2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Generalizable Movement Intention Recognition with Multiple Heterogeneous EEG Datasets

Xiao Gu, Jinpei Han, Guang-Zhong Yang, Benny Lo

***Abstract*— Human movement intention recognition is important for human-robot interaction. Existing work based on motor imagery electroencephalogram (EEG) provides a non-invasive and portable solution for intention detection. However, the data-driven methods may suffer from the limited scale and diversity of the training datasets, which result in poor generalization performance on new test subjects. It is practically difficult to directly aggregate data from multiple datasets for training, since they often employ different channels and collected data suffers from significant domain shifts caused by different devices, experiment setup, etc. On the other hand, the inter-subject heterogeneity is also substantial due to individual differences in EEG representations. In this work, we developed two networks to learn from both the shared and the complete channels across datasets, handling inter-subject and inter-dataset heterogeneity respectively. Based on both networks, we further developed an online knowledge co-distillation framework to collaboratively learn from both networks, achieving coherent performance boosts. Experimental results have shown that our proposed method can effectively aggregate knowledge from multiple datasets, demonstrating better generalization in the context of cross-subject validation.**

I. Introduction

Understanding human movement intention plays a critical role in human-robot interaction. Especially, for rehabilitation and assistive robotics, successfully recognizing movement intention is a prerequisite for assistive tool control and therapeutic motor training [1], [2], [3], [4]. For social robotics, anticipating upcoming movements can help improve the safety in human-level engagement over the course of human-robot collaboration [5], [6]. Thus far, it has received increasing attention to develop intelligent brain-computer interface (BCI) systems for intention recognition, and particularly, electroencephalogram (EEG) has been a popular measurement method due to its non-invasiveness and convenience of data acquisition.

Advances in deep learning have enabled automatic feature extraction and intention prediction from raw EEG signals, with considerable research efforts dedicated to effective computational architecture designs, such as EEGNet and DeepConvNet. However, raw EEG representations are quite complex, and are dominated by individual specific characteristics. Current research is prone to overfitting and being biased towards training subjects, which results in suboptimal cross-subject generalization [7], [8].

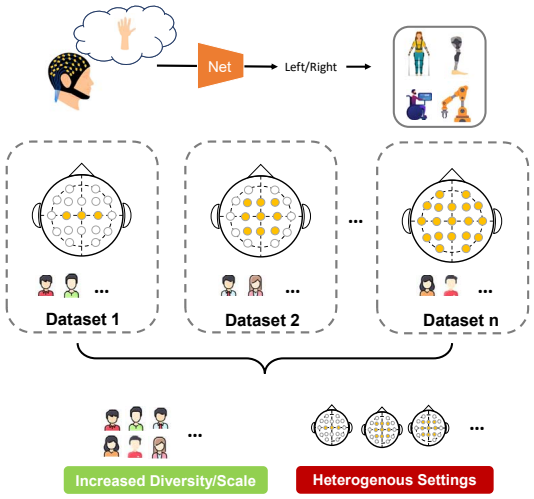


Fig. 1. **Illustration of benefits and challenges of leveraging multiple datasets for neural network training.** a) Accumulating multiple relatively-small motor-imagery EEG datasets can increase the diversity and scale of the datasets b) However, there exist heterogeneous settings across different datasets, including channel number, device, experiment paradigms, etc., limiting the applicability of utilizing multiple datasets.

On the other hand, research efforts have also been devoted to curating motor imagery datasets for algorithm development and benchmarking [9]. However, the diversity and scale of every single dataset are usually, if not always, relatively small, as caused by several factors such as tedious calibration and ethical issues. The limited data diversity makes the cross-subject generalization issue even more severe [7].

The increased diversity enabled by the increased number of available subjects is a potential countermeasure, as shown in Figure 1. One intuitive strategy is to directly aggregate multiple existing EEG datasets for model training. However, there are several issues hindering such practice. First of all, the number of electrode channels across datasets is varied, and such heterogenous settings lead to the failure of fixed implementation of existing computational architectures [10], [11], since the dimensionality of the input data is different. Although selecting common electrodes of different datasets could be a practical solution, this might drop out useful information from those dataset-unique channels [12]. Hence, corresponding solutions are necessary to more effectively deal with the heterogeneity of input dimensionality, which is rarely explored in the existing literature.

