Article:

Zhang, J., Li, K., Yang, B. et al. (2025) Cross-dataset motor imagery decoding — A transfer learning assisted graph convolutional network approach. Biomedical Signal Processing and Control, 102. 107213. ISSN: 1746-8094

https://doi.org/10.1016/j.bspc.2024.107213

Cross-Dataset Motor Imagery Decoding - A Transfer Learning Assisted Graph Convolutional Network Approach

Jiayang Zhang, Kang Li, *Senior Member, IEEE*, Banghua Yang and Zhengrun Zhao

***Abstract*— The proliferation of portable electroencephalogram (EEG) recording devices has made it practically feasible to develop the motor imagery (MI) based brain-computer interfaces (BCIs). However, the low signal-to-noise ratio of EEG signals for abstract MI tasks, limited data, limited EEG channels, and strong inter- and intra-subject variability pose significant challenges for MI-task recognition. This paper proposes a transfer learning assisted graph convolutional network (GCN) modeling approach for cross-dataset MI decoding, one of the most challenging issues in this field. In the experiments, a multi-channel dataset with 62 electrodes and a few-channel dataset with 8 electrodes are utilized for cross-dataset modeling. To harness multi-channel information, we utilize the GCN module to aggregate topological features. The pre-trained model is guided with few-channel signals as inputs through a knowledge distillation framework. Subsequently, the pre-trained model is adapted to the few-channel dataset using a transfer learning strategy with minimal data training. Experiment results show that the proposed model achieves 3.92% and 3.83% more accuracy improvement compared with state-of-the-art models in the cross-validation and cross-session scenario respectively, demonstrating the effectiveness of the proposed approach in cross-dataset MI-EEG decoding, thus enabling more effective MI-BCI applications.**

***Index Terms*— Brain-computer interface, Cross-dataset, Graph convolution network, Transfer learning, Motor imagery.**

I. Introduction

THE brain-computer interface (BCI) is a neurotechnological system enabling direct communication between brains and external devices by recognizing patterns of brain activities [1]. Motor imagery (MI), as one of the important paradigms in the BCI field, allows people to self-generate massive electroencephalogram (EEG) signals by stimulating the related motor cortex without actual movements. Due to the emphasis on achieving functional independence and returning to a normal life without reliance on supportive devices, the MI-BCI system has consistently been prioritized in motor rehabilitation therapy [2].

Decoding algorithms play a vital role in developing an effective and resilient MI-BCI system. The commonly used traditional machine learning method namely Common Spatial Pattern (CSP) [3] extracted the spatial feature of EEG signals by identifying a set of filters that maximize or minimize the variance of signals. The core idea is also employed by subsequent variations of the CSP approach like Common Spatio-Spectral Pattern (CSSP) [4] and Filter Bank Common Spatial pattern (FBCSP) [5]. However, the process of manually selecting classifiers and feature extractors in machine learning constrains the model's classification accuracy and generalization capability.

Deep learning (DL), as an end-to-end data-driven model, has been widely employed in BCI-based applications. Classic model structures such as ConvNets [6] and EEGNet [7] have effectively leveraged the flexibility of Convolution Neural Networks (CNNs), capturing important temporal-spatial features in EEG signals. Dai et al [8] utilized multi-scale CNN filters to extract the temporal features across different windows. Mane et al [9], proposed the FBCNet model by utilizing multi-frequency band inputs and incorporating a variance layer for highly generalizable discriminative feature extraction. In addition to the extended research in the time and frequency domains, Ju et al [10] introduced the framework of geometric deep learning, namely Tensor-CSPNet. The model characterized spatial covariance matrices obtained from EEG signals on symmetric positive definite (SPD) manifolds and successfully captured the temporal-spatial-frequency patterns. Attention mechanism-based approaches such as transformer [11] have been widely applied to MI-EEG decoding recently. For instance, Yang et al [12] constructed interdependencies among the deep-temporal and multi-spectral domains through SE-Block to highlight channel-wise feature responses. Zhang et al [13] adopted the local and global schemes of the transformer to capture temporal features dynamically.Altaheri et al [14] employed multi-head self-attention to features and CNNs to capture high-level temporal features based sliding window. The adapted model with dynamic convolution allowed richer information to be learned at the kernel level without increasing parameters according to [15]. Although significant successes have been achieved in decoding EEG signals in the temporal domain, learning spatial features by CNNs is not effective [16], especially given the EEG signals are full of rich topological information [17].

Graph neural networks (GNNs), operating on graph domain

J. Zhang, K. Li and Z. Zhao are with the School of Electronic and Electrical Engineering, University of Leeds, Woodhouse Lane, Leeds, West Yorkshire, LS2 9JT, UK (e-mail: eljzh@leeds.ac.uk; k.li1@leeds.ac.uk; elzzhao@leeds.ac.uk).

B. Yang is with the School of Mechatronic Engineering and Automation, Shanghai University, Shangda Road 99, Shanghai, 200444, China (e-mail: yangbanghua@shu.edu.cn).

with convincing performance, have been applied across various areas such as social networks [18], traffic forecasting [19], and neuropsychological studies [20]. In the BCI field, Ding et al [16] proposed the local-global-graph network (LGGNet) to classify cognitive tasks by learning the different functional areas of the brain. Vivek et al [21] replaced depthwise CNNs with the Graph Convolutional Network (GCN) to harness the functional connectivity among channels. Sun et al [22] employed multiple GCN layers and an aggregation-selection method to choose an optimal set of channels. Ma et al [23] utilized a double-brand GCN to filter channels, weakening the spatial gap across different subjects. Therefore, GNNs perform well in aggregating and selecting EEG channels, as well as in acquiring spatial knowledge. Despite remarkable advancements in the classification accuracy of MI tasks using the DL-based methods, these models require substantial amounts of data for training and considerable time for calibration [24]. In the practical application of BCI systems, achieving good classification accuracy and robustness across different subjects with minimal or no retraining on new data poses a significant challenge.

Transfer learning (TL) methods such as domain adaptation (DA) and domain generalization (DG) are gradually beginning to be employed. DA-based approaches train adapted classifiers with limited data from the target domain by leveraging massive training data from source domains [25]. Hang et al [26] adopted maximum mean discrepancy (MMD) to align the source and target domain features. Chen et al [27] considered both the class-related and time-related labels and transferred a single subject to another one. DG-based methods are also an important research direction in BCI applications, such as driver status detection [28], mental workload classification [29] and MI [30]. The core idea of DG is to find the invariant features among the source domain without acquiring new data for retraining. Although the accuracy is not yet sufficient, a considerable amount of experimental and calibration time is saved. Fine-tuning is also an effective approach to transfer parameters by freezing parts of layers in the pre-training models based on the source domain [31]. However, these studies only used single datasets for validation, which are often collected using numerous wet electrodes, ensuring higher data quality. In contemporary EEG applications, there is an increasing prevalence of portable EEG acquisition devices [32]. To save experimental time, these devices typically have fewer channels and employ dry electrodes, resulting in lower data quality [33], bringing a significant challenge for decoding MI-EEG signals. Additionally, due to variations in the number of channels, high-quality datasets cannot be directly utilized for transfer learning without channel selection. Research on MI-based cross-dataset studies remains limited. Zaremba and Atyabi [34] used three different datasets and filtered data with the same 11 channels existing across these datasets. However, this method merely consolidated subjects from different datasets for the cross-subject training, following a leave-one-out experiment, and did not address the differences between various datasets. Xu [35] and Xie [36] only choose three channels (C3, CZ, C4) across different datasets. The former method employed Riemannian Procrustes Analysis (RPA) [37] to align the Riemannian center among subjects within different datasets and train the DL-based model. The latter approach adopted fine-tuning technology when applying models to new MI paradigms. However, these methods do not consider the channel differences caused by various datasets or devices, as well as the potential transfer from a multi-channel dataset to one with fewer channels.

