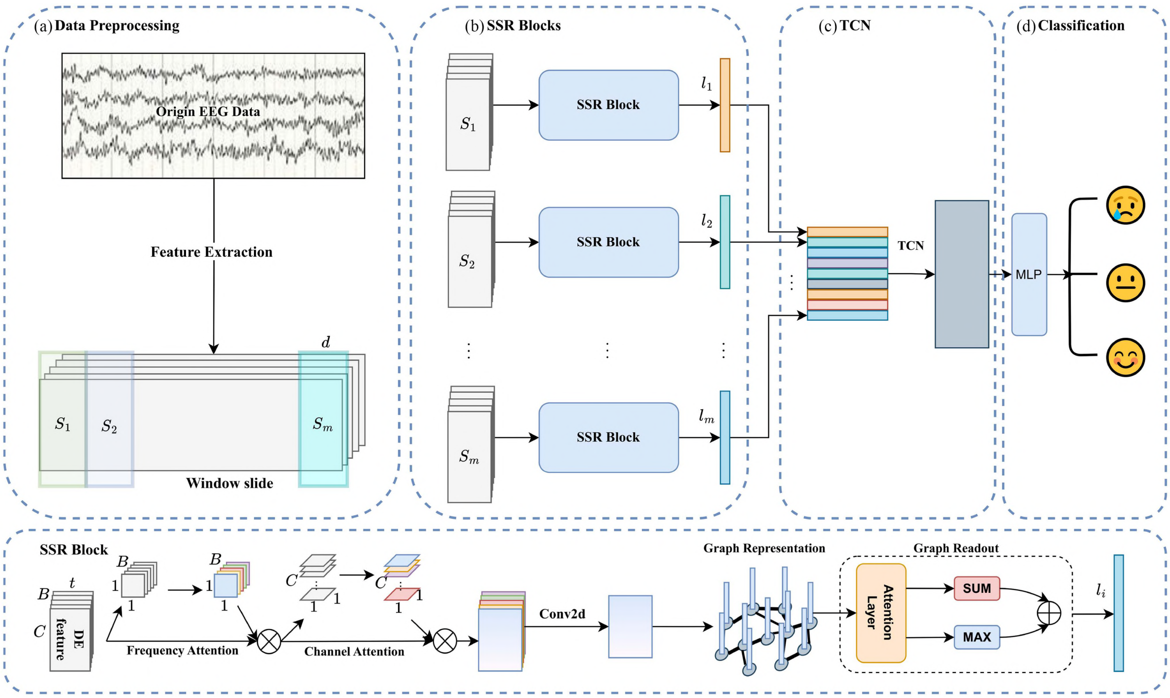

Fig. 1. Hierarchical attention-based temporal graph representation learning for EEG-based emotion recognition.

leads to significant improvements in emotion recognition accuracy. Furthermore, TCN offers the advantage of parallel computation, which substantially accelerates the processing speed. Therefore, we opt to utilize TCN due to its status as a promising tool for extracting the time dependency of EEG signals in the proposed ATGRNet. The TCN architecture can be represented as follows:

$$TCN = \text{1D FCN} + \text{causal convolutions}\,. \tag{7}$$

In order to capture distant past information, TCN incorporates dilated convolutions and residual layers. For a 1D input sequence $X = \{x_1, x_2, \ldots, x_T\}$ and filter $F = (f_1, f_2, \ldots, f_K)$, the dilated convolution of F at position t is defined as:

$$F(x_t) = \left(F_d^* X\right)(x_t) = \sum_{k=1}^{K} f_k x_{t-(K-k)d}, \tag{8}$$

where K denotes the filter size, d represents the dilation factor, and the expression $t-(K-k)d$ indicates the direction of the past. Formally, the TCN is represented as a function $\tau : X \rightarrow Y$, where $X = \{x_1, x_2, \ldots, x_T\}$ is the input sequence, K is the filter size, and d is the dilation. This function produces a mapping from the input sequence to the output sequence Y as follows:

$$\hat{y}_1, \ldots, \hat{y}_T = \tau\left(x_1, \ldots, x_T, k, d, A\right), \tag{9}$$

where $Y = \{\hat{y}_1, \ldots, \hat{y}_T\}$, T is the sequence length, and A controls the dimension of each TCN layer.

According to (8), the dilation factor d and the depth of the TCN determine the receptive field size (shown in Fig. 2). The generic residual block used in TCN (shown in Fig. 3) addresses the issues of vanishing gradients and exploding gradients that can arise when expanding shallow networks into deep networks.

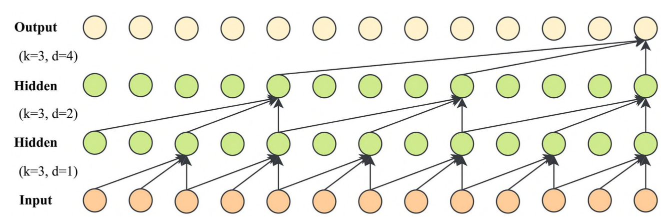

Fig. 2. Receptive field coverage of dilated causal convolution is affected by dilation factors d and filter size k.

III. Methodology

In this section, we give a brief overview of the proposed ATGRNet for EEG emotion recognition. Next, we introduce the EEG signal preprocessing method. Finally, we provide a comprehensive account of each module incorporated in the proposed ATGRNet.

A. Framework Overview

Generally, most existing EEG-based emotion recognition work [9], [10], [12], [13], [14], [15] focuses primarily on using single features or combinations of two features; as a result, these works overlook the complementarity of different domain features and are unable to learn spatial distribution information. Meanwhile, most graph neural network-based works [23], [24], [25], [26], [27], to some extent, overlook time-frequency domain

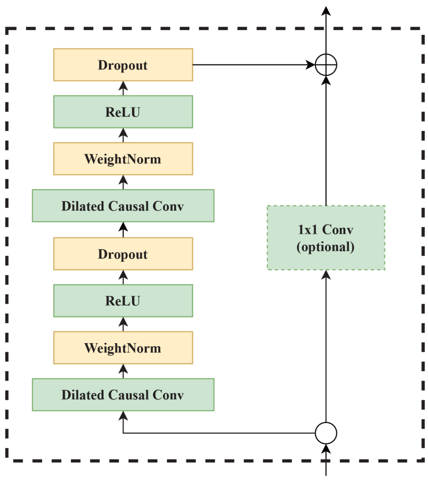

Fig. 3. TCN residual block diagram showing 1x1 convolution for dimensionality matching between residual input and output.

features to some extent. However, EEG recordings contain multidimensional data representing brain activity, including robust temporal, spatial, and spectral information. They may offer complementary perspectives that, when integrated, can allow for more comprehensive characterization of the affective states.

The proposed ATGRNet applies a hierarchical attention mechanism for spatial-frequency analysis, aggregates spatial data via a graph representation, and employs TCN for temporal integration. Its process begins with preprocessing EEG signals to extract and smooth features. These features are then processed through a hierarchical attention mechanism and a graph convolutional neural network to understand relationships between EEG electrodes and capture temporal dependencies. Emotion classification is done using a multilayer perceptron (MLP) [51], optimized with Adam and cross-entropy loss.

B. Data Preprocessing

Data preprocessing involves removing the baseline, extracting features, and sliding window operation. Typically, recorded EEG signals consist of both baseline signals and stimulus stage signals. To remove the baseline signals, an average is computed for the points in the baseline period and then subtracted from each point in the waveform [52].

