Multiscale Temporal Self-Attention and Dynamical Graph Convolution Hybrid Network for EEG-Based Stereogram Recognition

Lili Shen, Mingyang Sun, Qunxia Li, Beichen Li, Zhaoqing Pan, and Jianjun Lei

***Abstract*— Stereopsis is the ability of human beings to get the 3D perception on real scenarios. The conventional stereopsis measurement is based on subjective judgment for stereograms, leading to be easily affected by personal consciousness. To alleviate the issue, in this paper, the EEG signals evoked by dynamic random dot stereograms (DRDS) are collected for stereogram recognition, which can help the ophthalmologists diagnose strabismus patients even without real-time communication. To classify the collected Electroencephalography (EEG) signals, a novel multi-scale temporal self-attention and dynamical graph convolution hybrid network (MTS-DGCHN) is proposed, including multi-scale temporal self-attention module, dynamical graph convolution module and classification module. Firstly, the multi-scale temporal self-attention module is employed to learn time continuity information, where the temporal self-attention block is designed to highlight the global importance of each time segments in one EEG trial, and the multi-scale convolution block is developed to further extract advanced temporal features in multiple receptive fields. Meanwhile, the dynamical graph convolution module is utilized to capture spatial functional relationships between different EEG electrodes, in which the adjacency matrix of each GCN layer is adaptively tuned to explore the optimal intrinsic relationship. Finally, the temporal and spatial features are fed into the classification module to obtain prediction results. Extensive experiments are conducted on collected datasets i.e., SRDA and SRDB, and the results demonstrate the proposed MTS-DGCHN achieves outstanding classification performance compared with the other methods. The datasets are available at https://github.com/YANGeeg/TJU-SRD-datasets and the code is at https://github.com/YANGeeg/MTS-DGCHN.**

***Index Terms*— EEG, DRDS, self-attention, multi-scale convolution, dynamical graph convolution.**

Manuscript received January 5, 2022; revised April 5, 2022; accepted May 3, 2022. Date of publication May 9, 2022; date of current version May 16, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62171319, in part by the Science and Technology Found of Tianjin Health Commission under Grant ZC20187, and in part by the Science and Technology Found of Tianjin Eye Hospital under Grant YKZD2001. *(Lili Shen and Mingyang Sun contributed equally to this work.) (Corresponding author: Mingyang Sun.)*

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

Lili Shen, Mingyang Sun, Beichen Li, Zhaoqing Pan, and Jianjun Lei are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: sll@tju.edu.cn; sunmy@tju.edu.cn; bcli@tju.edu.cn; zqpan3-c@my.cityu.edu.hk; jjlei@tju.edu.cn).

Qunxia Li is with the School of Economics and Management, University of Science and Technology Beijing, Beijing 100098, China (e-mail: liqx@ustb.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2022.3173724

I. Introduction

STEREOPSIS, as the basis of depth perception, is the most advanced binocular vision function, which enables us to distinguish the distance through two-eye coordination [1]. When watching objects, the horizontal separation of two eyes results in a disparity in the retina and stereopsis is formed subsequently. As an important physiological indicator, stereopsis is necessary for carrying out better motor control and more accurate cognition [1]. In the clinical practice, stereopsis also refers to stereoacuity, which is the minimum parallax that can touch off depth perception [2]. The early methods for measuring stereoacuity include Howard-dolman test and Frisby–Davis Distance Test [3]. However, these methods exist some inherent limitations due to the effects of monocular cues and low measurement sensitivity. To address these issues, many excellent methods based on static random dot stereogram (RDS) [4], [5] are applied in stereoacuity test. Recently, research has found that human beings are more sensitive to dynamic random dot stereogram (DRDS), and adopting DRDS in stereoacuity test can achieve higher sensitivity [6]. For recognition tasks, one of the most popular work is developed using physiological signals as in [7]–[9]. Among them, EEG is widely applied in visual recognition field by offering high temporal resolution in noninvasive and cost-effective acquisition manner.

Many traditional machine learning algorithms for EEG classification [10], [11] are generally comprised of two primary stages, hand-crafted features extraction and the classification. For feature extraction, Liu *et al.* [12] adopted a short-time Fourier transform (STFT) with nonoverlapping Hanning window to extract time-frequency features. Bose *et al.* [13] utilized the Welch's method to capture power spectral density (PSD) as frequency features. Ang *et al.* [14] used filter bank CSP (FBSCP) to capture the optimal spatial features through a group of bandpass filters. Zeng [15] and Kakkos [16] utilized functional connectivity to construct brain feature space for EEG analysis. Subsequently, the extracted EEG features were sent to the classifiers like linear discriminant analysis (LDA) [17], random forest (RF) [18] or support vector machine (SVM) [19] for classification. However, these above methods rely on designer's prior knowledge in the specific domain, which might ignore some underlying information and

fail to capture the representative information from raw EEG signals.

Recently, deep learning has shown great potential in many fields, such as natural language processing [20], computer vision [21]–[23], speech recognition [24], on the basis of the ability to obtain deeper intrinsic representation from pristine data automatically. Besides, deep learning has also been applied to EEG signals classification tasks, and achieved great performance improvement [25], [26]. Some studies demonstrated that the spatial-temporal joint feature extraction strategy can better represent EEG signal characteristics [27], [28]. Therefore, EEG classification based on the spatial-temporal network model has become the mainstream research direction. For instance, Zhang *et al.* [28] introduced a spatial–temporal recurrent neural network (STRNN) to integrate both temporal and spatial EEG information for emotion recognition. Jeong *et al.* [29] designed a bidirectional long short-term memory (LSTM) network followed by multi-directional convolutional neural network (CNN) to capture the spatial-temporal EEG features for motor imagery (MI) classification. Zhang *et al.* [30] developed a parallel convolutional recurrent neural network for MI classification, in which the CNN was exploited to learn local spatial information, and the LSTM network was adopted to extract temporal information. In the above-mentioned models, the CNNs were utilized to explore spatial features, while the RNNs were applied to capture temporal dependencies in time-series EEG data. However, the RNN should traverse all units sequentially to acquire Long-term information before entering the current unit, which may lead to the gradient disappearance problem and thus it makes RNNs difficult to be trained.

To solve this issue, some scholars employed CNNs to obtain temporal information instead of RNNs, which achieved satisfying results. For example, Li *et al.* [31] proposed a channel-projection mixed-scale CNN to extract the mixed-scale spatial-temporal features for MI EEG classification. Li *et al.* [32] further proposed an attention-based multi-scale fusion CNN for MI EEG signals decoding, where used 1-D spatial and temporal CNNs to learn spatial-temporal information, and introduced the attention mechanism to make the features more distinguishable. Some methods [33], [34] transformed raw EEG sequences into mesh-like representations according to the spatial location of different electrodes, then captured the spatial and temporal characteristic with 2-D CNNs. These CNNs illustrate promising performance on capturing temporal contextual information and local discriminative features in spatial domain with the localized convolutional kernels. Despite of the advantage on local feature extraction, CNNs experience difficulty to learn global spatial functional relationships among electrodes. An intuitive solution is enlarging the receptive field, which however might damage functional dependencies among different EEG regions. Neuroimaging studies [35], [36] have found that visual pathways across in several brain regions are responsible for visual information processing. Besides, paper [37] has indicated that the complicated relationships among different EEG electrodes are significant for recognition tasks. Therefore, the characteristics of spatial connections in the brain need to be explored when designing networks for EEG classification.

With the development of graph theory, graph convolutional networks (GCN) [38] have been employed to learn the potential spatial connections between different nodes, which provides an effective way to acquire EEG spatial characteristics. Each EEG electrode can be regarded as a node of the graph, and the connection between the electrodes correspond to the edge of the graph. The weights (representing functional relationships among electrodes) of all edges constitute adjacent matrix of the graph. In this way, the spatial functional relationships between different EEG electrodes can be learned by GCN. A series of studies about GCN have been carried out. Song *et al.* [39] proposed a dynamic graph convolution neural network (DGCNN) to optimize a weighted graph to characterize the strength of functional relationships between different EEG electrodes for emotion recognition. Furthermore, Song *et al.* [40] introduced a variational instance-adaptive graph method with GCN to estimate the underlying uncertain information and learn the individual dependencies among different EEG electrodes, simultaneously. Zhang *et al.* [41] developed a sparse DGCNN model, which adds sparse constraints on the graph to make the weights localized and sparse to improve the performance. These GCN models firstly extracted hand-crafted EEG features such as power spectral density (PSD) and differential entropy (DE), and then fed them into GCN as nodes of graph to get the intrinsic relationships between EEG electrodes. In this kind of structure, the input features are designed in advance without optimization by network training and may ignore the heterogeneity among subjects, leading to suboptimal results. To address the issue, some end-to-end GCN models are proposed to automatically learn optimal spatial-temporal features of EEG signals. For instance, Wang *et al.* [42] designed an attention-based multi-scale convolutional neural network-dynamic graph convolutional network (AMCNN-DGCN) model to detect driving fatigue. Li *et al.* [43] adopted a spatial-temporal-spectral hierarchical graph convolutional network (STS-HGCN) to obtain the spatial and temporal features for seizure prediction. However, there are still some limitations in these EEG classification methods. First, existing end-to-end GCN methods usually extract temporal features from raw signals, and then apply GCN on these extracted features to capture the spatial connections between different brain regions, resulting in the original spatial connections be affected. Second, although some attention-based CNNs methods consider the local temporal importance in single EEG segment, they ignore the global importance dependences from the other EEG segments.