To remedy these limitations, we propose a GCN network based on the Knowledge Distillation [38] and fine-tuning methods, extracting the temporal-spatial-spectral features in the multi-channel public dataset with high data quality, aggregating specific channels information and transferring to the dataset with much fewer channels for decoding MI tasks. In the experiment, the public dataset [39] collected by 62 wet electrodes is regarded as the source domain while the few-channel dataset collected by only 8 dry electrodes is regarded as the target domain. First, we train the proposed model as the teacher network with 62-channel inputs. To reduce the 62-channel data to the same specific 8 channels in the target domain, GCN layers are adopted to aggregate spatial information. Then the student network with these 8-channel inputs in the source domain is guided by the teacher network to learn the aggregated information and effectively harness all the data in the source domain. Finally, the pre-trained student model is validated on the target domain by fine-tuning technology with minimal data of re-training, aiming to transfer the parameters learned from the source domain model. From the experiment results, the proposed model is shown to achieve the highest accuracies among the compared benchmarks. Furthermore, ablation studies and visualization experiments are conducted to understand the effectiveness of GCN layers and transfer learning strategies. The public BCI-IV-2a dataset was also adapted to verify the superiority of the proposed model [40].

The remainder of the article is organized as follows. Section II gives the dataset descriptions and the detailed structure of the proposed model. Section III presents the results including the ablation studies and visualization experiments. Discussions are given in Section IV and we conclude the article in Section V.

II. Methods

In this section, datasets collected from three different EEG devices are introduced. Subsequently, we present the preprocessing steps, the specific details of the proposed model and the experimental procedure.

A. Data Description

1) Korean University dataset [39] (KU dataset): The EEG signals were collected using a device(BrainAmp; Munich, Germany) with 62 wet electrodes whose impedances were maintained below 10 $k\Omega$. The EEG channel configuration (Fig. 1(a)) conformed to the International 10-20 system. This dataset comprised 54 healthy individuals performing left and right-hand motor imagery tasks. Each subject participated in two experimental sessions, with each session consisting of 200 trials. Consequently, each individual contributed a total of 400

trial data. In this experiment, we downsampled the sampling rate from $1000Hz$ to $250Hz$.

2) Few-channel dataset (8-channel dataset): The EEG signals were collected using a device (BlueBCI; Beijing, China) with only 8 dry electrodes (Fig. 1(b)) whose impedances were about 300 $k\Omega$. This device is portable and features a plug-and-play functionality, eliminating the need for bridging the gap between electrode pin and scalp with conductive electrolyte gel and significantly reducing experimental preparation time. However, its impedance is much higher compared to wet electrodes, leading to a decrease in data quality. This dataset included 22 healthy subjects performing left and right-hand motor imagery tasks. Each subject had data from 5 blocks, with a total of 80 trials across these blocks, sampled at a rate of 250 Hz. In the cross-session experiment, the first three blocks were used as the training set, while the last two blocks served as the testing set.

3) BCI-IV-2a Dataset: The EEG signals were collected using a device with only 22 Ag/AgCl electrodes (Fig. 1(c)) which also had low impedances and good data quality, making it one of the most popular public MI datasets in previous studies. This dataset included EEG data from 9 subjects, with each one having data from 2 sessions. Each session consisted of 6 runs, with each run comprising 48 trials (12 trials for each of the four possible MI classes namely left hand, right hand, both feet and tongue), resulting in a total of 288 trials per session. In this experiment, we focused on a binary classification of left and right-hand movements, leading to 144 trials per session. In the cross-session experiment, the first session was regarded as the training set, while the second session was the testing set.

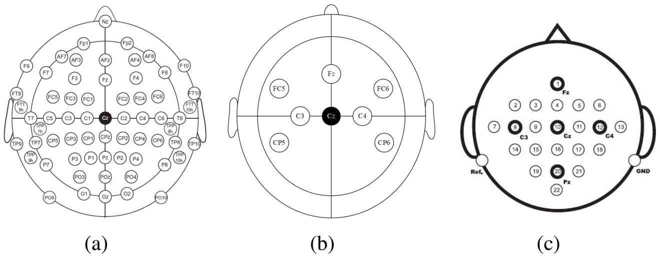


Fig. 1: The channel configuration of the International 10-20 system: (a) KU dataset with 62 channels. (b) Few-channel dataset with 8 channels. (c) BCI-IV-2a dataset with 22 channels.

B. Framework

The resistance of the portable device with 8 dry electrodes is much higher than the device with 62 wet electrodes used in the KU dataset. Moreover, the limited amount of data in the 8-channel dataset is insufficient for achieving robust classification results when utilized for within-subject modeling, particularly for deep learning models. In practical BCI applications, there is a greater need for models that can be used immediately or with minimal data calibration. Therefore, transferring model information from the KU dataset with high data quality and abundant data to the 8-channel dataset with fewer channels holds meaningful and valuable implications. However, the two datasets utilized devices with a different number of channels. While it is feasible to only use the 8 channels common to both devices like previous studies [34] [35] [36], this approach does not fully exploit the additional channel information present in the KU dataset. To address this problem, we divide the whole experiment framework into two parts. The first part employed data distillation, enabling the model to learn a compact representation that captures the task-specific feature representation when using all 62 channels, even though the model was trained with only 8 channels as input. The second step involves fine-tuning of the pre-trained model using parts of data in the target domain dataset for training, followed by validation of the remaining target domain data. BCI-IV-2a Dataset was also served as a target domain dataset for evaluating the performance of cross-dataset models.

C. Model Structure

The proposed model primarily consists of four components: Temporal Block, Dense Block, Graph Block and Feature Extraction (shown in Fig. 2).

1) Temporal Block: The EEG signals are denoted as $E = (X_i, Y_i)|i = 1, 2, ..., N$, where $X_i \in R^{C\times T}$ represents $i - th$ EEG trial with C channels and T samples. N is the total number of EEG signal trials. First, the EEG signals are sent into three parallel CNN layers with multi-scale temporal kernels. [8] demonstrated that the optimal kernel size differs among subjects and may vary over time for the same subject. Simultaneously conducting convolution at multiple scales and then aggregating allow for the extraction of features at different scales, resulting in richer temporal features and implying more accurate classification judgments during the final decision-making process. Define the ratio as $\alpha = \{\alpha^i \mid i = 1, 2, 3\}$, where i represents three parallel CNN layers. Hence, the kernel size is denoted as:

$$k^i = \left(1, \alpha^i \cdot f_s\right), i = 1, 2, 3 \tag{1}$$

where f_s is the sampling rate of the EEG signals. Then, three outputs with different scale temporal representations are concatenated and fed into the Dense Block for association and fusion.

2) Dense Block: The Dense Block consists of four CNN layers and two average pooling layers, fusing the concatenated temporal features and further refining useful features. To enhance information flow among the CNN layers, the outputs of each CNN filter were propagated to all subsequent layers, which generated the final output incorporating the extracted features from all preceding layers [41]. The connection between two common CNN layers is:

$$x_l = F_l(x_{l-1}) \tag{2}$$

where x_{l-1} and x_l are the input and output of the layer l. In the dense block, the lth layer receives the feature maps from all preceding layers:

$$x_l = F_l([x_0, x_1, ..., x_{l-1}]) \tag{3}$$

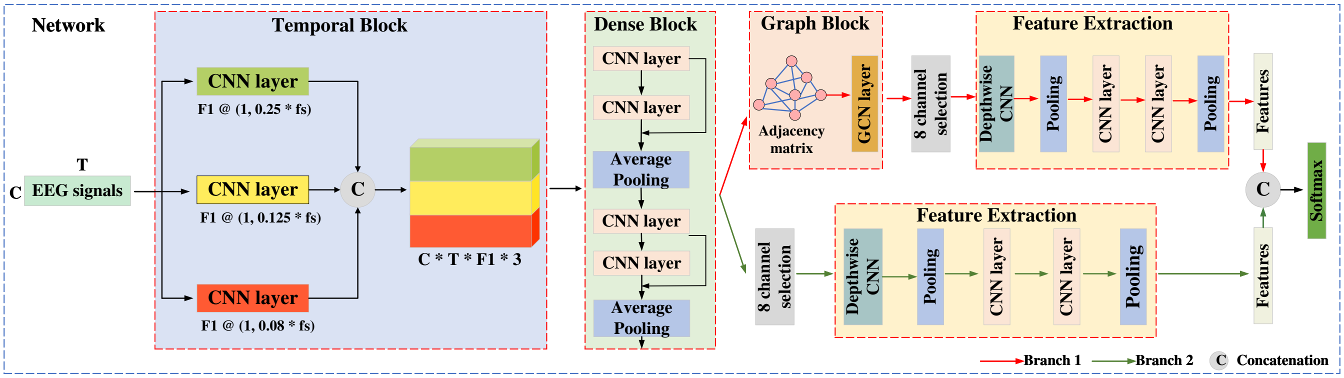


Fig. 2: The detailed structure of the proposed model.