In our experiment, the raw EEG signal is $X_R = [X_B, X_T]$, where $X_B \in \mathbb{R}^{C\times N_1}$ is the baseline signal, $X_T \in \mathbb{R}^{C\times N_2}$ is the stimulus stage signal, C is the number of EEG electrode nodes, and N_1, N_2 are the numbers of sampling points. Let H be the sampling frequency.

The mean value of the baseline per second can then be calculated via (10), where $\bar{X}_B \in \mathbb{R}^{C\times H}$ is the mean value of the baseline signal per second, and $T_1 = N_1/H$ is the number of seconds of X_B.

$$\bar{X}_B = \sum_{i=1}^{T_1} X_i/T_1. \tag{10}$$

We can represent the baseline-removed data using (11), where $X'_T = [X'_{T,1}, X'_{T,2}, \ldots, X'_{T,T_2}]$, and $T_2 = N_2/H$ denotes the number of seconds of X_T.

$$X'_{T,i} = X_{T,i} - \bar{X}_B. \tag{11}$$

After bandpass filtering preprocessing, five frquency-domain features are extracted with a window of 1 s on five frequency bands. The extracted features $X_E \in \mathbb{R}^{C'\times B\times F}$ can be expressed as follows:

$$X_F = \mathrm{FE}[\mathrm{BP}(X'_T)], \tag{12}$$

where $\mathrm{BP}(\cdot)$ denotes the bandpass filter, $\mathrm{FE}(\cdot)$ denotes the feature extraction operation, C' represents the new EEG channel after feature extraction, B represents the number of frequency bands, and F represents the extracted feature vector length.

To investigate the temporal dependence of EEG signals, the feature-extracted EEG data X_E is segmented into several temporal frames $X_S = \{X_1, X_2, \ldots, X_m\}$ using window slide processing without overlapping. Each window data frame $X_i (i = 1, 2, \ldots, m) \in \mathbb{R}^{C'\times B\times d}$ represents the i-th temporal slice, and m is the total number of windows. Here, $d = F/m$ represents the length of each window.

C. Proposed ATGRNet

The proposed ATGRNet comprises a hierarchical attention mechanism, a graph convolutional neural network with top-k operation, a graph readout module to aggregate spatial information, and a TCN module for temporal dependency extraction. The proposed ATGRNet architecture pipeline is depicted in Fig. 1.

1) Hierarchical Attention Mechanism: To determine the relative importance of the channels and frequency bands, we employ the SSR module for each segment X_i from the preprocessed EEG signal $X_S = \{X_1, X_2, \ldots, X_m\}$. A hierarchical attention mechanism is subsequently utilized to integrate feature representations from both frequency bands and channels ordered by priority, on the segment X_i.

We adopt squeeze-and-excitation blocks to implement the frequency band and channel-wise attention. Compared to Transformer self-attention, SE Layer is more computationally efficient, has fewer parameters, simpler structure, focuses on local correlations, and provides more interpretability of the attention weights. These characteristics make SE Layer well-suited for learning adaptive weighting of frequency bands and channels in our model. First, we apply average pooling for each frequency band to obtain a vector $a'_i = \mathrm{avgpooling}(X_i) \in \mathbb{R}^{1\times B}$.

A fully connected (FC) layer is then employed to compress the frequency bands with parameter w_1 and bias b_1, and the other FC layer is used to recover to the former dimension with parameter w_2 and bias b_2. Thus, the frequency bands attention score can be calculated as follows:

$$v'_i = \mathrm{softmax}(w_2 \cdot (ReLU(w_1 \cdot a'_i + b_1)) + b_2), \tag{13}$$

where $v'_i \in \mathbb{R}^{(1\times B)}$ denotes the attention mask across the frequency bands. w_1, w_2 are the weight matrices of the two fully connected (FC) layers, and b_1, b_2 are the corresponding biases. a'_i is the input to the frequency band attention module. Applying this mask yields the masked segment $X'_i = v'_i \otimes X^T_i$, with X'_i having identical dimensions to the original X_i.

Channel-wise attention is implemented in the same way as frequency bands attention. First, average pooling of each channel is $a''_i = avgpooling(X'_i) \in \mathbb{R}^(1 \times C')$. The channel-wise attention mask can be calculated as follows:

$$v''_i = \text{softmax}(w_4 \cdot (ReLU(w_3 \cdot a''_i + b_3)) + b_4). \quad (14)$$

Let w_3, w_4 be the parameters of the first and second FC layers respectively, and b_3, b_4 the corresponding biases. In the same way, the channel-wise attention masked segment can be represented as $X''_i = v'_i \otimes X'^T_i$ with the dimension $C' \times B \times d$.

2) Graph Representation Learning: To derive node-level representations from the EEG features, a Conv2d operation (15) is utilized to decrease the dimensionality of the feature map. The node-level representations aim to encode the feature information of each individual EEG channel node. In graph convolutional networks, these node representations serve as a medium for information propagation between neighboring nodes and for learning intrinsic relationships in the graph structure.

$$F_i = \text{Conv2d}(X''_i), \quad (15)$$

where $F_i \in \mathbb{R}^{C' \times d}$ denotes the feature map output from frequency bands and channels' attention. A graph convolutional neural network based on Chebyshev polynomials is then used to learn the graph representations of the new feature map F_i, which is used to integrate the information from each EEG channel related to emotion recognition. The K-th order Chebyshev polynomial graph layer can be represented as follows:

$$F'_i = \sigma\left(\sum k = 0^{K-1} \theta_k T_k(\hat{L}) F_i\right), \quad (16)$$

where $\widetilde{L} = 2\,L/\lambda_{\max} - I_n$ and $L = D - A$. The network parameters θ_k and adjacency matrix A are both trainable.

To emphasize edges exhibiting greater correlations, we applied a top-k technique that preserves only the k edges with the strongest connectivity throughout training. Since $|E| = N^2$, we set a parameter $\mathrm{k\ ratio}$ to determine the percentage of edges to be retained. Edges that are not retained are set to 1^{-10} to avoid computational issues. The application of the top-k operation is as follows:

$$\begin{cases} k = \lceil |E|/\text{k-ratio} \rceil \\ index = topk(A, k) \\ A[\overline{index}] = 1^{-10} \end{cases} \quad (17)$$

In the proposed ATGRNet, the graph layers aim to extract node-level representations from the input feature map, after which the readout layer is used to obtain its graph-level representation. The readout function is defined as follows:

$$G_i = \text{sigmoid}\left(F'_i \cdot w_5 + b_5\right) \odot \tanh\left(F'_i \cdot w_6 + b_6\right). \quad (18)$$

Equation (18) serves to transform node-level features utilizing soft attention weights and non-linear mapping, while w_5, w_6 denote the weights of linear transformations, and b_5, b_6 are biases. In order to aggregate information from all EEG channels, we use channel-wise average pooling to integrate the contribution of EEG electrodes to emotion recognition and maxpooling to capture the role played by critical EEG electrodes. To mitigate the vanishing gradient problem in deep learning networks, we introduced residual connections in the graph readout module, as shown in (19). The role of residual connections is crucial. Qin et al. employed three deep residual connections, allowing the network to focus on the residual mapping when learning the mapping, instead of directly learning the more complex original mapping, thereby simplifying the optimization objective and alleviating the gradient vanishing issue [53].

$$T_i = \frac{1}{|V|}\sum_{j=1}^{|V|} G_{ij} + \text{Maxpooling}\left(G_{i1} \ldots G_{i|V|}\right) + F_i. \quad (19)$$

After the operations described above are complete, we obtain a high-dimensional feature $T_i \in \mathbb{R}^{1\times l}$ for each time window. The proposed hierarchical attention architecture, encompassing mechanisms for EEG frequency bands and channels, graph representation learning, and graph readout, extracts high-dimensional EEG signal features across both spatial and frequency domains. The combination of these methods is referred to as the spatial-spectral representation (SSR) extraction module. The SSR modules can be represented as follows:

$$T = \text{SSRs}(X_S), \quad (20)$$

where $T = \{T_1, T_2, \ldots, T_m\}$.