To tackle the aforementioned issues, we propose a dual network framework, termed multi-scale temporal self-attention and dynamical graph convolution hybrid network (MTS-DGCHN), including multi-scale temporal self-attention module (MTSM), dynamic graph convolution module (DGCM) and classification module. The multi-scale temporal self-attention module is composed of a temporal self-attention block and a multiscale convolution block. The temporal self-attention block emphasizes more discrimina-

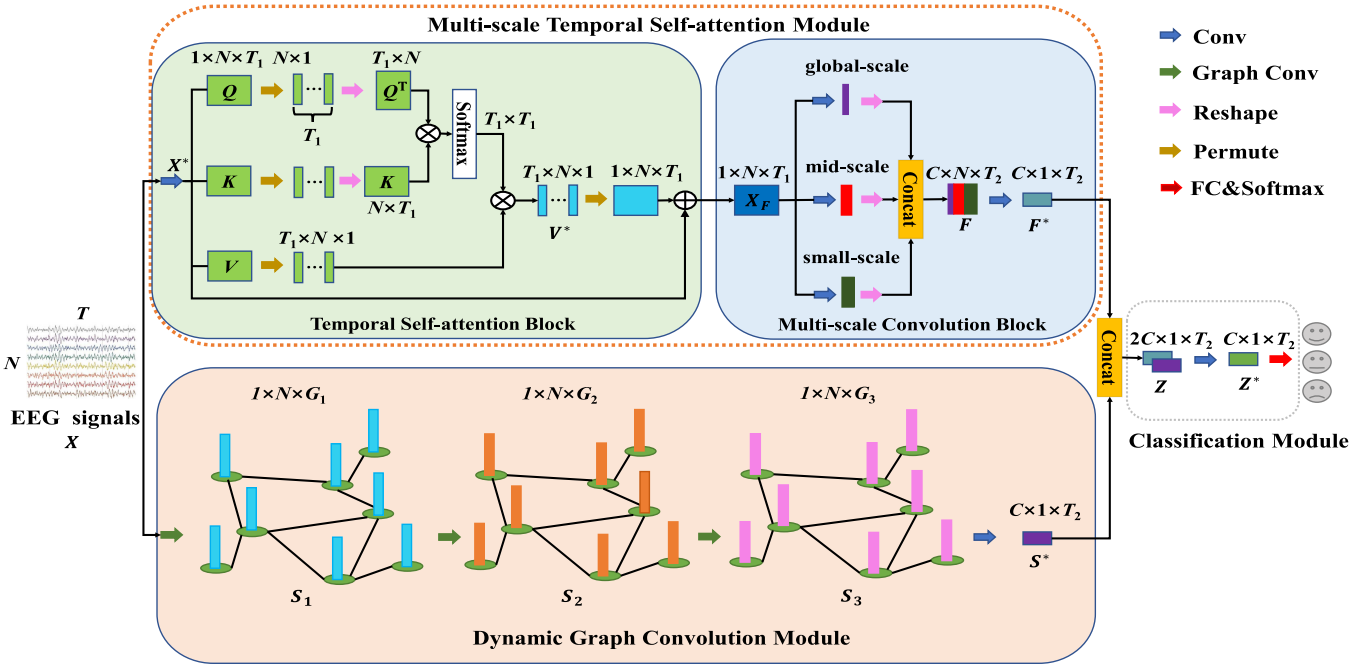

Fig. 1. An illustration of the proposed MTS-DGCHN framework, where 3-D feature maps are flatted.

tive segmented EEG information on the intact trail, and the multi-scale convolution block further extracts the global and local temporal context information. Simultaneously, the dynamic graph convolution module aims to acquire the potential spatial functional relationships between different EEG electrodes by dynamic GCN layers, in which a data-driven way is used to adjust the adjacency matrix adaptively, so as to make full use of the intrinsic relationships between electrodes. Finally, the extracted temporal and spatial features are sent to the classification module for predicting results.

The main contributions of this paper are summarized as follows:

1) In order to mine valuable temporal information, the temporal self-attention block and the multi-scale convolution block are proposed to highlight important time segments in an EEG trial and learn global and local information in multiple receptive fields. The multi-scale temporal self-attention module can fully capture significant time continuity representation information.
2) Aiming to obtain the spatial functional relationships between different EEG electrodes, the dynamic graph convolution module with several GCN layers is exploited, in which the adjacency matrix can update dynamically through the loss function in the process of network training.
3) The concurrent MTS-DGCHN can learn multi-scale time continuity and spatial functional dependences jointly. We evaluate the classification performance of the proposed MTS-DGCHN on the stereogram recognition EEG dataset A and B, and the experimental results demonstrate that the MTS-DGCHN is superior to the state-of-art methods.

The remaining of this paper is organized as follows. In Section II, we describe the proposed method in detail. Section III conducts extensive experiments on the dataset of stereogram recognition. Finally, a discussion is given in Section IV, and the conclusion is drawn in Section V.

II. Method

In this section, we first describe the notations and definitions used in this paper. Then, as shown in Fig. 1, overall framework of the proposed MTS-DGCHN is introduced. Finally, the implementation details of each module in MTS-DGCHN are described.

A. Notations and Definitions

Assume that the EEG signals of each subject is defined as $S = \{(X_i, y_i), i = 1, 2, \ldots, \mathrm{M}\}$, where $X_i \in R^{\mathrm{N}\times\mathrm{T}}$ is a two-dimension matrix representing the i-th trial of the EEG sample with N electrodes and T discretized time sampling points. M is the total number of EEG trials. y_i as the label of X_i takes values from the set Y, which contains H classes in the stereogram recognition task.For instance, the three-type stereogram recognition dataset contains corresponding label set $Y = \{y_1 ='' unclearly'', y_2 ='' slightly'', y_3 ='' clearly''\}$. In the network, the size of the feature map is detonated as $c \times h \times w$, which describes the number, height and width of the feature map respectively. For each EEG sample, the shape of X_i is regarded as $N \times 1 \times T$ in the model. We need to reshape it to $1 \times N \times T$ as input. The size of each convolutional kernel can be defined as $h_0 \times w_0$ where h_0 is the electrode (channel) dimension and w_0 is the time dimension.

B. Multiscale Temporal Self-Attention and Dynamical Graph Convolution Hybrid Network

To take advantage of temporal and spatial features of EEG signals, we propose a concurrent network, termed MTS-DGCHN, as shown in Fig. 1. Specifically, we design a multi-scale temporal self-attention module, in which a temporal self-attention block is utilized to emphasize more valuable time segments, and a multi-scale convolution block is adopted to learn temporal context information in global and local receptive fields. Meanwhile, the dynamic graph convolution module is developed to characterize the spatial functional dependencies among different EEG electrodes via GCNs. Finally, the discriminative temporal and spatial features are fed into the classification module for stereogram recognition. In the following, we will detailly exhibit each module of the proposed MTS-DGCHN.

1) Multiscale Temporal Self-Attention Module: EEG signals have a high time resolution, and each trail can be divided into a series of time segments. Different time segments are correlated with each other and play diverse roles in describing EEG time characteristics. In order to focus on valuable temporal information, the more useful time segments are assigned higher importance scores by the temporal self-attention block. In particular, the raw input EEG signal is $X = [x_1, \ldots, x_N] \in R^{N\times T}, x_N \in R^T$ where N is the number of electrodes, and T is the length of time sequence in each electrode, where $N = 30$ and $T = 256$. Firstly, raw signal X implements a convolution operation over time with the kernel size of 1×7 and the stride of 1×4 to get the shallow temporal features $X^* \in R^{1\times N\times T_1}$. Additionally, in order to calculate the importance scores in temporal self-attention block, the extracted temporal feature X^* is taken as query (Q), key (K) and value (V), so $Q = K = V = X^*$. The Q, K and V are further permuted into T_1 one-dimensional column vectors with the size of $N \times 1$ as queries, keys and values. Then, each query is conducted dot products with T_1 keys, and sent to a Softmax layer to obtain the importance scores of each value. In practice, the attention function on a set of queries is combined into a matrix Q^T, and the keys are reshaped into matrix K. The global importance score matrix φ can be calculated as follows.

$$\varphi(Q, K) = soft\max\left(Q^T K\right), \quad \varphi \in R^{T_1\times T_1} \tag{1}$$