where $[x_0, x_1, ..., x_{l-1}]$ are the feature maps before the layer l. In the experiment, we used two CNN layers and one pooling layer as one combination. Therefore, the output averaged feature maps from two preceding CNN layers:

$$x_l = F_{Average}\left(F_{l-1}\left(x_{l-1}\right) + F_{l-2}\left(x_{l-2}\right)\right) \tag{4}$$

Such connections create short paths which enhance the flow of information and feature reuse. Meanwhile, each CNN layer utilizes ELU as the activation function, followed by batch normalization and dropout techniques to suppress the overfitting problem. One CNN layer generates k feature maps contributing to the subsequent layer. Here we set $k = 10$ so that each combination produces 20 feature maps.

3) Graph Block: An undirected and weighted graph can be described as $G = (V, E)$ where V represents the nodes and E denotes the edges among the nodes. In the proposed model, each EEG channel is regarded as a node of the graph while edges are the relationship between channels. The adjacency matrix $W \in R^{N\times N}$ is built to describe the connection relationship between different nodes, where N is the number of channels. Nevertheless, the intricate nature of the activation states in the human brain during MI tasks poses a challenge in constructing an artificial matrix based on prior knowledge. Ma et al [23] learned the channel similarity based on semi-supervised learning and then manually selected 11 channels as inputs. EEG-GENet [17] set the edge between neighboring channels to 1 and those not neighboring to 0. Delvigne et al [42] used the distance as prior knowledge to create an adjacency matrix. To better adapt to the characteristics of end-to-end learning procedure in a DL model, the channel connections based on the temporal features learned for each channel are dynamically learned, ensuring a trainable adjacency matrix.

First, Pearson's correlation matrix (PCM) is adopted to initialize the matrix. PCM is an effective tool to capture the topological information among EEG channels [43] [17]. If one trial EEG data is defined as $X \in R^{C\times T}$ with C channels and T temporal features, the Pearson's correlation coefficient can be obtained by:

$$P_{ij} = \frac{cov\left(X_i, X_j\right)}{\sqrt{var\left(X_i\right) var\left(X_j\right)}} \tag{5}$$

where i and j denotes the i^{th} and j^{th} channel of the EEG signals. Therefore, the initialized adjacency matrix is:

$$A_{initial} = \begin{bmatrix} P_{1,1} & \cdots & P_{1,C} \\ \vdots & \ddots & \vdots \\ P_{C,1} & \cdots & P_{C,C} \end{bmatrix} \tag{6}$$

To make the adjacency matrix trainable and dynamically analyze the similarity among channels, a mask matrix of the same size consisting of trainable parameters is adopted:

$$A_{trainable} = \begin{bmatrix} W_{1,1} & \cdots & W_{1,C} \\ \vdots & \ddots & \vdots \\ W_{C,1} & \cdots & W_{C,C} \end{bmatrix} \tag{7}$$

where w is the weight initialized based on xavier uniform [44]. To make the symmetric trainable matrix, $A_{trainable}$ and its transposed are multiplied and applied to $A_{initial}$:

$$A = \Phi_{relu}\left(A_{Initial} \odot \left(A_{trainable} \cdot A_{trainable}^{T}\right)\right) + I \tag{8}$$

where Relu activation is employed to ensure the matrix is non-negative. The degree matrix is $\widetilde{D} = \sum_j A_{ij}, i \neq j$. The normalized adjacency matrix can be calculated as:

$$\widetilde{A} = \widetilde{D}^{-\frac{1}{2}} A \widetilde{D}^{-\frac{1}{2}} \tag{9}$$

Once the matrix $\widetilde{A}$ is obtained, a GCN layer with weights and bias vector is applied on the input feature maps:

$$X_{output} = \Phi_{elu}\left(\widetilde{A}XW + bias\right) \tag{10}$$

After the graph block operation, each EEG channel includes the information aggregated from other channels.

4) Channel Selection and Feature Extraction: If the input EEG data has 62 channels, a channel selection step is required; otherwise, it is not necessary. The channel selection procedure chooses the specific 8 channels (FC6, C4, Fz, C3, FC5, CP6, Cz and CP5) (Fig. 1(b)) same in the target domain data. When BCI-IV-2a dataset is adpoted as the target domain data, the channels that are common to both KU and BCI-IV-2a datasets will be selected. Subsequently, both the branch with the graph block and the branch without the graph block undergo further feature extraction. The depthwise CNN layer helps the model to extract global spatial features while reducing computational complexity compared with common CNN layers. Then two average pooling layers and CNN layers follow to fuse the feature maps and decrease dimensionality. The one with graph block aggregates other channels' features and topological information while the other one focuses on mining temporal-spatial features. The different views of feature representations brought by two parallel branches help enhance the model's robustness and classification performance.

D. Training Procedure

The training procedure was initially conducted on the source domain dataset, employing Knowledge Distillation and a two-stage training strategy [45] to build a pre-trained model. Then this model was validated on the target domain with fine-tuning technology.

In the source domain, the data consists of 62 channels, while the target domain has only 8 channels available. To leverage the extensive data in the source domain and transfer model parameters to the target domain, our proposed model is designed to learn feature representations and distributions by aggregating information from all 62 channels. However, the model is currently configured to accept 62-channel inputs, making it unsuitable for direct use in the target domain with only 8 channels. Training the model with the limited 8-channel data in the source domain would result in the loss of channel aggregation information. To address this issue, we employ a Knowledge Distillation framework, constructing both a teacher network and a student network with different numbers of channels as input (Fig. 3). First, all subject data in the source domain are divided into training data (80% of the total dataset) and validation data (the remaining 20% of the total dataset). In the first stage, the teacher network was trained with 62-channel data from the training set in the source domain as inputs and incorporated the channel selection step during training. The early-stopping tool monitored the validation set accuracy and stopped the training if there was no increase in the next 150 epochs. Then the best validation accuracy are saved for the next stage. In the second stage, the student network was trained with 8-channel data or 21-channel according to the number of channels in the target domain dataset from all source domain data (training set + validation set). During this procedure, the student network did not need the channel selection step but guidance from the teacher network to align the feature distributions between the two networks. A Mean Squared error loss was used to calculate the distance among the feature maps after the graph block:

$$\mathcal{L}_{mse} = \frac{1}{n}\sum_{i-1}^{n}\left(x_{teacher} - x_{student}\right)^2 \tag{11}$$

where n is the number of trials and x are the feature maps.$x_{teacher}$ had the channel aggregation knowledge based on 62-channel while $x_{student}$ had no other channels information. By minimizing $\mathcal{L}_{mse}$, the student network learned the feature representations with aggregation knowledge and abundant topological information extracted by the graph block in the teacher network. When it was fine-tuned in the target domain, useful pre-trained parameters were transferred effectively. The second stage used the saved validation set and stopped when the validation accuracy was higher than the one recorded in stage one. Even if the final accuracy can not reach the same value, the model will stop with the early stopping criteria to prevent the occurrence of infinite training.