3) Temporal Dependency Extraction: The TCN architecture has proven effective for EEG emotion recognition, owing to its capacity for learning temporally dependent features from EEG data [54]. Given that EEG data processed with SSR modules can be treated as sequential data, TCN is well-suited for extracting high-level emotion-related features. Given an input EEG feature sequence $T = T_1, T_2, \ldots, T_m$ generated by the SSR module, with a filter size K and dilation factor d, the TCN operation in ATGRNet can be defined as:

$$S = \tau(X, k, d, N), \quad (21)$$

where $S \in \mathbb{R}^{m\times r}$ is the output of TCN and $r = N[-1]$ denotes the dimension of the last layer of the TCN.

To obtain predicted emotions, a classifier is used to enhance the model's discriminability. The feature map S obtained from the TCN needs to be flattened before it is input into the classifier: $H = \text{flatten}(S)$. Equation (22) describes the architecture of the classifier.

$$\hat{y} = \text{softmax}(\sigma(\text{MLP}(H))). \quad (22)$$

The loss function utilized for parameter optimization consists of two components: a cross-entropy loss term for the classification outputs and an L2 regularization term weighted by α. The formula of the loss function is shown below; here, y is the label of the input sample, $\hat{y}$ is the predicted label, and θ denotes all parameters of ATGRNet.

$$Loss = \text{cross-entropy}\left(y, \hat{y}\right) + \alpha\|\theta\|_2^2. \quad (23)$$

IV. Experimental Results

In this section, we first introduce the two datasets used in our experiments (SEED [30], [31] and DEAP [32]), both of which are commonly leveraged for EEG-based emotion recognition research. The model configuration is then described. Subsequently, We present the results of the subject-dependent (SD) and subject-independent (SI) of the proposed ATGRNet on these two datasets. Finally, we analyze and discuss the experimental results.

A. Datasets

The SEED database includes EEG recordings from 15 subjects (seven males, eight females) as they viewed 15 film clips eliciting positive, neutral and negative emotions. Clip duration was restricted to 4 minutes to preclude subject fatigue. EEG signals were acquired from 62 electrodes positioned per the 10–20 system [55] and downsampled to 200 Hz. The EEG recordings corresponding to each stimulus were labeled as positive, neutral or negative based on the elicited emotion. Data were collected over three sessions per subject, with each session comprising 15 EEG trials, yielding 45 trials total per subject. To validate concordance between the presented clips and subjects' emotional states, self-reported assessments were administered after each session.

The DEAP [32] database collected multimodal physiological signals from 32 subjects (16 females and 16 males) who watched music videos limited to 1 minute in length. The sampling rate of recorded physiological signals was 512 Hz with 32 electrodes, down-sampled to 128 Hz. The data were annotated with the subjects' self-reported scores from 1-9 on arousal, valence, liking and dominance, obtained using the self-assessment manikins (SAM) technique. To facilitate binary classification, trials with scores above 5 were labeled as positive samples, while those with scores of 5 or below were labeled as negative samples.

The FACED dataset (Finer-grained Affective Computing EEG Dataset) [56] comprises EEG activity from 123 subjects recorded across 32 channels. These subjects viewed 28 videos, each 30 seconds long, designed to evoke one of nine emotion types. The dataset categorizes emotions into four positive (amusement, motivation, happiness, tenderness), four classical negative (anger, disgust, fear, sadness), and neutral.

B. Experiment Settings

The Adam optimizer was used to minimize the loss function in (23) for training ATGRNet, with a learning rate of 0.0001. Furthermore, a learning rate scheduler and an early stopping strategy were employed to adapt the training approach based on validation set performance. Specifically, the learning rate scheduler dynamically adjusted the learning rate, decreasing it by a factor if training metrics plateaued over epochs. Additionally, early stopping monitored validation set metrics, halting training if no improvement occurred within a defined number of epochs. The model was set to run for a maximum of 200 epochs, with a batch size of 16, a regularization coefficient of $\alpha = 0.0001$, and a division of 12 time window frames. The experiment was run in the Ubuntu 18.04 evironment of Python 3.8.8, Pytorch 1.9.1 on a workstation with an Intel Xeon Silver CPU and a NVIDIA RTX TITAN GPU.

To evaluate the proposed ATGRNet, we conducted SD and SI experiments, with some variations in the experimental settings between the SEED, DEAP, and FACED datasets.

In the SD experiment on SEED, we selected two sessions of a subject for training and one session for testing. Each subject was tested three times, after which the average result was computed as the SD experimental result. The final SD result of the ATGRNet was obtained by averaging the results of 15 subjects. For the subject-independent (SI) experiments on SEED, leave-one-subject-out (LOSO) cross-validation was used, where one subject's EEG data served as the test set and all other subjects' data comprised for training. The final SI result was obtained by averaging the results of all tests.

For the SD experiment with the DEAP dataset, a ten-fold cross-validation was employed. This involved splitting the 40 samples per participant into 10 equal groups, each time using 36 samples for training and 4 for testing. In the SI experiment, aligned with the SEED dataset protocol, we treated each subject's data as a separate test set, resulting in 32 individual experiments. The final classification result for the SI experiment is the average of these 32 trials.

Given the extensive subject numbers in the FACED dataset, we employed a 10-fold cross-validation approach in the SI experiment to evaluate ATGRNet's performance. In addition, 5-fold and 3-fold cross-validation experiments were also conducted. For the SD experiment, we utilized LOSO experiment protocol. The final classification accuracy was determined by averaging the LOSO accuracies across all 123 subjects.