Furthermore, though different datasets could share identical imagination classes (such as left/right hand movement), their experiments may be conducted under paradigms that

Xiao Gu and Benny Lo are with the Hamlyn Centre, Imperial College London, London SW7 2AZ, United Kingdom. (email:xiao.gu17@imperial.ac.uk; benny.lo@imperial.ac.uk)

Jinpei Han is with the Brain & Behaviour Lab, Imperial College London, London SW7 2AZ, United Kingdom. (email:j.han20@imperial.ac.uk)

Guang-Zhong Yang is with the Institute of Medical Robotics and School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. (email: gzyang@sjtu.edu.cn)

are not exactly the same, with different instructional cue types, session-trial settings, etc. For instance, in [13], the subject was asked to follow the direction of an arrow to perform imagination, whereas in [14] the subject was asked to follow the instruction text displayed on the screen. This might result in domain shifts across datasets. Thus, the direct combination, if not addressed properly, might also lead to negative knowledge transfer from auxiliary datasets.

In our work, to deal with the heterogeneity existing at the individual level, as well as the dataset level, we propose a framework consisting of two networks to handle both heterogeneity issues respectively. A fixed network with data from common channels of multiple datasets as input is applied to handle inter-subject heterogeneity. On the other hand, a dynamic network that can adaptively take data of varied channel numbers is applied to handle inter-dataset heterogeneity. With these two nets, we further develop an online knowledge co-distillation framework to transfer knowledge from each other. Our contributions are listed below,

Training with Heterogeneous Datasets. We proposed a dynamic neural network architecture to handle channel heterogeneity in EEG datasets. To the best of our knowledge, this is the first work that focuses on channel heterogeneity across EEG datasets, and our proposed framework can effectively aggregate samples from heterogeneous datasets, without simply selecting the common channels across them.

Knowledge Co-Distillation. We proposed an online knowledge distillation framework to simultaneously transfer knowledge between the networks trained with the shared channels and with the complete channels. It can achieve coherent performance boosts for both networks.

Enhanced Cross-Subject Generalization. Our proposed framework is able to implicitly enhance the cross-subject generalization performance by aggregating data from multiple datasets, overcoming the small-size issue of most motor imagery benchmarks.

II. Related Work

A. Deep Learning for BCIs

The advances in deep learning have brought a paradigm shift to the way brain signals are interpreted. Till now, different deep learning architectures have been applied in various BCI applications. The basic architectures like Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) have demonstrated their capability of handling original EEG time series and its transformations [10], [15], [16], [17]. Attention mechanisms [18], [19] have also been introduced to deal with the temporal and spatial dynamics of the data, to empower the representations of discriminative features. Albeit these advances, in conventional BCI classification settings, computational models are trained and evaluated on the same subject's data. The performance on the data collected from new subjects thus cannot be evaluated. Furthermore, large amounts of works [20], [21], [7] have proved that the EEG data are prone to variations caused by individual differences. Such variations would result in low generalization performance if the training is under a vanilla end-to-end supervised manner with limited data diversity/scale [7].

B. Cross-Subject Transfer Learning in BCIs

Instead of applying vanilla end-to-end supervised training in BCIs, increasing attention has been paid to the challenging inter-subject variability issue. Transfer learning, especially domain adaptation, dominates this field [21], [8], [22], [23]. It aims to reduce calibration efforts on the new subjects (target domain) by leveraging the knowledge derived from existing subjects (source domains). Wu *et al.* [24] summarizes the general pipeline of transfer learning in motor imagery based BCI tasks and highlights the importance of integrating data alignment and sophisticated transfer learning approaches. Recently, Wei *et al.* [22] proposed a new task of performing transfer learning in multiple BCI datasets. However, the solutions introduced in [22] simply selected common channels across datasets; the trade-off between keeping more common electrode channels or including more subjects from more datasets limited the effectiveness.