Subsequently, the student network was used as a pre-trained model being validated in the target domain. Despite the model learning to classify MI tasks using few common channels in the source domain, calibration based on limited target data was still necessary due to the utilization of two datasets from entirely different devices. The source domain dataset employs wet electrodes with good data quality, while the target domain dataset uses dry electrodes with an impedance reaching 300 $k\Omega$, resulting in poor data quality. Directly using the pre-trained model without verification yields poor classification model accuracy. Hence, fine-tuning is applied, retraining the parameters of the model based on little target domain data. We design two scenarios in the target domain validation: 1) A 5-fold cross-validation (CV) was employed with 3 folds for training, 1 fold for validation, and the rest for testing (Fig. 4(a)). 2) The dataset is split into two sessions, with one session for training while the other one for testing. The cross-session modeling, despite the data coming from the same subject, still encounter significant differences between sessions. Moreover, the limited amount of trainable data brings substantial challenges for cross-session MI classification. In this experiment, the aforementioned two-stage training approach will also be employed. During the training session, 75% of the data from the first session will be used for the first stage of training, while the remaining 25% will serve as validation data. In the second stage, all training session data will be fed into the model for training (Fig. 4(b)).

Since each model contains several operation layers, whether the parameters were frozen is based on the functionality of each block. Five schemes for fine-tuning strategy were conducted in the experiment (Fig.5). Different schemes froze different blocks in the pre-trained model while the rest blocks were adaptive and re-trained based on limited data in the target domain. Scheme 1 froze all layers namely no data in the target domain was used in training. This scheme no longer required any new data, greatly reducing validation time. However, it resulted in a decrease in classification accuracy due to significant data distribution differences caused by different EEG acquisition devices. Scheme 5 made all layers adaptive namely all parameters in the model were updated to match the target data.

E. Training Setup

The cross-entropy function was adopted to evaluate the distance between the probability distribution of the model prediction values y_p and the true labels y_t:

$$L\left(y_p, y_t\right) = -\sum_{m} y_{p,m} \log y_{t,m}. \tag{12}$$

where m is the index of y. Adam optimizer was used with 0.001 as the learning rate. The computer used in this experiment had 22 Intel processors and 80 GB RAM. GTX 4090 GPU with 24 GB memory was used for training and testing MI-EEG signals. Pytorch 1.10.0 was used for building the proposed model.

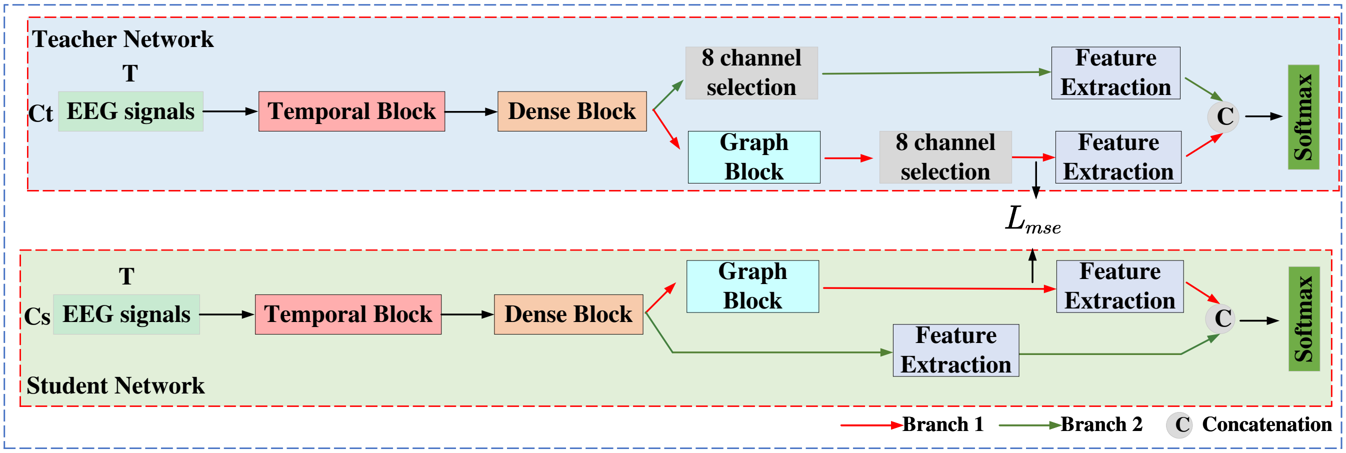
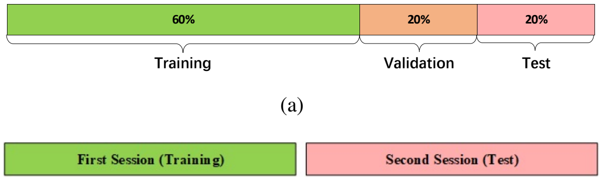


Fig. 3: The framework of the Knowledge Distillation.



(a)

(b)

Fig. 4: Two scenario descriptions. (a) Scenario 1 with 5-fold CV, (b) Scenario 2 with cross-session modelling

III. Results

We use two traditional machine learning methods (CSP [3] and FBCSP [5]), three CNN-based models (Shallow ConvNet [6], Deep ConvNet [6], and EEGNet [7]), two GCN-based models (EEG-GENet [17] and EEG-ARNN [22]) and two transformer-based models (EEGConformer [46] and ATCNet [14])as benchmarks to demonstrate the effectiveness of our proposed method. All the baseline methods used the parameters and structures suggested by their authors for a fair comparison. The details of the baseline models are described as follows:

1) Machine learning methods: CSP and FBCSP are two classic machine learning algorithms. The core idea is to find a set of optimal spatial filters that can separate features after projection. FBCSP goes a step further by dividing the data into multiple sub-bands and identifying informative and discriminative pairs of sub-bands. Both models are lightweight, easily modifiable, and widely applied. In the experiment, the Support Vector Machine (SVM) was employed as the classifier.
2) CNN-based models: Shallow ConvNet, Deep ConvNet and EEGNet have excellent performance on MI-EEG classification and robustness. They utilize a 1-D CNN and a deepwise CNN layer to extract temporal-spatial features. Then, the Deep ConvNet model combines several common CNN layers and pooling layers before the classifier while the Shallow ConvNet only adopts a squaring layer with the log operation. EEGNet uses the pointwise CNN layer to reduce amounts of calculation resources while ensuring informative learned features.
3) GCN-based models: The EEG-GENet model is built based on the structure of EEGNet. After extracting the temporal features by a CNN layer, a GCN layer is followed to capture the topology information according to the EEG electrodes. EEG-ARNN combines one CNN layer and one average pooling layer as a module. The GCN layers are added after each module with a trainable adjacency matrix which is initialized with one. Both of them perform well on the public BCI-IV-2a dataset [47].
4) Transformer-based models: EEGConformer compacts convolutional Transformer to learn local and global features from EEG signals. The CNN module is adopted to capture low-level temporal features while self-attention module is used to extract the correlation within temporal features. ATCNet employs a CNN-based sliding window approach to efficiently augment MI data and enhance the performance of MI classification [14].