Guided by this global importance score matrix φ, each one-dimensional column vector in V can aggregate valuable information to update itself, which is essentially a weighted sum updating process in the unit of each vector. The updated result is denoted as.

$$V^* = \varphi(Q, K)\, V \in R^{T_1\times N\times 1} \tag{2}$$

The skip connection is utilized to prevent gradient vanishing problem. The output feature X_F of temporal self-attention block is obtained.

$$X_F = X^* + permute\left(V^*\right) \in R^{1\times N\times T_1} \tag{3}$$

To further capture advanced temporal context features, three kinds of convolution kernels are applied in multi-scale convolution block, including global-scale, mid-scale and small-scale. To be more specific, global-scale convolution kernels are used to capture global features F_1 while mid-scale and small-scale convolution kernels are employed for local features F_2 and F_3. The three transformations $\mathcal{F}_1 : X_F \to F_1 \in R^{C_1\times N\times t_1}$, $\mathcal{F}_2 : X_F \to F_2 \in R^{C_2\times N\times t_2}$, $\mathcal{F}_3 : X_F \to F_3 \in R^{C_3\times N\times t_3}$ are performed via a set of 1-D convolution kernels with three kinds of kernel sizes $k, k/2r, k/4r$, where $k = T_1 = 43$, reduction ratio r is set to 4 in this paper, $C_1 = 348, C_2 = 12, C_3 = 12$ are the number of convolution kernels. Then, three kinds of features are reshaped to $F_1 \in R^{C\times N\times(C_1\times t_1/C)}$, $F_2 \in R^{C\times N\times(C_2\times t_2/C)}$, $F_3 \in R^{C\times N\times(C_3\times t_3/C)}$ and concatenated into the feature F along the channel dimension. F is defined as follows.

$$F = concat\,(F_1, F_2, F_3) \in R^{C\times N\times T_2} \tag{4}$$

where $C = 18, T_2 = (C_1t_1 + C_2t_2 + C_3t_3)/C = 43$. Finally, F is fed into a one-dimensional convolutional layer of size $N \times 1$ to average the influence of different electrodes. Therefore, discriminative temporal feature $F^* \in R^{C\times 1\times T_2}$ is obtained after the multiscale temporal self-attention module, which contains the significant temporal context information in local and global receptive fields.

2) Dynamic Graph Convolution Module: The functional relationships between EEG electrodes play a crucial role in EEG classification. Previous studies [39], [42] have indicated that GCNs can be employed to describe the dependencies among nodes. In order to stimulate the functional relationships, the dynamic graph convolution module is designed by adopting GCN layers. Based on GCN, where each EEG electrode is referred to one node of the graph whereas the connection between different EEG electrodes is corresponded to the edge of the graph.

Concretely, an undirected and weighted graph is represented as $G = (V, A)$. $V = \{v_1, v_2, \ldots, v_i, \ldots, v_N\}$ is the node set, in which v_i represents an electrode. The $N \times N$ matrix A is the adjacency matrix of G, describing the edge weight between nodes in V. Each a_{ij} in A denotes the connection importance from node i to node j. The key to build a better graph structure is how to determine the appropriate adjacency matrix. Phase locking value (PLV) measures phase synchronous change of two signals over a period of time, and contains interactive information to some extent between signals [44]. Therefore, PLV is applied to depict the functional connectivity of the EEG electrodes in the dynamic graph convolution module, where each a_{ij} in adjacency matrix A is represented by PLV. The PLV values of signal p and signal q are calculated by the following formula.

$$PLV_{pq} = \frac{1}{T}\left|\sum\nolimits_t^T e^{(\varphi_p(t)-\varphi_q(t))}\right| \tag{5}$$

where $\varphi(t)$ is the signal phase at time t, and T is the length of the signal, and PLV value range [0,1]. Based on PLV, the element a_{ij} in adjacency matrix A is determined by.

$$a_{ij} = \begin{cases} PLV_{ij} & PLV_{ij} \geq \tau \\ 0 & otherwise \end{cases} \tag{6}$$

where τ is an artificial threshold to make the adjacency matrix sparse. In this paper, τ is set to 0.5. About 70 percentage on

average of PLV values were above threshold. We use A to further compute the Laplace matrix L as follows.

$$L = \hat{D}^{-1/2}\,\hat{A}\hat{D}^{-1/2} \tag{7}$$

where $\hat{A} = A + I_N$, $\hat{A}$ is identity matrix, $\hat{D} = diag([d_1, d_2, \ldots, d_N])$ is the degree matrix of A, and $d_i = \sum_j a_{ij}$. The GCN update formula for each layer can be defined as.

$$S = \sigma\,(LX\theta) \in R^{N\times G} \tag{8}$$

where the input is $X = [x_1, \ldots, x_{\mathrm{N}}] \in R^{\mathrm{N}\times\mathrm{T}}$, $\theta \in R^{T\times G}$ is the filter's parameter matrix of GCN, σ is the RELU activation function, and G is the number of nodes after the GCN layer. Therefore, the spatial feature S_3 can be obtained through three GCN layers for original EEG signal X. The transformation is $\mathcal{G}_1 : X \to S_1 \in R^{1\times N\times G_1}$, $\mathcal{G}_2 : S_1 \to S_2 \in R^{1\times N\times G_2}$, $\mathcal{G}_1 : S_2 \to S_3 \in R^{1\times N\times G_3}$, where $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$ are the operation of GCN, and $G_1 = 192, G_2 = 128, G_3 = 43$ are the feature numbers obtained after each GCN layer.

In order to fuse the spatial and temporal feature map in the classification module, the number of C convolution kernels with the size of $N \times 1$ is employed to change the dimension of the feature map S_3. Therefore, we get the final feature map $S^* \in R^{C\times 1\times T_2}$ after dynamic graph convolution module, which is corresponding to the dimension of output F^* from the multi-scale temporal self-attention module. Besides, the adjacency matrix A of the graph convolution module is constantly updated during the training of this network. The process of dynamic updating A will be introduced in Algorithm 1.

3) Classification Module: In this module, the captured temporal and spatial features is firstly fused after multiscale temporal self-attention module and dynamic graph convolution module. Specifically, the temporal feature F^* and the spatial feature S^* are connected among the channel dimension to obtain the discriminative feature Z by the following formula.

$$Z = concat\left(S^*, F^*\right) \in R^{2C\times 1\times T_2} \tag{9}$$

Then, Z is further refined among channel dimension by 1×1 convolution to obtain the final spatial-temporal feature Z^*.

$$Z^* = Conv\,(Z) \in R^{C\times 1\times T_2} \tag{10}$$

where $Conv\,(\cdot)$ is the convolution operation. Finally, the feature Z^* is sent to two fully connected layers and one Softmax layer to get the predicted label $\hat{y}$. The calculation formula is as follows.

$$\hat{y} = Soft\max\left(W_2\left(\mathrm{Re}\,lu\left(W_1 Z^* + b_1\right)\right) + b_2\right) \tag{11}$$

where W_i, b_i are the weight matrix and bias matrix of the fully connected layers, $\hat{y} \in R^H$ is the predicted label with H classes.

In summary, the detailed steps of the optimizing procedure of the model are shown in Algorithm 1.

Specifically, supposed that we are given M labeled training data set $\{S, Y\} = \{X_i, y_i\}_{i=1}^M$, where $S \in R^{M\times N\times T}$, $Y \in R^{M\times H}$. In the MTS-DGCHN model, X_i is fed into the multi-scale temporal self-attention module and dynamic graph convolution module simultaneously to capture the temporal feature F^* and spatial feature S^*. Then, the temporal and spatial features are refined to get the spatial-temporal feature Z^*. Finally, the prediction label $\hat{y}$ is obtained by the classification module based on Eq. (11).

Algorithm 1 The Optimizing Procedure of the MTS-DGCHN

Input: A labeled EEG dataset $\{S, Y\} = \{X_i, y_i\}_{i=1}^M$ The number of epoch E, the batch size of each epoch B and other hyperparameters.

Output: The optimal set of the model parameter Θ and the learned adjacency matrix A.