C. SI and SD Experiments

To assess ATGRNet's performance, we compared it against several baselines, including traditional machine learning algorithms like SVM [30] and DBN [57], deep learning methods incorporating graph neural networks, such as DGCNN [23] and SOGNN [26], and other recent models, including, RGNN [24], MdGCNN [27], and IAG [58]. We conduct SI LOSO cross-validation experiments on the SEED dataset and summarize the results in Table I. The mean emotion recognition accuracy and standard deviation (STD) obtained by ATGRNet across the five frequency bands are reported. For a single-band experiment, the SSR module in the proposed ATGRNet removes the band attention. For experiments across all bands, the complete ATGRNet architecture was utilized for training and validation, with mean classification accuracy and standard deviations computed. Notably, our proposed model outperforms all other comparison baselines. Specifically, our model achieves a classification accuracy 0.74% higher than SOGNN whereas maintaining a lower standard deviation (5.25). The main factors contributing to this improvement are as follows: 1) the hierarchical attention mechanism highlights the contribution of certain frequency bands and channels to emotion generation; 2) the use of graph readout preserves node features whereas capturing the role of critical

TABLE I
COMPARISON OF MEAN ACCURACIES (%) AND STD OF SI EXPERIMENT (ACC/STD) ON THE SEED DATASET USING TEN CLASSIFIERS: SVM, DBN, EEG_GCN, DGCNN, RGNN, MDGCNN, IAG, SOGNN, TMLP+SRDANN, GRU-CONV AND ATGRNET

Method	δ **band**	θ **band**	α **band**	β **band**	γ **band**	**all**(δ,θ,α,β,γ)
SVM [30]	43.06/08.27	40.07/06.50	43.97/10.89	48.64/10.29	51.59/11.83	56.73/16.29
DBN [57]	57.12/13.22	65.94/14.41	72.40/16.46	67.22/11.71	64.28/10.96	58.50/10.94
EEG_GCN [25]	-	-	-	-	-	77.30/08.21
DGCNN [23]	49.79/10.94	46.36/12.06	48.29/12.28	56.15/14.01	54.87/17.53	79.95/09.02
RGNN [24]	64.88/06.87	60.69/5.79	60.84/7.57	74.96/8.94	77.50/08.10	85.30/06.72
MdGCNN [27]	56.63/6.99	61.64/03.57	64.75/04.24	79.35/06.98	85.40/02.13	83.72/02.94
IAG [58]	-	-	-	-	-	86.30/06.91
SOGNN [26]	70.37/07.68	76.00/6.92	66.22/11.52	72.54/08.97	71.70/08.03	86.81/05.79
TMLP+SRDANN [59]	-	-	-	-	-	81.04/06.28
GRU-Conv [60]	-	-	-	-	-	87.04/13.35
ATGRNet	**81.93/04.93**	**83.41/05.73**	**79.85/03.08**	**85.33/05.31**	**87.56/05.43**	**87.55/05.25**

The bold values represent the maximum value in the current column.

TABLE II
COMPARISON OF MEAN ACCURACIES (%) AND STD OF SI EXPERIMENT (ACC/STD) ON THE SEED DATASET FOR FIVE DIFFERENT FEATURES USING FIVE CLASSIFIERS: SVM, DBN, EEG_GCN, DGCNN, RGNN, AND ATGRNET

Features	DE	PSD	DASM	RASM	DCAU
SVM [30]	48.58/4.24	48.57/4.24	56.56/9.22	56.56/9.22	56.56/9.22
DBN [57]	51.64/7.84	52.82/8.71	51.88/9.67	52.56/9.31	48.60/10.71
DGCNN [23]	79.95 / 9.02	**64.27 / 13.80**	52.50 / 11.92	58.46 / 10.08	65.19 / 10.49
ATGRNet	**87.55/05.25**	59.37/07.11	**77.71/05.69**	**80.59/4.01**	**81.63/4.83**

The bold values represent the maximum value for accuracy in the current column.

TABLE III
COMPARISON OF MEAN ACCURACIES (%) AND STD OF SD EXPERIMENT (ACC/STD) ON THE SEED DATASET USING SIX CLASSIFIERS: SVM, DBN, EEG_GCN, DGCNN, RGNN, MDGCNN, AND ATGRNET

model	SVM [30]	DBN [57]	EEG-GCN [25]	DGCNN [23]	MdGCNN [27]	ATGRNet
acc/std	83.99 / 9.72	86.08 / 8.34	85.65/7.49	90.40/8.49	91.54/3.99	**92.59/8.73**

The bold values represent the maximum value in the current row.

TABLE IV
COMPARISON OF MEAN ACCURACIES (%) AND STD OF SD EXPERIMENT (ACC/STD) ON THE DEAP DATASET FOR TWO EMOTION CLASSES (NEGATIVE AND POSITIVE) AND TWO DIMENSIONS(VALENCE AND AROUSAL) USING FIVE CLASSIFIERS: SVM, DBN, DGCNN, TMLP+SRDANN, AND ATGRNET

Models	SI		SD	
	valence	arousal	valence	arousal
SVM [30]	48.58/4.24	50.75/4.87	51.60/6.32	55.80/9.49
DBN [57]	51.64/7.84	56.68/13.28	51.58/6.35	62.43/11.87
DGCNN [23]	58.46/7.85	61.65/13.34	**86.32/6.04**	**83.68/5.68**
TMLP+SRDANN [59]	57.70/7.23	61.88/5.55	-	-
ATGRNet	**68.65/9.09**	**68.75/7.85**	78.22/18.33	76.46/19.48

The bold values represent the maximum value in the current column.

electrodes; 3) the top-k operation enables the network to emphasize connections with greater correlations, underscoring the influence of inter-regional connections on emotion recognition.

On the SEED dataset, we also conducted sentiment classification experiments using different features, and the results of which are shown in Table II. Our proposed ATGRNet achieved the highest accuracy across all features. Among the five extracted EEG features, the DE feature showed the highest recognition accuracy, which is consistent with some previous studies. Notably, the accuracy of ATGRNet was on average 7.60% higher than that of DGCNN and the standard deviation was also lower, indicating that the proposed method achieved better stability. Furthermore, as can be seen from Table III, our model outperformed other graph convolutional neural network methods in classifying the data in the SD experiment. Overall, We achieved a notable increase in accuracy compared to baseline models.

We conducted SI and SD experiments on the DEAP dataset and present the results in Table IV. In the SI experiments, our classification accuracies for valence and arousal were 68.65% and 68.75%, respectively. Compared to DGCNN, our model achieved higher accuracy, with an increase of 10.19% for valence and 7.10% for arousal. However, in SD experiments, the

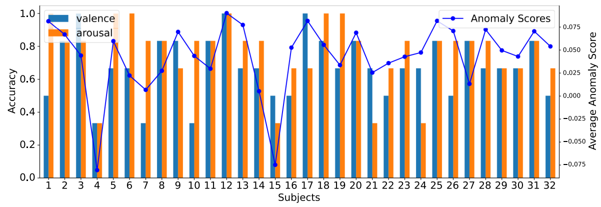

Fig. 4. SD classification accuracy for valence and arousal labels and average anomaly score per subject on the DEAP dataset.

TABLE V
COMPARISON OF MEAN ACCURACIES (%) AND STD OF SD EXPERIMENT (ACC/STD) ON THE FACED DATASET FOR TWO EMOTION CLASSES (NEGATIVE AND POSITIVE) AND TWO FEATURES(DE AND PSD) USING FIVE CLASSIFIERS: SVM, KNN, DBN, DGCNN, AND ATGRNET

Models	SI		SD	
	DE	PSD	DE	PSD
SVM [30]	69.30/01.50	54.97/05.67	78.80/01.00	67.44/07.62
KNN [62]	52.44/10.33	50.19/08.79	65.71/12.28	61.15/09.69
DBN [57]	60.32/14.48	52.63/07.13	70.25/11.07	65.64/17.49
DGCNN [23]	70.94/03.79	53.58/03.40	71.28/14.60	66.75/15.07
ATGRNet	**75.76/05.58**	**67.33/09.47**	**87.97/13.99**	**75.83/20.41**

The bold values represent the maximum value in the current column.

performance of our model was better than traditional models but slightly lower than DGCNN. It is noteworthy that the standard deviation of these results was relatively high, which implies a significant degree of variability among subjects. From the Fig. 4 , it is evident that some subjects have substantially lower classification accuracy than others. We hypothesized that some subjects in the DEAP dataset might show lower data quality and more variable distributions as outliers. To validate this, we used outlier detection experiments employing the isolation forest algorithm, a proven method for identifying outliers [61]. This approach is based on the idea that outliers, due to their rarity, are easier to isolate compared to regular data points. The procedure began by normalizing EEG DE features for all subjects, then constructing an isolation forest model. This model was created using randomly selected data subsets, forming several isolation trees. Each tree split the data by randomly selecting features and cut-off values until a specific criterion was met. We calculated the average path length for each subject in the isolation forest to determine an anomaly score. To account for the effects of random data selection, multiple experiments were conducted with cross-validation, averaging anomaly scores per subject for accurate outlier identification.