A. Overall performance

We conducted the experiments based on the two scenarios on the 8-channel dataset and BCI-IV-2a public dataset. To ensure fairness as much as possible, we used the original preprocessing methods and parameters of the comparison algorithms. During transfer, all parameters of both the comparison algorithms and the proposed algorithm were trainable to maximize the model's performance. Before transfer, ShallowConvNet performed best in the 5-fold CV scenario which reached to 88.89% while the proposed model was only 0.98% lower than that of theirs. ATCNet performed best in the cross-session scenario which was 2.35% higher than our model. However, after applying fine-tuning technology on the models, the proposed model with aggregated channel information from teacher network achieved the highest score in the both of two scenarios. The statistical significance tests including an Analysis of Variance (ANOVA) test and paired t-tests showed that the proposed model performed better than the baseline model ($p < 0.05$) except the ATCNet. The table I showed that almost all deep learning models got better results than the

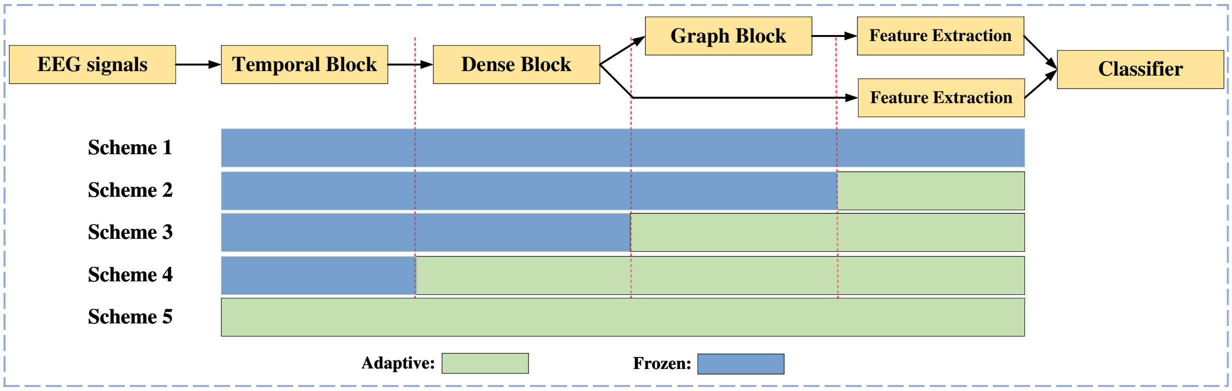


Fig. 5: The schemes of the fine-tuning framework.

TABLE I: Comparison of classification accuracy and kappa value on the BCI-IV-2a dataset.

	BCI-IV-2a Dataset							
	Before transfer				After transfer			
	cross-validation		cross-session		cross-validation		cross-session	
	accuracy (%)	kappa value	accuracy (%)	kappa value	accuracy (%)	kappa value	accuracy (%)	kappa value
CSP [3]	73.74	0.477	71.68	0.434	/	/	/	/
FBCSP [5]	74.58	0.528	82.18	0.644	/	/	/	/
ShallowConvNet [6]	**88.89**	**0.778**	83.20	0.664	90.42	0.808	86.73	0.735
DeepConvNet [6]	70.29	0.406	67.21	0.334	88.89	0.778	81.40	0.628
EEGNet [7]	83.78	0.676	81.20	0.621	87.78	0.756	80.79	0.616
EEG-GENet [17]	74.89	0.498	65.43	0.309	81.81	0.636	72.84	0.456
EEG-ARNN [22]	61.40	0.113	58.48	0.170	89.02	0.781	84.18	0.684
EEGConformer [46]	84.95	0.667	79.63	0.593	84.17	0.683	78.47	0.570
ATCNet [14]	80.71	0.615	**85.99**	**0.718**	91.30	0.826	85.88	0.718
Proposed model	87.91	0.758	83.64	0.673	**91.81**	**0.836**	**87.27**	**0.745**

TABLE II: Comparison of classification accuracy and kappa value on the 8-channel dataset.

	8-channel Dataset							
	Before transfer				After transfer			
	cross-validation		cross-session		cross-validation		cross-session	
	accuracy (%)	kappa value	accuracy (%)	kappa value	accuracy (%)	kappa value	accuracy (%)	kappa value
CSP [3]	51.2	0.04	54.54	0.091	/	/	/	/
FBCSP [5]	64.15	0.227	62.27	0.246	/	/	/	/
ShallowConvNet [6]	50.45	0.009	53.41	0.068	56.42	0.139	56.96	0.139
DeepConvNet [6]	55.68	0.114	57.39	0.148	56.02	0.12	62.78	0.256
EEGNet [7]	55.51	0.11	56.53	0.131	58.52	0.17	58.66	0.173
EEG-GENet [17]	54.60	0.092	56.53	0.131	56.59	0.179	58.95	0.179
EEG-ARNN [22]	53.98	0.176	63.35	0.267	62.27	0.245	58.81	0.176
EEGConformer [46]	56.82	0.136	56.53	0.131	58.81	0.176	57.95	0.159
ATCNet [14]	63.40	0.268	60.62	0.213	67.27	0.345	65.20	0.310
Proposed model	**67.67**	**0.353**	**65.20**	**0.303**	**71.19**	**0.42**	**69.03**	**0.381**

traditional machine learning methods in the BCI-IV-2a dataset. In the 8-channel Dataset, the proposed model performed best in all scenarios. Before transfer, the proposed model achieved 67.67% and 65.20% in the CV and cross-session scenarios respectively. With the help of aggregated information learned by knowledge distillation and fine-tuning technology, the proposed model's classification accuracy improved by 3.52% and 2.7% before and after transfer respectively. In the traditional machine learning methods, FBCSP performed best, surpassing even other DL models. The possible reason is that the 8-channel dataset is more sensitive to frequency band filtering. FBCSP was the only method among these baseline models that divided the original data into multiple sub-bands and performed band selection. In the DL models, ATCNet performed better and got an accuracy of 67.27% and 65.2% after transfer. In comparison, the proposed model improved by 3.92% and 3.83% over the ATCNet in two different scenarios.

B. Analysis of Different Schemes

To further validate the influence of different blocks in the model during the fine-tuning process, we conducted experiments based on different schemes (Fig. 5). The box plot (Fig. 6) illustrated that the more parameters involved in adaptive tuning, the better the model's performance. Scheme 1 froze all layers so that no weights could be updated to adapt to the target domain, leading to the worst classification accuracy. The data distribution divergence across datasets and devices limited the model's performance and robustness. Although updating all parameters increases the computational load, achieving a 71% accuracy in a limited dataset collected from only 8 dry electrodes is worthwhile and makes it effectively applicable to portable devices.

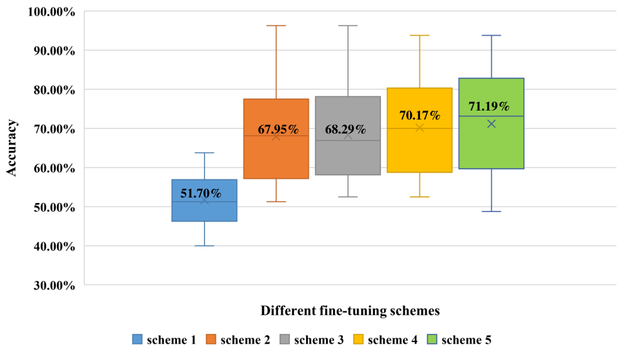


Fig. 6: Different schemes of the fine-tuning framework.

C. Analysis of Training Proportion

In practical applications of BCI, obtaining a large amount of data has always been challenging. Therefore, calibrating the model with little or no data is necessary. In the 8-channel dataset, only 80 trials were collected for each individual, significantly less than the data in public datasets. Following Scenario 2, we divided the entire 8-channel dataset into two sessions, using one session for training and the other one for testing. The training set was used to adapt the parameters of the pre-trained model based on fine-tuning method. The training data started from 25% and increased by 25% of the total data volume in each experiment, up to 100%, for a total of 4 experiments. The result (shown in Fig. 7) indicated that with more training data used for adaptation, the model's classification performance improved. With 25% of the data namely 12 trials used for adaptation, the model's classification accuracy can reach to 59.23%. When the training data volume reached 75% of the whole training dataset, the classification accuracy was 65.06% which surpassed the results of most models based on the cross-session scenario. When using all data in the training session, the classification result reached to 69.03% which was higher than the 5-fold CV before transfer. Therefore, with the assistance of fine-tuning techniques, the proposed model can achieve performance surpassing the use of within-subject models in the target domain, even with the adaptation and calibration using a small amount of target domain data, further demonstrating the practicality of the proposed model.