Initialize the parameters in the proposed MTS-DGCHN, including Θ and other hyperparameters, and Initialize adjacency matrix A according to Eq.(5)(6).

for epoch $= 1 : E$ **do**
  while this epoch is not complete **do**
    Sample one batch size of samples X_B and y_B from S and Y, respectively.
    Calculate the temporal feature map $F^*{}_B$ by passing X_B into the MTSM based on Eq. (1)(2)(3)(4).
    Calculate the Laplacian matrix L based on Eq. (7).
    Calculate the spatial feature map $S^*{}_B$ by passing X_B into the DGCM based on Eq. (8).
    Calculate the fused spatial-temporal feature map $Z^*{}_B$ by passing $F^*{}_B$ and $S^*{}_B$ into the classification module based on Eq. (9)(10).
    Calculate the prediction label $\hat{y}_B$ by passing $Z^*{}_B$ into Eq. (11).
    Utilize y_B and $\hat{y}_B$ to calculate the loss function based on Eq. (12).
    Update the model parameter set Θ and the adjacency matrix A via SGD optimizer according to the loss function.
  end while
end for

In the output of MTS-DGCHN, the cross-entropy loss L is utilized to evaluate the inconsistencies between real label y_i and predicted label $\hat{y}$.

$$L = -\frac{1}{M}\sum_{k=1}^{K}\sum_{i=1}^{M} y_i{}^k \log\left(\hat{y}_i^k\right) + \lambda\|\Theta\|_1 + \mu\|A\|_1 \tag{12}$$

where Θ is all model parameters during the training process, $\|\cdot\|_1$ represents the $l_1 - norm$, λ and μ are constants. The regularization term $\lambda\|\Theta\|_1 + \mu\|A\|_1$ is used to prevent overfitting and dynamically update adjacency matrix A. In the process of implementation, the batch size is 32, the learning rate is set to 0.01 and the SGD optimizer is adopted. The λ and μ in Eq. (5) are 0.00001 and 0.2, respectively. The peoposed MTS-DGCHN is trained on the NVIDIA GPU (RTX 3090) with Python 3.6 and Pytorch 1.9. After 100 epochs, it takes 14.5 mins to get the optimal model. The five-fold cross validation is used to evaluate the MTS-DGCHN comprehensively.

TABLE I
THE DATA FORMAT FOR EACH SUBJECT

Array name	Array shape	Array contents
Data	432×30×1024	trial×channel×data
Label	432×3	trial×label(unclearly, slightly, clearly)

III. EXPERIMENTAL RESULTS

A. EEG Datasets

In this study, two stereogram recognition EEG datasets are built to evaluate the performance of the proposed MTS-DGCHN. The details of the EEG datasets are described as follows.

1) Stereogram Recognition EEG Dataset a: (SRDA) contains EEG data of 5 subjects (2 males and 3 females), which are recorded by 32 EEG electrodes at 1000 Hz when the subjects are watching 24 high-contrast DRDS vides with three parallax patterns, i.e., stereoscopic graphics in DRDS can be recognized clearly, slightly and unclearly. The EEG acquisition consists of 6 sessions, and each session contains 72 trials, so that each subject owns a total of 432 EEG trials. More details of the SRDA can be found in [45].

2) Stereogram Recognition EEG Dataset B: (SRDB) contains EEG data of 8 subjects (5 males and 3 females), which are collected when they are watching 24 low-contrast DRDS vides with three parallax patterns. The SRDB is established following the details of SRDA, and the subjects in the two databases do not overlap.

In this paper, the sampling frequency in SRDA and SRDB are down-sampled from 1000 Hz to 256 Hz. The invalid information and power line interference are filtered out through a 1-40hz band-pass filter. Then, the baseline removal is performed through the EEGLAB toolbox in MATLAB, the artifacts are removed by independent component analysis (ICA). Only 30 electrodes, except for the reference A1, A2, are adopted. The data format is illustrated in Table I. For data augmentation, we take overlapping slices of the signals (80% overlap), and obtain a series of 1s samples. Finally, the trials will be expanded from 432 to 6,912 for each subject.

B. Overall Performance

The proposed MTS-DGCHN is a competitive model, which can effectively capture the temporal context information and the spatial intrinsic relationships between different electrodes. To verify the performance advantage of MTS-DGCHN, eight state-of-the-art models were chosen for comparison on the SRDA and SRDB datasets. These models are briefly described in the following.

1) MCS-STWCSG [45]: A traditional machine learning model we proposed previously, which is based on multi-channel selection and CSP for the same work.
2) EEGNet [46]: A compact convolutional neural network model for EEG classification tasks, which combined deep convolution and separable convolution closely.
3) RCNN [47]: A CNN combined with RNN network structure for EEG classification.
4) DGCNN [39]: The handcrafted features, such as DE and PSD, were fed into dynamic graph convolutional network for emotion recognition.
5) AMCNN-DGCN [42]: A serial framework combining an attention-based multiscale CNN with a dynamical GCN for detecting driving fatigue from EEG signals.
6) TSception [48]: A spatial-temporal multi-scale CNN framework for emotion recognition.
7) AttnSleep [49]: An attention-based deep learning model named AttnSleep for EEG classification of sleep stages.
8) TS-SEFFNet [50]: An EEG decoding framework, which applies squeeze and-excitation feature fusion network to capture temporal-spectral features for motor imagery task.

TABLE II
THE OVERALL COMPARISON RESULTS OF AVERAGE CLASSIFICATION PERFORMANCE

Datasets	Methods	ACC (%)	PRE (%)	REC (%)	F1-score	Kappa
SRDA	MCS-STWCSG	87.50	86.13	85.85	0.860	0.813
	RCNN	86.80	84.64	83.53	0.841	0.802
	EEGNet	81.56	80.35	83.06	0.817	0.723
	TSception	93.66	93.78	93.05	0.934	0.905
	DGCNN	90.82	90.84	90.26	0.905	0.862
	AMCNN-DGCN	92.97	92.8	91.65	0.922	0.895
	AttnSleep	93.59	91.91	92.36	0.921	0.904
	TS-SEFFNet	94.65	94.10	93.89	0.940	0.920
	MTS-DGCHN	**95.47**	**94.91**	**94.06**	**0.945**	**0.932**
SRDB	MCS-STWCSG	86.94	86.06	84.95	0.855	0.804
	RCNN	84.88	84.12	83.71	0.839	0.773
	EEGNet	79.40	79.55	81.15	0.803	0.691
	TSception	92.94	92.69	92.53	0.926	0.894
	DGCNN	89.36	89.27	88.91	0.891	0.840
	AMCNN-DGCN	90.49	91.16	89.17	0.902	0.857
	AttnSleep	92.85	92.04	91.62	0.918	0.893
	TS-SEFFNet	94.03	93.47	92.19	0.928	0.910
	MTS-DGCHN	**95.19**	**94.65**	**95.17**	**0.949**	**0.928**

where the bold values indicate the best results.

Table II summarizes the overall classification results of different methods on both datasets. We selcet the accuracy, precision, recall, F1-score and Kappa values as evaluation metrics to comprehensively evaluate the classification performance.

From Table II, it can be seen that, on the SRDA dataset, the MTS-DGCHN achieves the highest average accuracy of 95.47%, which is 7.97%, 8.67%, 13.91%, 1.81%, 4.65%, 2.50%, 1.88% and 0.82% higher than MCS-STWCSG, RCNN, EEGNet, TSception, DGCNN, AMCNN-DGCN, AttnSleep and TS-SEFFNet, indicating the ability of MTS-DGCHN to learn multi-scale temporal representations and spatial functional dependences. Meanwhile, for the other evaluation metrics such as the precision, recall, F1-score and Kappa values, the proposed MTS-DGCHN yields 94.91%, 94.06%, 0.945 and 0.928, respectively, which are the highest among all the methods. For the attention-based models, the MTS-DGCHN outperforms the AMCNN-DGCN, AttnSleep and TS-SEFFNet on all

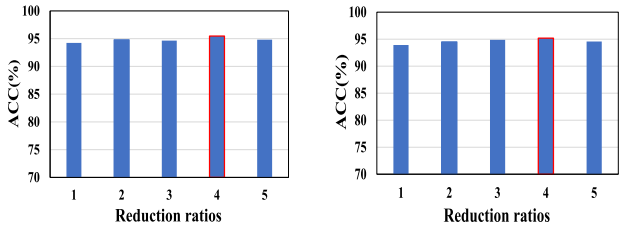

Fig. 2. average accuracy comparison between different reduction ratios in MCB on (a) SRDA, (b) SRDB.

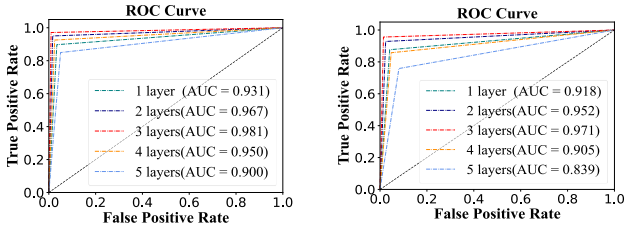

Fig. 3. ROC curves comparison between different numbers of GCN layers on (a) SRDA, (b) SRDB.

evaluation metrics. Therefore, the model achieves best results and proves excellent performance for EEG classification.

To demonstrate the adaptability of the proposed MTS-DGCHN, the performance of the MTS-DGCHN and other methods is also evaluated on SRDB dataset. Table II clearly shows that the average classification accuracy of MTS-DGCHN is encouraging as it gains the highest accuracy of 95.19%. Moreover, the proposed model yields average precision of 94.65%, average recall of 95.17% and average F1-score of 0.949, which outperform all the other methods. Besides, the proposed MTS-DGCHN can achieve promising results of 0.928 in Kappa value, which is at least 0.018 higher than the compared methods. In general, the above experimental results shows the outstanding performance of the MTS-DGCHN for stereogram recognition on SRDA and SRDB datasets.