Our experimental results, depicted in Fig. 4, show that subjects 4, 7, and 15, with lower classification accuracies, had the lowest anomaly scores. Subjects like 21 and 24, despite having slightly higher anomaly scores, still fell below the majority, aligning with their lower accuracies. This trend highlights how outlier anomalies affect overall model performance and supports our hypothesis about the impact of data quality and distribution variability on effectiveness.

Our ATGRNet model was benchmarked against SVM, KNN, DBN, and DGCNN on the FACED dataset (see Table V). The results demonstrate that ATGRNet surpasses these methods

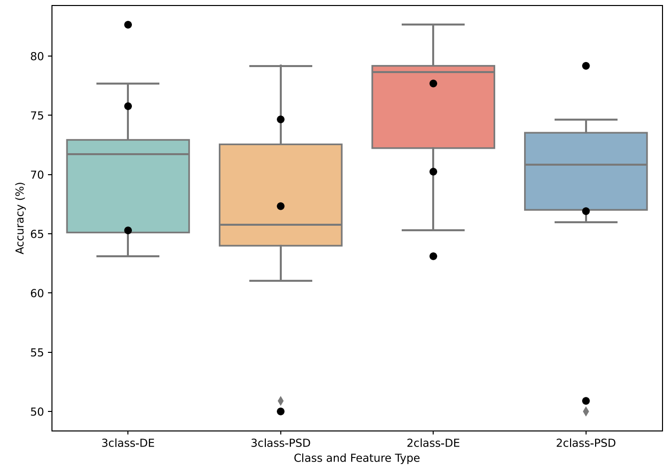

Fig. 5. SI classification accuracy for 2-class and 3-class on DE and PSD features on the FACED dataset.

in both SI and SD scenarios. In the SI experiment, ATGRNet achieved average accuracies of 75.76% (DE features) and 67.33% (PSD features), with standard deviations of 5.58 and 9.47, respectively. These figures not only reflect higher accuracy but also greater stability compared to other methods. In the SD experiment, ATGRNet's performance further excels, achieving 87.97% accuracy for DE features and 75.83% for PSD features, significantly outperforming the compared methods.

The aforementioned results were based on the dichotomous label setting of the FACED dataset. We also extended our experiments to a 3-class classification task (refer to Fig. 5), utilizing 10-fold cross-validation. In the 2-class task, ATGRNet achieved an average accuracy of 75.76% (DE features) and 67.33% (PSD features), with standard deviations of 5.58 and 9.47, respectively. For the 3-class task, the model recorded average accuracies of 70.24% (DE features, STD: 5.31) and 67.11% (PSD features, STD: 9.47). The model demonstrated notably stable and superior performance on DE features. The variation in classification accuracies across different tasks is depicted in a box plot. ATGRNet performs well in binary and tertiary emotion classification with DE features, while PSD features show variability in binary tasks, due to EEG signals' spectral complexity in diverse emotions. The accuracy decline in three-class tasks suggests added difficulty in discerning emotions, particularly neutral ones.

V. DISCUSSION

In this section, we first conduct an ablation study to verify the validity of each component of the model. Next, ATGRNet is evaluated using a confusion matrix to assess metrics including accuracy and recall. Subsequently, the model's performance is analyzed by varying the slide window size and compression coefficients in the SSR module. Lastly, t-SNE analysis is utilized to assess the model's feature extraction capabilities.

A. Ablation Study

We conducted an ablation study to investigate the contribution of each key component in our proposed ATGRNet. The

TABLE VI
ABLATION STUDY FOR SI AND SD CLASSIFICATION ACCURACY (MEAN/STD) ON SEED AND DEAP

Model	SEED(SD)	SEED(SI)	DEAP(SD)	DEAP(SI)
ATGRNet	**92.59/8.73**	**87.55/05.25**	**78.22/18.33**	**68.65/9.09**
–band attention	84.34/6.72	85.36/4.33	67.29/8.55	59.69/8.32
–channel-wise attention	85.77/8/23	82.58/5.41	67.38/6.83	62.35/5.67
–readout layer	88.54/9.47	81.17/7.39	66.70/8.27	63.38/7.56
–top k operation	90.29/5.56	85.59/6.62	67.38/8.55	65.58/5.21
–TCN(relpaced with LSTM)	88.32/9.44	83.57/8.61	68.46/7.46	66.67/9.44

Symbol "–" indicates the following component is removed.
The bold values represent the maximum value in the current cloumn.

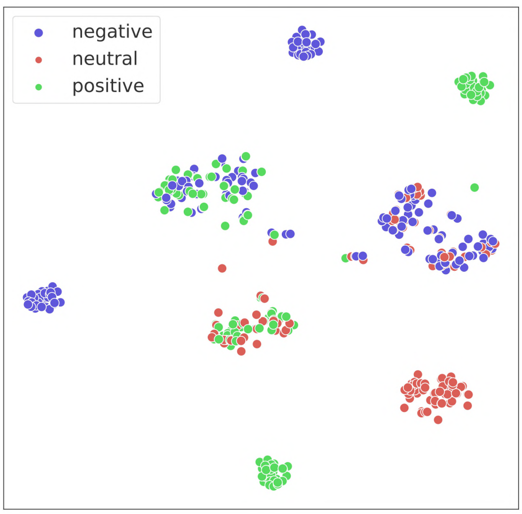

Fig. 6. t-SNE visualization of EEG feature representations in the SEED dataset before training.

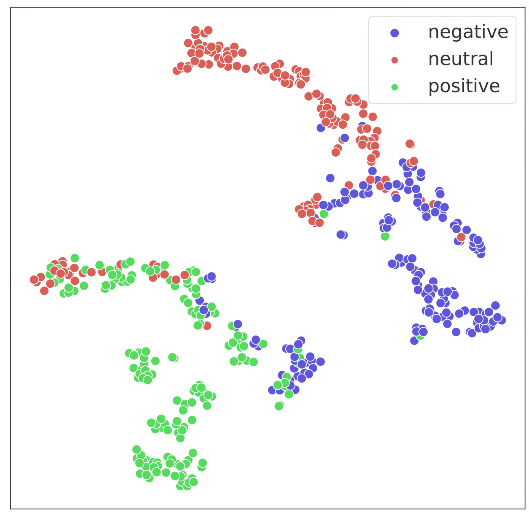

Fig. 7. t-SNE visualization of EEG feature representations in the SEED dataset after training.