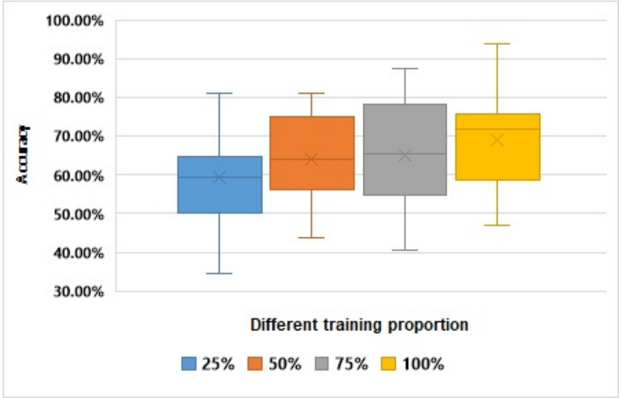


Fig. 7: The results using different training data volumes.

TABLE III: The classification accuracy (%) of the ablation study

	W/O D_Block	W/O G_Blcok	Proposed model
Avg	65.79	70.11	71.19
Std	13.18	13.20	13.30

D. Ablation Study

To validate the contribution of the Dense Block and Graph Block which were important components in the proposed model, an ablation study was conducted: 1) Without Dense Block (W/O D_Block): Dense blocks were utilized to capture information from the fused feature maps extracted by the Temporal Block. The Dense Block included 4 CNN layers and 2 average pooling layers. All of them were abandoned in the ablation experiment. 1) Without Graph Block (W/O G_Block): Graph Block was adopted to learn the topological information based on electrodes and transfer the knowledge from data with 62 channels to data with 8 channels. In the ablation experiment, we prohibited the transmission of topological information learned by the graph block in the knowledge distillation framework. The results in Table III showed that the proposed model had an accuracy of 5.4% $(p = 0.02)$ and 1.08% $(p = 0.03)$ higher than the W/O D_Block model and W/O G_Block, respectively. The Dense Block contributed more because it refined the temporal block and involved more parameters while only one GCN layer existed in the Graph Block.

E. Influence of Aggregated Channels

The source domain has 62 channels while the target domain only has 8 channels. To transmit the topological information learned from 62-channel data, the features from the rest of the k channels were aggregated in the GCN layer. To validate the influence of K-aggregated channels, we adjusted the adjacency matrix. First, C_s defined as one of the 8 specific channels was selected. Then we sorted the other channels that were not included in these 8 channels in descending order based on the weights obtained by the trainable adjacency matrix. The m channels with the smallest weights, specifically those least correlated with these 8 specific channels, were set to 0 in the adjacency matrix, ensuring that the GCN layer did not consider their information when aggregating channel features. We conducted 7 experiments from using all 62 channels to only 9 channels. The result in Fig. 8 indicated that when the number of fused channels decreased and the limited information was captured, the overall classification performance of the model also decreased accordingly.

F. Visualization

1) Adjacency matrix visualization: We recorded the weights of the trainable adjacency matrix to validate the channel relationships learned by the proposed model. Fig. 9 shows the heatmaps of the adjacency matrix in the teacher network model which was trained based on the source domain. In Fig. 9(a), These 8 channels were only connected to themselves because the initialization matrix included a self-loop step. Following

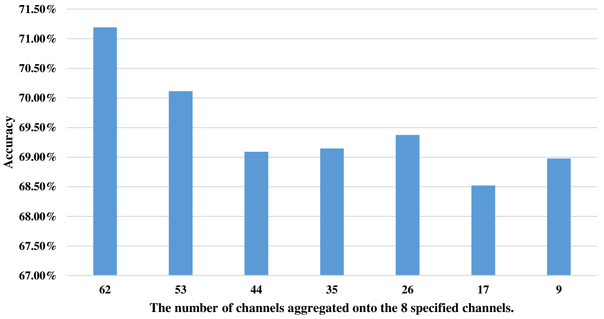


Fig. 8: The accuracy of the proposed model with different numbers of aggregated channels.

training and validation based on early stopping criteria, the channels established relationships with each other and aggregated based on the trained adjacency matrix, contributing to the final classification of MI tasks (Fig. 9(b)). FC6, FC5, C3, C4 and Cz have stronger connections with their neighboring electrodes. For instance, connections between 1) FC6 and (FC4 and FC2), 2) FC5 and F9, 3) C3 and (C1 and FC1), 4) Cz and (CP2 and C2), and 5) C4 and (FC2 and C2). CP5, CP6 and Fz have more relations with channels P7, P3, PO3 and PO4 which belong to the parietal and parieto-occipital lobes. It can be observed that many channels strongly correlated with these specific 8 channels are not part of the same set. Without aggregating by the GCN layer, the information associated with these correlated channels will be missing, leading to a decrease in model classification accuracy.

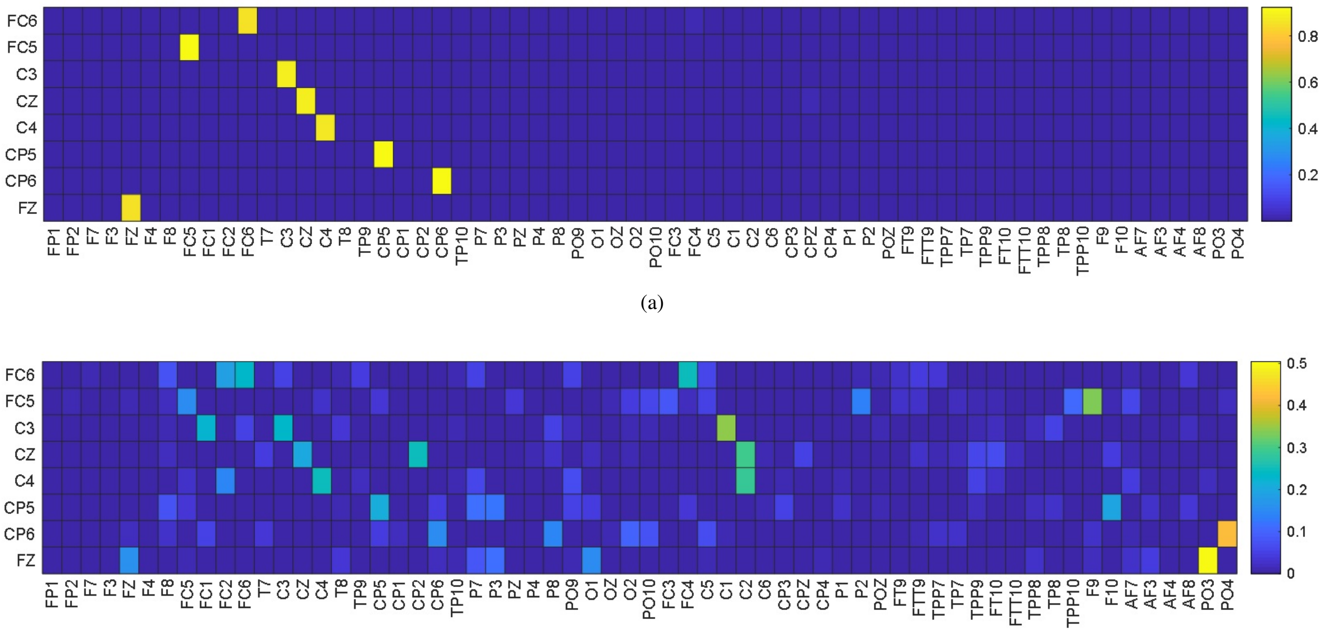
During the knowledge distillation procedure, the student network was guided by the teach network and better extracted the features based on the 8-channel inputs in the source domain. Compared with the trained adjacency matrix (Fig. 10(b)) and untrained matrix (Fig. 10(a)), some connections were strengthened like C3 and Cz. Fig. 10(c) is the fine-tuned model of the 21^{st} subject which reached an accuracy of 93.75% in the 8-channel dataset. Compared with the pre-trained model (Fig. 10(b)), the fine-tuning method allowed the adjacency matrix to further adapt to the 8-channel dataset and reconstruct the whole relations among channels. The channels Cz and C3 still maintained a strong connection while the relationships between CP6 and FC5, CP6 and C3 decreased. Some connections were activated like Cz and Fz, CP5 and CP6, and C4 and FC5. Due to the fine-tuning method applied to each subject in the pre-trained model, the reconstructed relationships among channels varied. However, further research is needed to explore the relationship between the activated channels and the classification performance of each within-subject model.