C. The Impact of Model Parameters

In this section, we will conduct experiments on the SRDA and SRDB to explore the impact of different parameters on the overall performance of the model. The main parameters include the reduction ratios of multi-scale convolution block (MCB) in MTSM, the number of GCN layers in DGCM and the parameter μ in the loss function in Eq. (12). When changing one of the above parameters, the other parameters are fixed to ensure the experimental effectiveness.

With high time resolution, EEG contains rich time context information. Generally, the reduction ratios corresponds to the receptive fields of EEG information. In the MCB of the MTSM, the convolution kernels of size k are used to capture global information, while the convolution kernels of size $k/2r, k/4r$ are used to capture local information, and r is the reduction ratio. The impact of the reduction ratio r in MCB are discussed particularly in Fig 2. The results show that the convolution kernels with $r = 4$ has achieved the best classification accuracy.

To verify the impact of the number of GCN layers in DGCM, we set the number ranging from 1 to 5. Table III depicts the average results of all subjects with different number of GCN layers, and it can be observed that the MTS-DGCHN reaches the optimal performance with three GCN layers. Besides, the ROC curves are further draw in Fig. 3. We can see that the MTS-DGCHN with three GCN layers obtains the highest AUC of 0.981 and 0.971 on SRDA and SRDB, showing strong robustness. Therefore, a conclusion can be drawn that the three GCN layers is the best for the MTS-DGCHN.

TABLE III
AVERAGE ACCURACY COMPARISON BETWEEN DIFFERENT NUMBERS OF GCN LAYERS

layers	SRDA		SRDB	
	ACC(%)	Kappa	ACC(%)	Kappa
1 GCN layer	89.56	0.84	89.43	0.84
2 GCN layers	93.17	0.90	92.50	0.89
3 GCN layers	**95.47**	**0.93**	**95.19**	**0.93**
4 GCN layers	89.92	0.85	88.29	0.82
5 GCN layers	87.34	0.81	86.75	0.80

where the bold values indicate the best results.

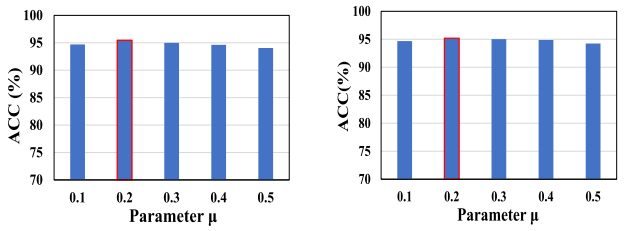

Fig. 4. average accuracy comparison between different μ in the loss function on (a) SRDA, (b) SRDB.

Apart from the parameters of the network structure, the parameters μ in Eq. (12) will also affect classification performance. The parameter μ in the loss function controls the adaptive adjustment of the adjacency matrix A in GCN layers, and determines the trends of network optimization. We conduct experiments on value $\mu \in (0.1, \ldots, 0.5)$ with step 0.1 to search for the best setting. Fig. 4 shows results of the proposed MTS-DGCHN with parameter μ. Experimental results indicates that the MTS-DGCHN can reach the highest classification accuracy when μ is set to 0.2.

D. Ablation Study

In this section, we conducted ablation experiments to verify the impact of each part in the MTS-DGCHN. Firstly, we explore the importance of MTSM and DGCM in the proposed model. Table IV shows the result of each subject on SRDA and SRDB. It is observed that the average accuracy and the Kappa value of MTS-DGCHN are both improved compared with the MTS-DGCHN without DGCM or MTSM. The results obviously proved that MTSM and DGCM are both significant in the MTS-DGCHN model. The reason is that

TABLE IV
THE COMPARISON RESULTS OF MTS-DGCHN W/O DGCM OR MTSM

Datasets	Subject	MTS-DGCHN w/o DGCM		MTS-DGCHN w/o MTSM		MTS-DGCHN	
		Acc	kappa	Acc	kappa	Acc	kappa
SRDA	S1	88.15	0.822	85.34	0.780	**89.72**	**0.846**
	S2	92.98	0.895	90.57	0.859	**95.25**	**0.929**
	S3	94.69	0.920	90.83	0.862	**96.11**	**0.942**
	S4	94.26	0.914	92.12	0.882	**97.56**	**0.963**
	S5	95.30	0.930	93.09	0.896	**98.70**	**0.981**
	AVE	93.08	0.896	90.39	0.856	**95.47**	**0.932**
SRDB	S1	93.74	0.906	88.24	0.824	**97.69**	**0.965**
	S2	90.29	0.854	86.64	0.800	**94.70**	**0.921**
	S3	93.64	0.905	90.33	0.855	**96.17**	**0.943**
	S4	82.96	0.744	85.14	0.777	**89.16**	**0.837**
	S5	95.85	0.938	92.43	0.886	**98.38**	**0.976**
	S6	91.90	0.879	89.35	0.840	**95.94**	**0.939**
	S7	88.58	0.829	90.95	0.864	**92.33**	**0.885**
	S8	94.63	0.919	89.79	0.847	**96.90**	**0.954**
	AVE	91.45	0.872	89.11	0.837	**95.19**	**0.928**

where the bold values indicate the best results.

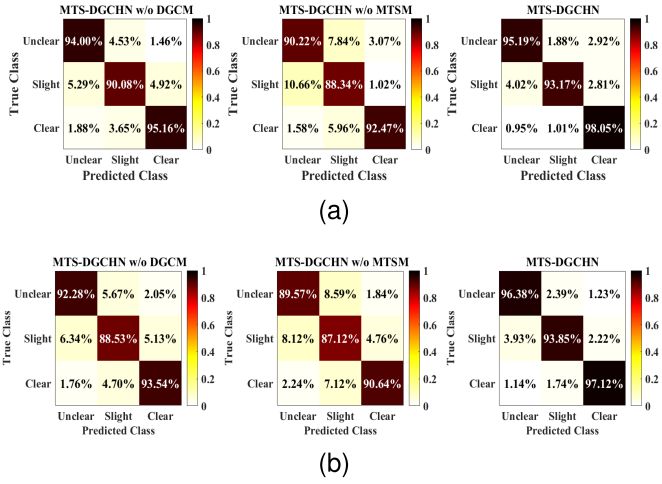

Fig. 5. The confusion matrixes on both (a) SRDA and (b) SRDB.

the MTSM mainly captures global and local time continuity information, while the DGCM learns the spatial relationships between different EEG electrodes. Therefore, missing any one of them will make the classification performance decline inevitably.

In addition, we construct the confusion matrix with the data of all subjects in Fig. 5, where the values on the diagonal represent the correct recognition, and the others are wrong. As we can see, the classification accuracy of the MTS-DGCHN in each category is better than that of models without MTSM or DGCM. Therefore, the MTSM and DGCM in the MTS-DGCHN are both critical to the overall success of the proposed model. Furthermore, the temporal self-attention block (TSB) is performed ablation validation to illustrate the function in the MTSM. According to Fig. 6, the classification performance of each subject is improved due to the existence of the TSB, which demonstrate the effectiveness of the TSB.

In order to further demonstrate the benefits of the temporal self-attention block, we compared our self-attention block with other two attention blocks. Specifically, we apply two existing attention blocks in temporal domain, which are combined with our multi-scale convolution block (MCB) for classification

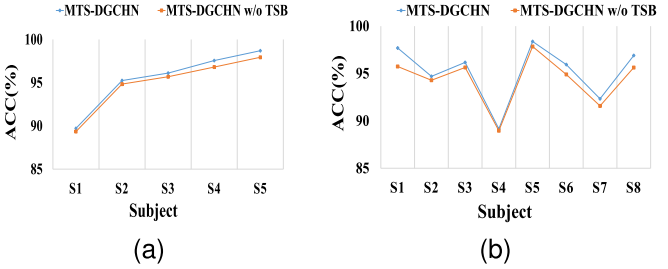

Fig. 6. Performance comparison of MTS-DGCHN w/o TSB on (a) SRDA, (b) SRDB.

TABLE V
PERFORMANCE COMPARISON OF DIFFERENT MODULE COMPONENTS

Module combination	SRDA		SRDB	
	Acc	kappa	Acc	kappa
Attention block in [42] + MCB	92.51	0.888	90.89	0.863
SE block [51] + MCB	92.89	0.893	91.04	0.866
our self-attention block + MCB	**93.08**	**0.896**	**91.45**	**0.872**

where the bold values indicate the best results.

TABLE VI
AVERAGE ACCURACY(%) COMPARISON BETWEEN DIFFERENT ADJACENCY MATRICES

Datasets	Random		PLV	
	fixed	dynamic	fixed	dynamic
SRDA	89.36	95.00	93.26	**95.47**
SRDB	87.14	94.92	92.63	**95.19**

where the bold values indicate the best results.

on SRDA and SRDB. As shown in Table V, the proposed self-attention block can achieve the highest average accuracy and Kappa value on both two datasets, which illustrates the advantages of the multi-scale temporal self-attention module.