SD classification results on the SEED dataset are reported in Table VI. Regarding the removal of components, we observed a decrease in classification accuracy when both band attention and channel-wise attention were removed; this suggests that the attention mechanism effectively removes redundant information from EEG signals. Additionally, the results indicate that the graph readout mechanism is more effective at extracting spatial information from the graph compared to dimensional transformation. We further compared the top-k adjacency matrix with the adjacency matrix without connection removal and found that the top-k adjacency matrix performed better. Specifically, the top-k operation allows a network to concentrate on electrodes with higher relevance; it disregards insignificant connectivity patterns and eliminates redundant information, leading to improved classification accuracy. Additionally, the TCN module was substituted with LSTM under identical experimental conditions to validate the efficacy of TCN. Our experimental results demonstrate that TCN exhibits stronger time-dependent feature extraction capabilities compared to LSTM in this scenario.

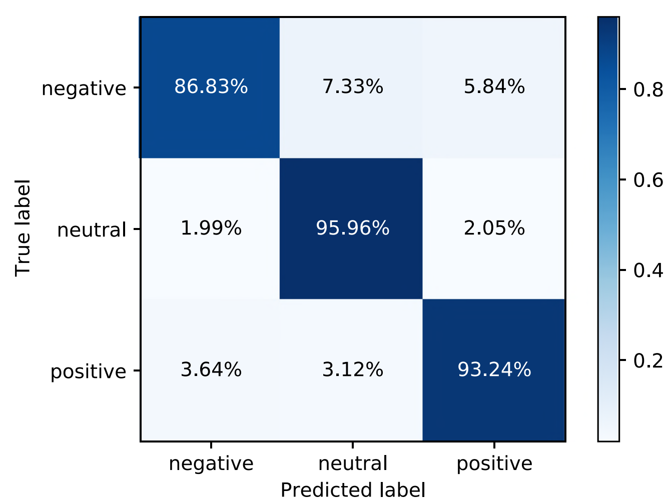

Fig. 8. Confusion matrices for SD experiment using the proposed ATGRNet on the SEED dataset.

B. Confusion Matrix

The proposed ATGRNet's performance on the SEED dataset was evaluated using a confusion matrix and the results are shown in Figs. 8 and 9. The SD experiments achieved higher classification accuracy than the SI experiments for negative, neutral, and positive emotions, with accuracies of 86.83%, 95.96%, and 93.24%, respectively. This difference in accuracy may be attributed to inter-individual differences that challenge the model's generalization ability, causing the model to learn specific emotional patterns of individual subjects in the SD

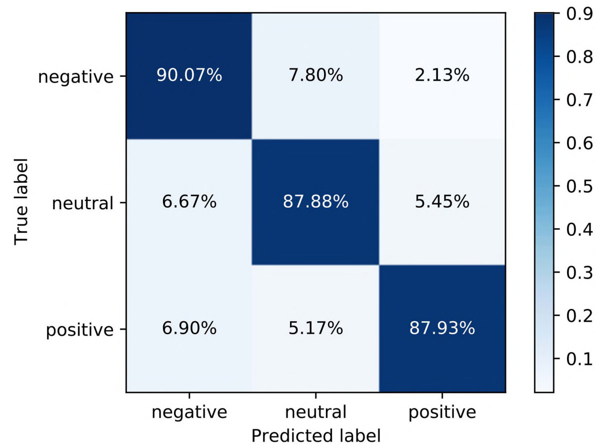

Fig. 9. Confusion matrices for SI experiment using the proposed ATGRNet on the SEED dataset.

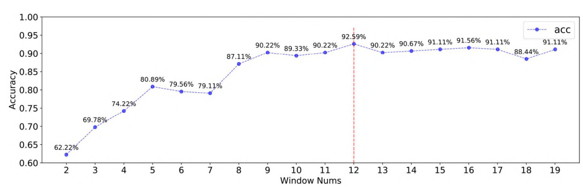

Fig. 10. Classification accuracy of our proposed ATGRNet under various window sizes in SD experiments on the SEED dataset.

experiments. For neutral classification results, SD experiments results were much higher than SD experiments results, but the opposite was true for negative results. This may be because the model cannot rely on specific personal information in SD experiments and thus loses some useful information. However, the characteristics of negative emotions may exhibit some commonality among different subjects, and the model captures this commonality.

C. Parameter Optimization

We found that the sliding window length and the k-ratio of the top-k operation had a significant effect on the experimental results.

The model's performance was analyzed by varying the window slide number and compression coefficients in the SSR module, as illustrated in Fig. 10. For the SEED dataset, increasing the sliding window numbers from 5 to 12 generally increases the model performance. However, as the number of windows continues to increase, the performance of the model decreases significantly. Our hypothesis is that increasing the number of time windows enables TCN to extract more time-related features. However, this may limit the spatial information input into each SSR module, resulting in reduced spatial information extraction. Conversely, fewer time windows allow for greater spatial information input into each SSR module, but may hinder time-related feature extraction. Thus, a balance must be struck between these factors. Through experimentation, we found that a window size of 23 and dividing the signal into 12 windows resulted in the highest classification accuracy. This optimal balance between temporal and spatial feature extraction

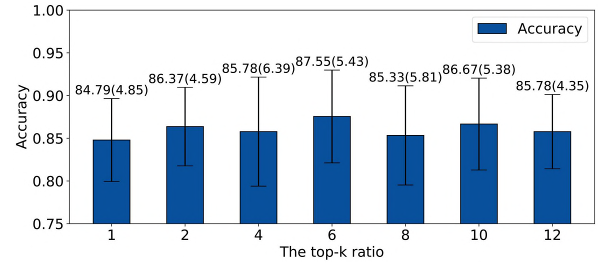

Fig. 11. Classification accuracy of our proposed ATGRNet under various k-ratio in SD experiments on the SSED dataset.

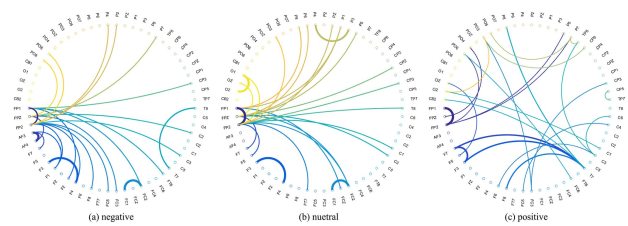

Fig. 12. EEG adjacency matrix visualization in the SEED dataset on SI experiment.

enhances our proposed ATGRNet's classification performance on the SEED dataset.

The k-ratio is a parameter that determines the number of edges removed in the top-k operation. Increasing the k-ratio removes more irrelevant connections, but may also result in the loss of useful information for emotion recognition tasks. We conducted experiments with different k-ratios ranging from 1 to 12 the present the results in Fig. 11. We found that setting the k-ratio to 6 resulted in the best performance. This suggests that a k-ratio of 6 effectively eliminates unnecessary connection information in EEG graph structural data whereas valuable information for emotion recognition tasks is retained.

D. Visualization

Relational ring diagrams were employed to illustrate the EEG electrode connections across different emotional states. These visualizations focus on the top-k connection nodes with higher weights in negative, neutral, and positive emotions (see Fig. 12).

In the negative emotion state, strong connections were identified between electrodes FP1, FPZ, F3, F4, FC1, FC2, T7, and T8, located in the frontal and temporal lobes, which are crucial for emotion processing. The frontal lobe activity is significantly associated with the processing of negative emotions such as sadness or fear [63]. Simultaneously, the temporal lobes (T7, T8) are implicated in emotional memory and recognition, suggesting their enhanced role in negative states. In neutral emotional states, strong connections were observed between occipital (O1, O2) and frontal electrodes, indicating a mix of visual information processing and emotion regulation [64]. For positive emotions, the pattern shifted to strong connections among FP1, FP2, F7, F3, and T7, covering various frontal and temporal regions. This suggests these areas' involvement in the cognitive processing and emotional memory associated with positive emotions [65].