2) Feature Visualization: The t-distributed Stochastic Neighbor Embedding (t-SNE) method was utilized to visualize the feature maps of the fully connected layer before the final classifier of the proposed model. Fig. 11(a) and Fig. 11(b) are the teacher network and student network trained in the source domain. Fig. 11(c) is the feature map of the fine-tuned model based on the 21^{st} subject in the target domain. Each subject in the target domain only has 80 trials so the limited points are shown in Fig. 11(c). Based on the t-SNE analysis, the proposed model demonstrated strong capabilities in EEG signal classification. The classification boundaries of the teacher network's feature map appear more distinguishable than those of the student network. One reason for this is that the teacher network took the data with 62 channels as input, incorporating more information, while the student network only had 8 channels as input. Although the student network learned topological features from the teacher network, there was still a slight loss in classification performance.

IV. Discussion

With the proliferation of portable devices, the research and application of BCI has gained much more momentum. In real-life applications, it is challenging to collect large amounts of high-quality data, and there is a strong demand for reducing experimental preparation time. Therefore, it is crucial to ensure excellent accuracy in MI-task classification while reducing calibration time and the amount of required training data. DL models have shown promising results in decoding EEG signals, and transfer learning has been effectively applied to shorten verification times. However, few models can be generalized across datasets, especially when the datasets are collected using different devices with different channels. The target data in our experiment were collected using dry electrode devices, which have limited quantity, lower quality, and a very restricted number of channels, making it challenging to directly use the models trained with past public datasets. Therefore, we first utilized GCN to learn the topological knowledge of EEG channels on a public dataset with 62 channels. Subsequently, through a knowledge distillation framework, the feature distribution obtained from classifying MI tasks based on 62-channel data in the source domain was adopted to guide the proposed model with 8-channel inputs. Finally, the pre-trained model employed fine-tuning for adapting target domain data.

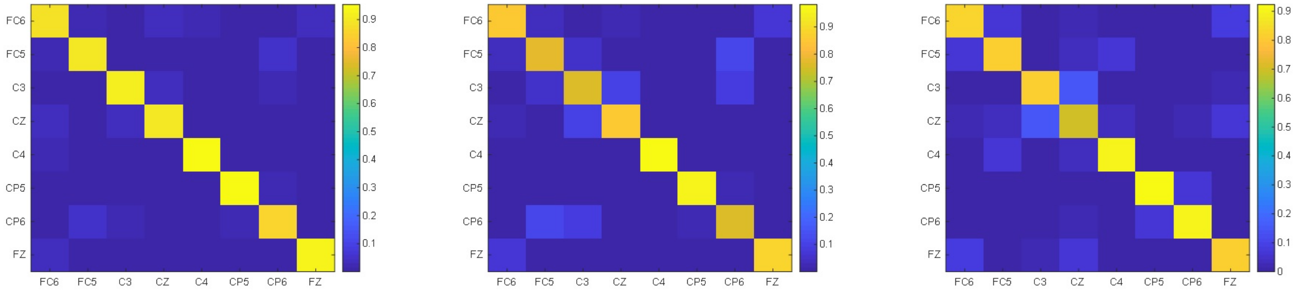
In our experiments, the proposed model achieved the highest classification accuracy compared with machine learning methods, CNN-based, GCN-based and other transformer-based models. To better validate the practicability of the model, two scenarios including cross-validation and cross-session were conducted to examine the model's performance and robustness. Besides that, we found that the model achieved a classification accuracy of 65.06% when the training data only constituted 75% of the training session, which was higher than the results of most baseline models using the whole training session. The model with fine-tuning technology built based on the source domain has a classification accuracy of 3.52% and 2.7% higher than the model built before the transfer, demonstrating the effectiveness of the transfer learning. The different schemes of fine-tuning also influenced the model performance. The more parameters involved in the adaptation, the better the model performed. Besides that, we also validated how the number of aggregated channels affects the model performance. When the teacher network captures features from more channels and guides the student network, the final fine-tuned model will achieve higher classification accuracy.



(a)

(b)

Fig. 9: The heatmaps of the adjacency matrix in the teacher network model: (a) Untrained model, (b) Trained model.

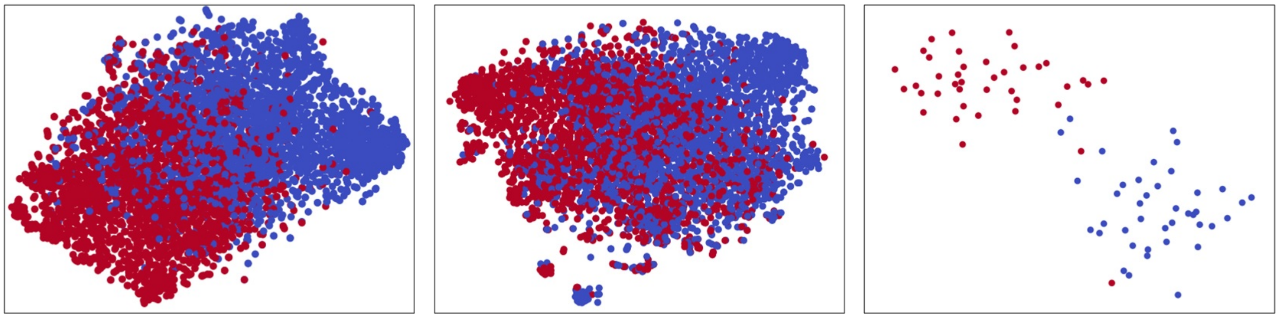


(a)

(b)

(c)

Fig. 10: The heatmaps of the adjacency matrix: (a) Untrained model (Student network), (b) Trained model (Student network), (c) Fine-tuned model (the 21^{st} subject in the 8-channel dataset).



(a)

(b)

(c)

Fig. 11: The feature map of the proposed model: (a) Teacher network, (b) Student network, (c) Fine-tuned model.

Although the proposed model has achieved the best classification performance than the state-of-the-art models and validated the design of the framework, there are still limitations. For instance, the adjacency matrix was trainable to dynamically capture the topological information, but the construction of the graph was simply initialized using the PCC, lacking prior knowledge, including the relationship between the motor brain area and the other brain areas, as well as individual channel connectivity. When using knowledge distillation, the transfer of knowledge from graph convolutions to temporal-spatial features is insufficient. Further research is needed on how to better integrate multi-channel information into a reduced number of channels. At the same time, we only considered binary classification for cross-dataset modeling and did not address multi-class cross-scenario modeling, which is also an important challenge for the future practical application of brain-computer interfaces.

V. Conclusion

This paper has proposed GCN based transfer learning method for cross-dataset MI EEG decoding. The proposed model combines both the CNN and GCN layers, aggregating topological information from 62 channels into only 8 specific channels and guiding a pre-trained model by knowledge distillation. Fine-tuning technology has been used to adapt the target dataset. The results show that the proposed model achieved an accuracy of 71.19% in the cross-validation and 69.03% in the cross-session scenario, at least 3.92% and 3.83% higher than the state-of-the-art approaches across-dataset. The results obtained based on the public BCI-IV-2a dataset also showed good results in both of the two scenarios. The feature visualization and heatmaps indicate excellent performance of the proposed model on EEG decoding and channel relation reconstruction, demonstrating its potential to enhance the effectiveness of BCI applications with portable devices.

References

[1] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for eeg-based brain–computer interfaces," *Journal of neural engineering*, vol. 4, no. 2, p. R1, 2007.

[2] R. Mane, T. Chouhan, and C. Guan, "Bci for stroke rehabilitation: motor and beyond," *Journal of neural engineering*, vol. 17, no. 4, p. 041001, 2020.

[3] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1123–1134, 2001.