Finally, we explore the impact of different types of adjacency matrix of GCN layers in Table VI, in which the fixed means that adjacency matrix A will not be optimized by loss function in Algorithm 1, while the dynamic means that adjacency matrix A will be updated as learnable network parameter. For fixed adjacency matrix, the classification accuracy of PLV matrix is obviously higher than random matrix. The reason is that the random matrix will ignore the correlation among EEG electrodes. On the contrary, the PLV matrix quantifies the phase synchronization between signals, which can provide some useful information about the relationships between different EEG electrodes. For dynamically updated adjacency matrix, the classification accuracy of PLV matrix is also higher than random matrix, which indicates the PLV matrix enables the DGCM take advantage of prior knowledge to optimize the network towards optimal results. Besides, we can find that the dynamically updated adjacency matrix can achieve better performance than the fixed adjacency matrix. It demonstrates that the dynamically updated adjacency matrix can learn the potential spatial connections of EEG electrodes, and help to improve the model performance of the proposed MTS-DGCHN.

E. Performance of Subject-Independent Experiments

To investigate the model generalizability across subjects for EEG-based stereogram recognition, we conduct subject-independent experiments on datasets SRDA and

TABLE VII
ACCURACIES (%) COMPARISON OF SUBJECT-INDEPENDENT EXPERIMENTS

Methods	Year	SRDA		SRDB	
		Acc(%)	F1	Acc(%)	F1
RCNN	2015	40.25	0.396	42.91	0.417
EEGNet	2018	44.76	0.437	45.10	0.444
TSception	2020	56.84	0.559	58.25	0.578
DGCNN	2020	61.31	0.611	63.47	0.631
AMCNN-DGCN	2021	53.78	0.518	56.65	0.564
AttnSleep	2021	55.69	0.547	57.92	0.571
TS-SEFFNet	2021	61.58	0.608	63.07	0.629
MTS-DGCHN	2022	62.16	0.619	63.63	0.642

TABLE VIII
SIGNIFICANCE TEST FOR THE COMPARISON METHODS ON SRDB

Comparison Methods	MTS-DGCHN
MCS-STWCSG	0.0152
RCNN	0.0043
EEGNet	0.0023
TSception	0.0137
DGCNN	0.0026
AMCNN-DGCN	0.0309
AttnSleep	0.0098
TS-SEFFNet	0.0002

SRDB, and the results are listed in Table VII. Specifically, the leave-one-subject-out cross-validation (LOSO-CV) is adopted in subject-independent experiments [52]. For example, assuming that S1 is the test subject. Except for S1, all the other subjects were used to train the proposed MTS-DGCHN, and the data from S1 was utilized for the performance evaluation.

Table VII shows that the MTS-DGCHN achieves the highest accuracy and F1-score across subjects on both two datasets, which indicates the proposed model outperforms the other methods. Compared with subject-dependent experiments, the subject-independent experiments will lead to significant performance degradation, which is caused by the individual differences among different subjects. However, the proposed model can still achieve promising accuracies and F1-scores (much better than the random results in three-class classification case), which illustrates the MTS-DGCHN owns the ability to learn individual differences to some extent in EEG-Based stereogram recognition tasks.

F. Statistical Analysis

Additionally, the Wilcoxon signed-rank test [53] is performed between the proposed MTS-DGCHN and other methods on SRDB to explore the statistical significance of the comparison results. The significance tests results are shown in Table VIII, it can be observed that the p-values are all smaller than 0.05, which indicates the differences of average classification performance between the MTS-DGCHN and every comparison model are statistically significant. Therefore, the proposed MTS-DGCHN significantly improve the performance for EEG stereogram recognition.

IV. DISCUSSION

A. Study on Computational Complexity

To demonstrate the benefits of the proposed MTS-DGCHN, the computational complexity and average classification performance of the MTS-DGCHN is compared with other seven methods, and the results are exhibited in Table IX. The MTS-DGCHN involves approximately 4.2×10^5 parameters, and is more than the EEGNet with only 3×10^4 parameters, observably less than RCNN and AMCNN-DGCN. Moreover, the model inference time, indicating the time a deep learning model required to give a recognition for each EEG trial, is evaluated and the results are shown in Table IX. We can see that the inference time of the MTS-DGCHN is 1.97 ms, which is slower than EEGNet, but significantly faster than RCNN, AMCNN-DGCN and TS-SEFFNet. Besides, the model decoding time defined as the duration from raw EEG to decoding results is reported. The decoding time of the proposed method on one EEG trial is 13.10 ms. which is also slower than EEGNet. We can see that EEGNet achieves the lowest computational complexity, but the poorest classification performance, about 15% lower than the proposed MTS-DGCHN on both two datasets. Therefore, the MTS-DGCHN has achieved the state-of-the-art performance with acceptable complexity, which demonstrates the advantages of the proposed MTS-DGCHN.

TABLE IX
COMPARISON OF COMPUTATIONAL COMPLEXITY

Methods	Parameter (million)	Inference time(ms)	Decoding time(ms)	ACC (%)	
				SRDA	SRDB
RCNN	2.26	30.59	150.76	86.80	84.88
EEGNet	0.03	0.32	4.28	81.56	79.40
TSception	0.33	0.85	8.44	93.66	92.94
DGCNN	0.06	1.02	11.17	90.82	89.36
AMCNN-DGCN	1.68	3.73	19.72	92.97	90.49
AttnSleep	0.93	2.56	15.07	93.59	92.85
TS-SEFFNet	1.34	3.29	18.48	94.65	94.03
MTS-DGCHN	0.42	1.97	13.10	95.47	95.19

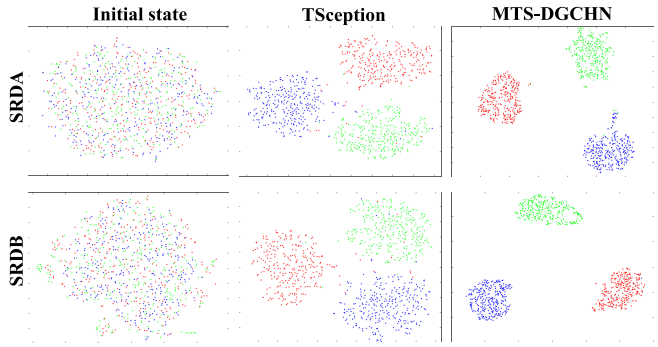

Fig. 7. The t-SNE visualization in 2D embedding space of features learned from the subject S1 on SRDA and SRDB. Red points denote recognized unclearly, green points denote recognized slightly and blue points denote recognized clearly.

B. T-SNE Visualization of MTS-DGCHN

In order to study the distribution of features captured by the proposed MTS-DGCHN, the extracted EEG features are transformed into a 2-D embedding dimension with t-SNE visualization technology. As shown in Fig. 7, three different types of EEG signals are mixed when they are in initial state.

After training, the EEG signals will be recognized efficiently. There are still a few samples cannot be distinguished,

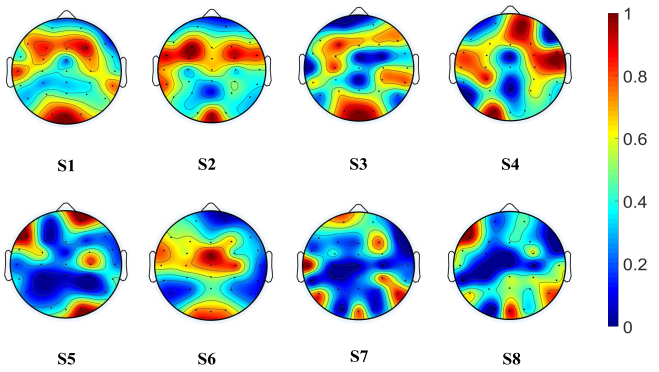

Fig. 8. Topographic maps learned from the MTS-DGCHN model on SRDB dataset for different subjects based on "clearly" recognition state.

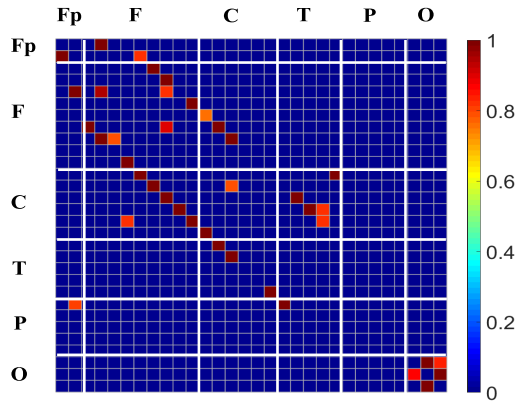

Fig. 9. Top-45 connections (about 5% of the total connections) between channels learned on SRDB with MTS-DGCHN model, which is the average of all subjects. Fp: frontal pole region; F: frontal region; C: central region; T: temporal region; P: parietal region and O: occipital region.

but most samples are predicted correctly. Compared with the TSception, the MTS-DGCHN generates more separable features, which can easily recognize the classes of EEG signals. Therefore, the t-SNE visualization results demonstrate that the MTS-DGCHN extracts the discriminative EEG features, indicating the proposed model is effective.