The results highlight distinct patterns of brain network activity across various emotional states. Visualizing the adjacency matrix of the proposed method enabled us to discern the linkage between emotional states and regional brain activities. These observations are consistent with previous neuroscience research [66], [67], [68].

To evaluate the model's feature extraction ability, we used t-SNE to analyze the features extracted by the model. We utilized the features from the last layer of the FC layer of the model for t-SNE clustering analysis, and present the results in Figs. 6 and 7. The t-SNE visualization results indicate that most samples are well-clustered according to their respective categories. However, there are a few samples from different categories that lack clear boundaries between them. This may result from individual variances in emotional expression, which can impact feature extraction quality.

VI. Conclusion

This paper presents a hierarchical attention-based temporal graph representation network (ATGRNet) for EEG emotion recognition. The ATGRNet was motivated by neuroscience theories stipulating high correlation between EEG signals from specific brain regions and particular emotions. Based on these theories, our model fully exploits the complementarity among spatial features, temporal dependencies, and frequency domain features in EEG signals. We subsequently validated the effectiveness of the proposed hierarchical multi-attention mechanism, graph representation learning, and TCN in EEG-based emotion recognition. Extensive evaluations on two public datasets exhibit the proposed ATGRNet architecture outperforming baseline methods.

References

[1] F. Schoeller, A. Haar, A. Jain, and P. Maes, "Enhancing human emotions with interoceptive technologies," *Phys. Life Rev.*, vol. 31, pp. 310–319, 2019.

[2] T. B. Norboevich, "Analysis of psychological theory of emotional intelligence," *Eur. J. Res. Reflection Educ. Sci.*, vol. 8, no. 3, pp. 99–104, 2020.

[3] E. Yadegaridehkordi, N. F. B. M. Noor, M. N. B. Ayub, H. B. Affal, and N. B. Hussin, "Affective computing in education: A systematic review and future research," *Comput. Educ.*, vol. 142, 2019, Art. no. 103649.

[4] L. Li and J.-h. Chen, "Emotion recognition using physiological signals," in *Proc. 16th Int. Conf. Artif. Reality Telexistence*, Hangzhou, China, 2006, pp. 437–446.

[5] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors*, vol. 20, no. 3, 2020, Art. no. 592.

[6] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 374–393, Jul.–Sep. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/7946165/

[7] K. Kamble and J. Sengupta, "A comprehensive survey on emotion recognition based on electroencephalograph (EEG) signals," *Multimedia Tools Appl.*, vol. 82, no. 18, pp. 27269–27304, 2023.

[8] M. Jafari et al., "Emotion recognition in EEG signals using deep learning methods: A review," *Comput. Biol. Med.*, vol. 165, 2023, Art. no. 107450.

[9] W. Tao et al., "EEG-based emotion recognition via channel-wise attention and self attention," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 382–393, Jan.–Mar. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9204431/

[10] Y. Li, W. Zheng, L. Wang, Y. Zong, and Z. Cui, "From regional to global brain: A novel hierarchical spatial-temporal neural network model for EEG emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 568–578, Apr.–Jun. 2022.

[11] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial–temporal recurrent neural network for emotion recognition," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 839–847, Mar. 2019.

[12] Y. Ding, N. Robinson, S. Zhang, Q. Zeng, and C. Guan, "TSception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2238–2250, Apr.-Sep. 2023.

[13] S. Liu et al., "3DCANN: A spatio-temporal convolution attention neural network for EEG emotion recognition," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 11, pp. 5321–5331, Nov. 2022.

[14] P. V. and A. Bhattacharyya, "Human emotion recognition based on timefrequency analysis of multivariate EEG signal," *Knowl.-Based Syst.*, vol. 238, Feb. 2022, Art. no. 107867.

[15] S. K. Khare and V. Bajaj, "Time-frequency representation and convolutional neural network-based emotion recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 2901–2909, Jul. 2021.

[16] S. Liu, Y. Zhao, Y. An, J. Zhao, S.-H. Wang, and J. Yan, "Glfanet: A global to local feature aggregation network for eeg emotion recognition," *Biomed. Signal Process. Control*, vol. 85, 2023, Art. no. 104799.

[17] D. Li, L. Xie, Z. Wang, and H. Yang, "Brain emotion perception inspired EEG emotion recognition with deep reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 2023, doi: 10.1109/TNNLS.2023.3265730.

[18] S. Liu, Z. Wang, Y. An, J. Zhao, Y. Zhao, and Y.-D. Zhang, "Eeg emotion recognition based on the attention mechanism and pre-trained convolution capsule network," *Knowl.-Based Syst.*, vol. 265, 2023, Art. no. 110372.

[19] C. Li, Z. Bao, L. Li, and Z. Zhao, "Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multimodal emotion recognition," *Inf. Process. Manage.*, vol. 57, no. 3, 2020, Art. no. 102185.

[20] F. Li et al., "Differentiation of Schizophrenia by combining the spatial EEG brain network patterns of rest and task P300," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 4, pp. 594–602, Apr. 2019.

[21] Y. Yin, X. Zheng, B. Hu, Y. Zhang, and X. Cui, "EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM," *Appl. Soft Comput.*, vol. 100, 2021, Art. no. 106954.

[22] P. A. Kragel and K. S. LaBar, "Decoding the nature of emotion in the brain," *Trends Cogn. Sci.*, vol. 20, no. 6, pp. 444–455, Jun. 2016. doi: 10.1016/j.tics.2016.03.011.

[23] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 532–541, Jul.–Sep. 2020.

[24] P. Zhong, D. Wang, and C. Miao, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1290–1301, Jul.–Sep. 2022.

[25] Y. Gao, X. Fu, T. Ouyang, and Y. Wang, "EEG-GCN: Spatio-temporal and self-adaptive graph convolutional networks for single and multi-view eeg-based emotion recognition," *IEEE Signal Process. Lett.*, vol. 29, pp. 1574–1578, 2022.

[26] J. Li, S. Li, J. Pan, and F. Wang, "Cross-subject EEG emotion recognition with self-organized graph neural network," *Front. Neurosci.*, vol. 15, 2021, Art. no. 61165.

[27] J. Bi, F. Wang, X. Yan, J. Ping, and Y. Wen, "Multi-domain fusion deep graph convolution neural network for EEG emotion recognition," *Neural Comput. Appl.*, vol. 34, pp. 22241–22255, 2022.

[28] X. Lin, J. Chen, W. Ma, W. Tang, and Y. Wang, "EEG emotion recognition using improved graph neural network with channel selection," *Comput. Methods Prog. Biomed.*, vol. 231, 2023, Art. no. 107380.

[29] J. Li, W. Pan, H. Huang, J. Pan, and F. Wang, "STGATE: Spatial-temporal graph attention network with a transformer encoder for eeg-based emotion recognition," *Front. Hum. Neurosci.*, vol. 17, 2023, Art. no. 1169949.

[30] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Ment. Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.

[31] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *Proc. IEEE/EMBS 6th Int. Conf. Neural Eng.*, 2013, pp. 81–84.

[32] S. Koelstra et al., "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012.

[33] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 327–339, Jul./Sep. 2014.