[4] S. Lemm, B. Blankertz, G. Curio, and K.-R. Muller, "Spatio-spectral filters for improving the classification of single trial eeg," *IEEE transactions on biomedical engineering*, vol. 52, no. 9, pp. 1541–1548, 2005.

[5] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (fbcsp) in brain-computer interface," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, 2008, pp. 2390–2397.

[6] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[7] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018.

[8] G. Dai, J. Zhou, J. Huang, and N. Wang, "Hs-cnn: a cnn with hybrid convolution scale for eeg motor imagery classification," *Journal of neural engineering*, vol. 17, no. 1, p. 016025, 2020.

[9] R. Mane, N. Robinson, A. P. Vinod, S.-W. Lee, and C. Guan, "A multi-view cnn with novel variance layer for motor imagery brain computer interface," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 2950–2953.

[10] C. Ju and C. Guan, "Tensor-cspnet: A novel geometric deep learning framework for motor imagery classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] 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 Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1534–1545, 2021.

[13] J. Zhang, K. Li, B. Yang, and X. Han, "Local and global convolutional transformer-based motor imagery eeg classification," *Frontiers in Neuroscience*, vol. 17, 2023.

[14] H. Altaheri, G. Muhammad, and M. Alsulaiman, "Physics-informed attention temporal convolutional network for eeg-based motor imagery classification," *IEEE transactions on industrial informatics*, vol. 19, no. 2, pp. 2249–2258, 2022.

[15] ——, "Dynamic convolution with multilevel attention for eeg-based motor imagery decoding," *IEEE Internet of Things Journal*, vol. 10, no. 21, pp. 18 579–18 588, 2023.

[16] Y. Ding, N. Robinson, C. Tong, Q. Zeng, and C. Guan, "Lggnet: Learning from local-global-graph representations for brain–computer interface," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[17] H. Wang, H. Yu, and H. Wang, "Eeg_genet: A feature-level graph embedding method for motor imagery classification based on eeg signals," *Biocybernetics and Biomedical Engineering*, vol. 42, no. 3, pp. 1023–1040, 2022.

[18] Y. Wu, D. Lian, Y. Xu, L. Wu, and E. Chen, "Graph convolutional networks with markov random field reasoning for social spammer detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 01, 2020, pp. 1054–1061.

[19] Y. Wang, S. Fang, C. Zhang, S. Xiang, and C. Pan, "Tvgcn: Time-variant graph convolutional network for traffic forecasting," *Neurocomputing*, vol. 471, pp. 118–129, 2022.

[20] M. Graña and I. Morais-Quilez, "A review of graph neural networks for electroencephalography data analysis," *Neurocomputing*, p. 126901, 2023.

[21] B. Vivek, A. Adarsh, J. Gubbi, K. Muralidharan, R. K. Ramakrishnan, and A. Pal, "St-gnn for eeg motor imagery classification," in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2022, pp. 01–04.

[22] B. Sun, Z. Liu, Z. Wu, C. Mu, and T. Li, "Graph convolution neural network based end-to-end channel selection and classification for motor imagery brain-computer interfaces," *IEEE transactions on industrial informatics*, 2022.

[23] W. Ma, C. Wang, X. Sun, X. Lin, and Y. Wang, "A double-branch graph convolutional network based on individual differences weakening for motor imagery eeg classification," *Biomedical Signal Processing and Control*, vol. 84, p. 104684, 2023.

[24] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 843–852.

[25] X. Xie, S. Sun, H. Chen, and J. Qian, "Domain adaptation with twin support vector machines," *Neural Processing Letters*, vol. 48, pp. 1213–1226, 2018.

[26] W. Hang, W. Feng, R. Du, S. Liang, Y. Chen, Q. Wang, and X. Liu, "Cross-subject eeg signal recognition using deep domain adaptation network," *IEEE Access*, vol. 7, pp. 128 273–128 282, 2019.

[27] Y. Chen, R. Yang, M. Huang, Z. Wang, and X. Liu, "Single-source to single-target cross-subject motor imagery classification based on multisubdomain adaptation network," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 1992–2002, 2022.

[28] L.-l. Chen, A. Zhang, and X.-g. Lou, "Cross-subject driver status detection from physiological signals based on hybrid feature selection and transfer learning," *Expert Systems with Applications*, vol. 137, pp. 266–280, 2019.

[29] J. Zhang, Y. Wang, and S. Li, "Cross-subject mental workload classification using kernel spectral regression and transfer learning techniques," *Cognition, Technology & Work*, vol. 19, pp. 587–605, 2017.

[30] D.-K. Han and J.-H. Jeong, "Domain generalization for session-independent brain-computer interface," in *2021 9th International Winter Conference on Brain-Computer Interface (BCI)*. IEEE, 2021, pp. 1–5.

[31] S. An, S. Kim, P. Chikontwe, and S. H. Park, "Dual attention relation network with fine-tuning for few-shot eeg motor imagery classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[32] F. Liu, P. Yang, Y. Shu, N. Liu, J. Sheng, J. Luo, X. Wang, and Y.-J.

Liu, “Emotion recognition from few-channel eeg signals by integrating deep feature aggregation and transfer learning,” *IEEE Transactions on Affective Computing*, 2023.

[33] M. Soufineyestani, D. Dowling, and A. Khan, “Electroencephalography (eeg) technology applications and available devices,” *Applied Sciences*, vol. 10, no. 21, p. 7453, 2020.

[34] T. Zaremba and A. Atyabi, “Cross-subject & cross-dataset subject transfer in motor imagery bci systems,” in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–8.

[35] L. Xu, M. Xu, Y. Ke, X. An, S. Liu, and D. Ming, “Cross-dataset variability problem in eeg decoding with deep learning,” *Frontiers in human neuroscience*, vol. 14, p. 103, 2020.

[36] Y. Xie, K. Wang, J. Meng, J. Yue, L. Meng, W. Yi, T.-P. Jung, M. Xu, and D. Ming, “Cross-dataset transfer learning for motor imagery signal classification via multi-task learning and pre-training,” *Journal of Neural Engineering*, vol. 20, no. 5, p. 056037, 2023.

[37] P. L. C. Rodrigues, C. Jutten, and M. Congedo, “Riemannian procrustes analysis: Transfer learning for brain–computer interfaces,” *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 8, pp. 2390–2401, 2018.

[38] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge distillation: A survey,” *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 2021.

[39] M.-H. Lee, O.-Y. Kwon, Y.-J. Kim, H.-K. Kim, Y.-E. Lee, J. Williamson, S. Fazli, and S.-W. Lee, “Eeg dataset and openbmi toolbox for three bci paradigms: An investigation into bci illiteracy,” *GigaScience*, vol. 8, no. 5, p. giz002, 2019.

[40] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, “Bci competition 2008–graz data set a,” *Institute for knowledge discovery (laboratory of brain-computer interfaces), Graz University of Technology*, vol. 16, pp. 1–6, 2008.

[41] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[42] V. Delvigne, H. Wannous, T. Dutoit, L. Ris, and J.-P. Vandeborre, “Phydaa: Physiological dataset assessing attention,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2612–2623, 2021.

[43] Y. Hou, S. Jia, X. Lun, Z. Hao, Y. Shi, Y. Li, R. Zeng, and J. Lv, “Gcns-net: a graph convolutional neural network approach for decoding time-resolved eeg motor imagery signals,” *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[44] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[45] R. Mane, E. Chew, K. Chua, K. K. Ang, N. Robinson, A. P. Vinod, S.-W. Lee, and C. Guan, “Fbcnet: A multi-view convolutional neural network for brain-computer interface,” *arXiv preprint arXiv:2104.01233*, 2021.

[46] Y. Song, Q. Zheng, B. Liu, and X. Gao, “Eeg conformer: Convolutional transformer for eeg decoding and visualization,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 710–719, 2022.

[47] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. Mueller-Putz *et al.*, “Review of the bci competition iv,” *Frontiers in neuroscience*, p. 55, 2012.