C. The Visualization of Critical Connections in the Graph

To further investigate obtained spatial relationship among EEG electrodes, we try to employ Degree Centrality to visualize the graph connections. The degree centrality has been commonly used to evaluate the connectivity of the nodes, which measures the importance of a node with the other nodes. In the proposed MTS-DGCHN, the learned adjacency matrix A characterizes the connections between EEG electrodes. The values in the i-th row and the j-th column of A represent the weights associated with the i-th node. Thus, the degree centrality C_i of the i-th EEG electrode can be calculated by.

$$C_i = \sum_{m=1}^{30} A_{m,i} + \sum_{n=1}^{30} A_{i,n} - 2A_{i,i} \quad (i = 1, \ldots, 30) \quad (13)$$

As shown in Fig. 8, the degree centrality C of eight subjects on SRDB is visualized based on clear recognition state on SRDB dataset. For better visualization, all values are normalized to [0,1]. We can observe that the learned spatial connections are different among subjects, which illustrates the proposed MTS-DGCHN can adaptively capture the spatial relationship of EEG electrodes for each subject. Consistently, the degree centrality is high in the occipital lobe since the stereogram recognition has high relations with the occipital lobe.

We also try to visualize the critical connections in graph based on the average of all subjects for DRDS recognition, as shown in Fig. 9. In adjacency matrix A, only the 45 highest edge weights, termed the top-45 connections (about 5% of the total connections) are preserved for better illustration. From Fig. 9, we can see that the frontal-frontal, frontal-central, Temporal-central and occipital-occipital are closely related. This phenomenon can also be observed in Fig. 8 for most subjects. Previous studies have shown that the occipital region is related to visual tasks [54], and the frontal region is often associated with attention-related activity [57]. Besides, neurophysiological researches [55], [56] indicate that object shape perception is processed by the ventral visual pathway across the occipital region to the temporal region. In the stereogram recognition task, subjects are required to judge shapes of the DRDS, which exactly refer to attention-related shape perception task with binocular vision function. The visualization results demonstrate the consistency of active brain regions between the critical connection in the graph learned by MTS-DGCHN and neurophysiological evidence.

D. The Visualization of Weights on EEG Time Segments

To demonstrate the effectiveness of temporal self-attention block, we further visualize the attention weights added to the EEG time segments by the self-attention block. In our stereogram recognition task, there are three kinds of recognition states, clearly, slightly and unclearly. We visualized the weight sequence of two different trials for S1 and S2 in each state. As shown in Fig. 10, each row represents one trial, and each column corresponds to a weight added to one time segment. From Fig. 10, we can see that some time segments will be highlighted by assigning higher attention weight. Besides, the weights for each subject are different between trials and between states.Moreover, the discrepancy in the same recognition state also exists between subjects. The visualization has illustrated that the proposed model can highlight important time segments within and between subjects, thus achieving better EEG stereogram recognition performance.

E. Limitations and Future Directions

Although the proposed MTS-DGCHN has achieved outstanding EEG classification performance, there are still two main limitations in our present work. First, Phase Locking Value is applied to depict the functional connectivity for constructing the adjacency matrix in the dynamic graph convolution module. However, studies [58], [59] have shown that the Phase Locking Value exists some problems, such as active reference electrodes and volume conduction when measuring

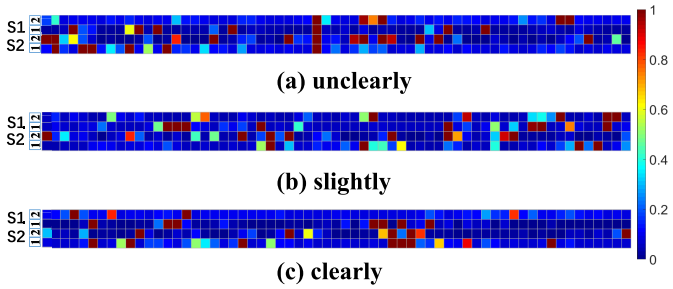

Fig. 10. The visualization of weight sequences learned by the self-attention block. In each state, two different trials (1, 2) of two subjects (S1, S2) are visualized. Each row represents one trial, and each column corresponds to a weight added to one time segment.

functional connectivity. Therefore, in our future work, we will research for the optimal metrics to assess functional connectivity for the stereogram recognition task. Second, the proposed model shows the effectiveness in subject-independent EEG classification. However, the performance is reduced by about 30% compared with subject-dependent experiments. The reasons may be complicated and diverse. On one hand, our EEG datasets do not have enough subjects for an deep learning model to carry out subject-independent experiments. On the other hand, there are significant individual differences in the SRDA and SRDB, which causes drifting issue in data distributions among different subjects, so that the proposed method cannot commendably address it. Therefore, some solutions like the transfer learning will be considered to improve the MTS-DGCHN model in the feature.

V. Conclusion

In this paper, we propose a dual network termed MTS-DGCHN for EEG-Based stereogram recognition. The proposed model consists of the multi-scale temporal self-attention module and the dynamic graph convolution module. On one hand, the multi-scale temporal self-attention module learns global importance of different time segments in an EEG trial with self-attention block, and the local and global temporal continuity features of EEG signals with multi-scale convolution. On the other hand, the dynamic graph convolution module obtains the potential spatial functional relationships of different electrodes. The results indicate that the proposed model achieves average accuracy of 95.47% and 95.19% on SRDA and SRDB, which are superior to other six state-of-the-art methods.In conclusion, the MTS-DGCHN is an effective model for stereogram recognition. The subjects we collected are all healthy groups with visual function. In the future, some subjects suffer from stereopsis disorders like strabismus will be added. We hope this study will make a substantial contribution to find patients in a timely manner.

References

[1] H. Wu *et al.*, "Evaluating stereoacuity with 3D shutter glasses technology," *BMC Ophthalmol.*, vol. 16, no. 1, pp. 1–8, Apr. 2016.
[2] H. R. Kim, D. E. Angelaki, and G. C. DeAngelis, "The neural basis of depth perception from motion parallax," *Phil. Trans. Roy. Soc. B, Biol. Sci.*, vol. 371, no. 1697, Jun. 2016, Art. no. 20150256.
[3] G. Westheimer, "Clinical evaluation of stereopsis," *Vis. Res.*, vol. 90, pp. 38–42, Sep. 2013.
[4] D. Kim, S. Choi, and K. Sohn, "Effect of vergence–accommodation conflict and parallax difference on binocular fusion for random dot stereogram," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 5, pp. 811–816, May 2012.
[5] K.-H. Lee and P.-L. Chiu, "Sharing visual secrets in single image random dot stereograms," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4336–4347, Oct. 2014.
[6] A. Budai *et al.*, "Validation of dynamic random dot stereotests in pediatric vision screening," *Graefe's Arch. Clin. Experim. Ophthalmol.*, vol. 257, no. 2, pp. 413–423, Feb. 2019.
[7] J. Han, X. Ji, X. Hu, L. Guo, and T. Liu, "Arousal recognition using audio-visual features and FMRI-based brain response," *IEEE Trans. Affect. Comput.*, vol. 6, no. 4, pp. 337–347, Oct. 2015.
[8] U. Cote-Allard *et al.*, "A transferable adaptive domain adversarial neural network for virtual reality augmented EMG-based gesture recognition," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 546–555, 2021.
[9] J. R. Paulo, G. Pires, and U. J. Nunes, "Cross-subject zero calibration Driver's drowsiness detection: Exploring spatiotemporal image encoding of EEG signals for convolutional neural network classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 905–915, 2021.
[10] A. Jiang, J. Shang, X. Liu, Y. Tang, H. K. Kwan, and Y. Zhu, "Efficient CSP algorithm with spatio-temporal filtering for motor imagery classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 4, pp. 1006–1016, Apr. 2020.
[11] C. Li *et al.*, "Seizure onset detection using empirical mode decomposition and common spatial pattern," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 458–467, 2021.
[12] Y.-J. Liu, M. Yu, G. Zhao, J. Song, Y. Ge, and Y. Shi, "Real-time movie-induced discrete emotion recognition from EEG signals," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 550–562, Jan. 2017.
[13] R. Bose, H. Wang, A. Dragomir, N. V. Thakor, A. Bezerianos, and J. Li, "Regression-based continuous driving fatigue estimation: Toward practical implementation," *IEEE Trans. Cognit. Develop. Syst.*, vol. 12, no. 2, pp. 323–331, Jul. 2019.
[14] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on BCI competition IV datasets 2A and 2B," *Frontiers Neurosci.*, vol. 6, no. 1, p. 39, 2012.
[15] L.-L. Zeng *et al.*, "Identifying major depression using whole-brain functional connectivity: A multivariate pattern analysis," *Brain A, J. Neurol.*, vol. 135, no. 5, pp. 1498–1507, May 2012.
[16] I. Kakkos *et al.*, "EEG fingerprints of task-independent mental workload discrimination," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 10, pp. 3824–3833, Oct. 2021.
[17] R. Fu, Y. Tian, T. Bao, Z. Meng, and P. Shi, "Improvement motor imagery EEG classification based on regularized linear discriminant analysis," *J. Med. Syst.*, vol. 43, no. 6, pp. 1–13, Jun. 2019.
[18] J. Luo, Z. Feng, J. Zhang, and N. Lu, "Dynamic frequency feature selection based approach for classification of motor imageries," *Comput. Biol. Med.*, vol. 75, pp. 45–53, Aug. 2016.
[19] E. Dong, C. Li, L. Li, S. Du, A. N. Belkacem, and C. Chen, "Classification of multi-class motor imagery with a novel hierarchical SVM algorithm for brain–computer interfaces," *Med. Biol. Eng. Comput.*, vol. 55, no. 10, pp. 1809–1818, Oct. 2017.
[20] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[21] Z. Pan, W. Yu, J. Lei, N. Ling, and S. Kwong, "TSAN: Synthesized view quality enhancement via two-stream attention network for 3D-HEVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 345–358, Jan. 2021.
[22] B. Peng, J. Lei, H. Fu, Y. Jia, Z. Zhang, and Y. Li, "Deep video action clustering via spatio-temporal feature learning," *Neurocomputing*, vol. 456, pp. 519–527, Oct. 2021.
[23] J. Lei, X. Li, B. Peng, L. Fang, N. Ling, and Q. Huang, "Deep spatial-spectral subspace clustering for hyperspectral image," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2686–2697, Jul. 2021.
[24] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 12, pp. 2263–2276, Dec. 2016.
[25] 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.*, early access, Jun. 14, 2019, doi: 10.1109/TAFFC.2019.2922912.