[34] S. K. Hadjidimitriou and L. J. Hadjileontiadis, "Toward an EEG-based recognition of music liking using time-frequency analysis," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 12, pp. 3498–3510, Dec. 2012.

[35] S. Phadikar, N. Sinha, and R. Ghosh, “Automatic eyeblink artifact removal from EEG signal using wavelet transform with heuristically optimized threshold,” *IEEE J. Biomed. Health Inform.*, vol. 25, no. 2, pp. 475–484, Feb. 2021.
[36] W. Liu, W.-L. Zheng, and B.-L. Lu, “Emotion recognition using multimodal deep learning,” in *Proc. 23rd Int. Conf. Neural Inf. Process.*, Kyoto, Japan, 2016, pp. 521–529.
[37] X.-W. Wang, D. Nie, and B.-L. Lu, “Emotional state classification from EEG data using machine learning approach,” *Neurocomputing*, vol. 129, pp. 94–106, 2014.
[38] W.-L. Zheng, J.-Y. Zhu, Y. Peng, and B.-L. Lu, “EEG-based emotion classification using deep belief networks,” in *Proc. IEEE Int. Conf. Multimedia Expo*, 2014, pp. 1–6.
[39] C. A. Frantzidis, C. Bratsas, C. L. Papadelis, E. Konstantinidis, C. Pappas, and P. D. Bamidis, “Toward emotion aware computing: An integrated approach using multichannel neurophysiological recordings and affective visual stimuli,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 3, pp. 589–597, May 2010.
[40] L. I. Goldfischer, “Autocorrelation function and power spectral density of laser-produced speckle patterns,” *J. Opt. Soc. Amer.*, vol. 55, no. 3, Mar. 1965, Art. no. 247.
[41] Y. Liu and O. Sourina, “Real-time fractal-based valence level recognition from EEG,” in *Transactions on Computational Science XVIII*. Berlin, Germany: Springer, 2013, pp. 101–120.
[42] Y.-P. Lin et al., “EEG-based emotion recognition in music listening,” *IEEE Trans. Biomed. Eng.*, vol. 57, no. 7, pp. 1798–1806, Jul. 2010.
[43] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, “Spectral networks and locally connected networks on graphs,” 2013, *arXiv:1312.6203*.
[44] M. Defferrard, X. Bresson, and P. Vandergheynst, “Convolutional neural networks on graphs with fast localized spectral filtering,” in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 3844–3852.
[45] S. Gannouni, A. Aledaily, K. Belwafi, and H. Aboalsamh, “Emotion detection using electroencephalography signals and a zero-time windowing-based epoch estimation and relevant electrode identification,” *Sci. Rep.*, vol. 11, no. 1, 2021, Art. no. 7071.
[46] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, “Recurrent neural networks for multivariate time series with missing values,” *Sci. Rep.*, vol. 8, no. 1, 2018, Art. no. 6085.
[47] M. Lechner and R. Hasani, “Learning long-term dependencies in irregularly-sampled time series,” 2020, *arXiv:2006.04418*.
[48] M. Sundermeyer, R. Schlüter, and H. Ney, “LSTM neural networks for language modeling,” in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 45–46.
[49] S. Bai, J. Z. Kolter, and V. Koltun, “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,” 2018, *arXiv:1803.01271*.
[50] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” 2014, *arXiv:1412.3555*.
[51] I. O. Tolstikhin et al., “MLP-mixer: An all-MLP architecture for vision,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 24261–24272.
[52] S. J. Luck, *An Introduction to the Event-Related Potential Technique*. Cambridge, MA, USA: MIT Press, 2014.
[53] F. Qin, N. Gao, Y. Peng, Z. Wu, S. Shen, and A. Grudtsin, “Fine-grained leukocyte classification with deep residual learning for microscopic images,” *Comput. Methods Prog. Biomed.*, vol. 162, pp. 243–252, 2018.
[54] L. Yang and J. Liu, “Eeg-based emotion recognition using temporal convolutional network,” in *Proc. IEEE 8th Data Driven Control Learn. Syst. Conf.*, 2019, pp. 437–442.
[55] T. C. Technologies, “10–20 system positioning - manual,” Website, 1996. [Online]. Available: https://trans-cranial.com/docs/10_20_pos_man_v1_0_pdf.pdf
[56] J. Chen, X. Wang, C. Huang, X. Hu, X. Shen, and D. Zhang, “A large finer-grained affective computing EEG dataset,” *Sci. Data*, vol. 10, no. 1, 2023, Art. no. 740.
[57] G. E. Hinton, “Deep belief networks,” *Scholarpedia*, vol. 4, no. 5, 2009, Art. no. 5947.
[58] T. Song, S. Liu, W. Zheng, Y. Zong, and Z. Cui, “Instance-adaptive graph for EEG emotion recognition,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 2701–2708.
[59] W. Li, B. Hou, X. Li, Z. Qiu, B. Peng, and Y. Tian, “TMLP+SRDANN: A domain adaptation method for EEG-based emotion recognition,” *Measurement*, vol. 207, 2023, Art. no. 112379.
[60] G. Xu, W. Guo, and Y. Wang, “Subject-independent EEG emotion recognition with hybrid spatio-temporal GRU-conv architecture,” *Med. Biol. Eng. Comput.*, vol. 61, no. 1, pp. 61–73, 2023.
[61] D. Germano et al., “Unsupervised detection of covariate shift due to changes in EEG headset position: Towards an effective out-of-lab use of passive brain–computer interface,” *Appl. Sci.*, vol. 13, no. 23, 2023, Art. no. 12800.
[62] Y. Wang and J. Mo, “Emotion feature selection from physiological signals using tabu search,” in *Proc. 25th Chin. Control Decis. Conf.*, 2013, pp. 3148–3150.
[63] R. Ouerchefani, N. Ouerchefani, M. R. Ben Rejeb, and D. Le Gall, “Role of the prefrontal cortex and executive functions in basic emotions recognition: Evidence from patients with focal damage to the prefrontal cortex,” *Cogn. Neurosci.*, vol. 14, no. 3, pp. 75–95, 2023.
[64] S. Borgomaneri et al., “Increasing associative plasticity in temporo-occipital back-projections improves visual perception of emotions,” *Nature Commun.*, vol. 14, no. 1, 2023, Art. no. 5720.
[65] A. Dehghani, H. Soltanian-Zadeh, and G.-A. Hossein-Zadeh, “Neural modulation enhancement using connectivity-based EEG neurofeedback with simultaneous FMRI for emotion regulation,” *Neuroimage*, vol. 279, 2023, Art. no. 120320.
[66] Z. Yin, M. Zhao, Y. Wang, J. Yang, and J. Zhang, “Recognition of emotions using multimodal physiological signals and an ensemble deep learning model,” *Comput. Methods Prog. Biomed.*, vol. 140, pp. 93–110, 2017.
[67] X. Wu, W.-L. Zheng, and B.-L. Lu, “Identifying functional brain connectivity patterns for EEG-based emotion recognition,” in *Proc. IEEE/EMBS 9th Int. Conf. Neural Eng.*, 2019, pp. 235–238.
[68] X. Wu, W.-L. Zheng, Z. Li, and B.-L. Lu, “Investigating EEG-based functional connectivity patterns for multimodal emotion recognition,” *J. Neural Eng.*, vol. 19, no. 1, 2022, Art. no. 016012.