[26] X. Du *et al.*, "An efficient LSTM network for emotion recognition from multichannel EEG signals," *IEEE Trans. Affect. Comput.*, early access, Aug. 3, 2020, doi: 10.1109/TAFFC.2020.3013711.

[27] W. Tao *et al.*, "EEG-based emotion recognition via channel-wise attention and self attention," *IEEE Trans. Affect. Comput.*, early access, Sep. 22, 2020, doi: 10.1109/TAFFC.2020.3025777.

[28] 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.

[29] J.-H. Jeong, K.-H. Shim, D.-J. Kim, and S.-W. Lee, "Brain-controlled robotic arm system based on multi-directional CNN-BiLSTM network using EEG signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 5, pp. 1226–1238, May 2020.

[30] D. Zhang, L. Yao, K. Chen, S. Wang, X. Chang, and Y. Liu, "Making sense of spatio-temporal preserving representations for EEG-based human intention recognition," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3033–3044, Jul. 2019.

[31] Y. Li, X.-R. Zhang, B. Zhang, M.-Y. Lei, W.-G. Cui, and Y.-Z. Guo, "A channel-projection mixed-scale convolutional neural network for motor imagery EEG decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 6, pp. 1170–1180, Jun. 2019.

[32] D. Li, J. Xu, J. Wang, X. Fang, and Y. Ji, "A multi-scale fusion convolutional neural network based on attention mechanism for the visualization analysis of EEG signals decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2615–2626, Dec. 2020.

[33] J. Chen, D. Jiang, Y. Zhang, and P. Zhang, "Emotion recognition from spatiotemporal EEG representations with hybrid convolutional recurrent neural networks via wearable multi-channel headset," *Comput. Commun.*, vol. 154, pp. 58–65, Mar. 2020.

[34] D. Li, B. Chai, Z. Wang, H. Yang, and W. Du, "EEG emotion recognition based on 3-D feature representation and dilated fully convolutional networks," *IEEE Trans. Cognit. Develop. Syst.*, vol. 13, no. 4, pp. 885–897, Dec. 2021.

[35] A. Khasnobish, A. Konar, D. N. Tibarewala, and A. K. Nagar, "Bypassing the natural visual-motor pathway to execute complex movement related tasks using interval type-2 fuzzy sets," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 1, pp. 91–105, Jan. 2017.

[36] T. Kim, S. Lee, and A. C. Bovik, "Transfer function model of physiological mechanisms underlying temporal visual discomfort experienced when viewing stereoscopic 3D images," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4335–4347, Nov. 2015.

[37] W. Hu *et al.*, "Interpretable multimodal fusion networks reveal mechanisms of brain cognition," *IEEE Trans. Med. Imag.*, vol. 40, no. 5, pp. 1474–1483, May 2021.

[38] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[39] 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. 2020.

[40] T. Song *et al.*, "Variational instance-adaptive graph for EEG emotion recognition," *IEEE Trans. Affect. Comput.*, early access, Mar. 9, 2021, doi: 10.1109/TAFFC.2021.3064940.

[41] G. Zhang, M. Yu, Y.-J. Liu, G. Zhao, D. Zhang, and W. Zheng, "SparseDGCNN: Recognizing emotion from multichannel EEG signals," *IEEE Trans. Affect. Comput.*, early access, Jan. 13, 2021, doi: 10.1109/TAFFC.2021.3051332.

[42] H. Wang, L. Xu, A. Bezerianos, C. Chen, and Z. Zhang, "Linking attention-based multiscale CNN with dynamical GCN for driving fatigue detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[43] Y. Li, Y. Liu, Y.-Z. Guo, X.-F. Liao, B. Hu, and T. Yu, "Spatio-temporal-spectral hierarchical graph convolutional network with semisupervised active learning for patient-specific seizure prediction," *IEEE Trans. Cybern.*, early access, May 25, 2021, doi: 10.1109/TCYB.2021.3071860.

[44] Z.-M. Wang, R. Zhou, Y. He, and X.-M. Guo, "Functional integration and separation of brain network based on phase locking value during emotion processing," *IEEE Trans. Cognit. Develop. Syst.*, early access, Jun. 11, 2020, doi: 10.1109/TCDS.2020.3001642.

[45] L. Shen, X. Dong, and Y. Li, "Analysis and classification of hybrid EEG features based on the depth DRDS videos," *J. Neurosci. Methods*, vol. 338, May 2020, Art. no. 108690.

[46] V. Lawhern, A. Solon, N. Waytowich, S. M. Gordon, C. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, 2018, Art no. 56013.

[47] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," 2015, *arXiv:1511.06448*.

[48] Y. Ding *et al.*, "TSception: A deep learning framework for emotion detection using EEG," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–7.

[49] E. Eldele *et al.*, "An attention-based deep learning approach for sleep stage classification with single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 809–818, 2021.

[50] Y. Li, L. Guo, Y. Liu, J. Liu, and F. Meng, "A temporal-spectral-based squeeze-and-excitation feature fusion network for motor imagery EEG decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1534–1545, 2021.

[51] H. Jie, S. Li, and S. Gang, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[52] O. Y. Kwon, M. H. Lee, C. Guan, and S. W. Lee, "Subject-independent brain-computer interfaces based on deep convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 3839–3852, Oct. 2020.

[53] D. Zhang, L. Yao, K. Chen, S. Wang, P. D. Haghighi, and C. Sullivan, "A graph-based hierarchical attention model for movement intention detection from EEG signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 11, pp. 2247–2253, Nov. 2019.

[54] X. Fan, Q. Zhou, Z. Liu, and F. Xie, "Electroencephalogram assessment of mental fatigue in visual search," *Bio-Med. Mater. Eng.*, vol. 26, no. s1, pp. S1455–S1463, Aug. 2015.

[55] S. Kastner and L. G. Ungerleider, "Mechanisms of visual attention in the human cortex," *Annu. Rev. Neurosci.*, vol. 23, no. 1, pp. 315–341, 2000.

[56] V. Zachariou, R. Klatzky, and M. Behrmann, "Ventral and dorsal visual stream contributions to the perception of object shape and object location," *J. Cognit. Neurosci.*, vol. 26, no. 1, pp. 189–209, Jan. 2014.

[57] G. Doniger *et al.*, "Activation timecourse of ventral visual stream object-recognition areas: High density electrical mapping of perceptual closure processes," *J. Cognit. Neurosci.*, vol. 12, no. 4, pp. 615–621, Jul. 2000.

[58] C. J. Stam, G. Nolte, and A. Daffertshofer, "Phase lag index: Assessment of functional connectivity from multi channel EEG and MEG with diminished bias from common sources," *Hum. Brain Mapping.*, vol. 28, no. 11, pp. 1178–1193, Nov. 2007.

[59] K. Yoshinaga *et al.*, "Comparison of phase synchronization measures for identifying stimulus-induced functional connectivity in human magnetoencephalographic and simulated data," *Frontiers Neurosci.*, vol. 14, p. 648, Jun. 2020.