C. Cross-Subject Generalization in BCIs

Different from transfer learning, which assumes the (partial) availability of target subjects for model adaptation, the task of domain generalization is expected to generalize on totally unseen subjects without any calibration. Domain generalization has received rapidly increasing attention in the computer vision field [25], [26], [27]. However, it has been much less explored in motor imagery based BCIs [28], especially compared to transfer learning. On the other hand, the generalization issue could be implicitly mitigated by more training subjects [29], which however often cannot be realized within each individual existing motor imagery benchmark [30], [31]. Although aggregating multiple datasets could promote the training subject diversity, it poses another novel challenge, namely the inter-dataset channel heterogeneity. Existing solutions such as picking up data from common channels [12] would drop out lots of useful information.

D. Knowledge Distillation in EEG

Knowledge distillation aims to transfer unique knowledge learned by one model to another model, which can be categorized as response-based, feature-based or relation-based distillation [32]. They distillate model output logits, intermediate feature representations, and relationships between different layers or data samples, respectively. In the literature, there have been several works on applying knowledge distillation to EEG data to handle different issues in varied applications. For instance, Wu *et al.* [33] applied knowledge distillation to bridge the gap between patient-specific and patient-independent models for EEG-based seizure detection. Zhang *et al.* [34] performed visual-to-EEG knowledge distillation for continuous emotion recognition. In contrast, we focused on the heterogeneity issue resulting from aggregating heterogeneous motor imagery datasets. We leveraged knowledge distillation to perform co-distillation between the models that are targeted at inter-subject heterogeneity and inter-dataset heterogeneity, respectively.

TABLE I
EXPERIMENTAL SETTINGS OF DIFFERENT TRAINING/TEST SPLITS.

Learning Paradigm	Dataset	Subject	Channel Number
Conventional validation	$d^{tr} = d^{te}$	$\{s\}^{tr} = \{s\}^{te}$	Same
Uni-dataset inter-subject adaptation	$d^{tr} = d^{te}$	$\{s\}^{tr} \supset \{s\}^{te}$	Same
Uni-dataset inter-subject generalization	$d^{tr} = d^{te}$	$\{s\}^{tr} \cap \{s\}^{te} = \varnothing$	Same
Multi-dataset inter-subject generalization	$\{d\}^{tr} \supset \{d\}^{te}$	$\{s\}^{tr} \cap \{s\}^{te} = \varnothing$	Heterogeneous

III. METHODS

A. Problem Formulation

Considering the left-right motor imagery EEG classification task, we denote the input as $\boldsymbol{x} \in \mathbb{R}^{C_d \times T}$, the dataset it belongs to as d, the subject as s_i, and the output as $y \in [0, 1]$, where C_d represents the channel number of existing domain d, class 0,1 represents left/right hand imagination, respectively. Thus in total there are $\{\{\boldsymbol{x}_i, y_i, s_i\}_{i=1}^{N_d}\}_{d=1}^{D}$.

Different from previous experimental settings, we would like to highlight the novelty of our learning paradigm.

1. Conventional Validation. Classic machine learning models are validated under intra-subject settings. With the training and testing data drawn from the same dataset and the same subject(s), the stages of model training and testing are deployed under the same data distribution. This paradigm has been utilized by several existing works [10], [35].

2. Intra-Dataset Inter-Subject Adaptation. Existing works on transfer learning mostly focus on intra-dataset inter-subject model adaptation. With both the data from source subjects and (un)labelled target subjects, fast and accurate model adaptation to the target subjects is expected [21], [24].

3. Intra-Dataset Inter-Subject Generalization. Another line of work is focused on generalizing to totally unseen subjects, yet is under-explored for motor-imagery EEG classification [20]. In other words, this protocol does not use any data from the target subjects during training. On the other hand, existing generalization methods, assume the data dimensionality to be the same, which can only apply to the data collected with the same electrode settings.

4. Multi-Dataset Inter-Subject Generalization. Beyond the above, our work involves collecting the data from multiple small-scale datasets and then training under the aggregated data, which forms another experiment pipeline.

We summarize the conceptual differences between the above evaluation protocols in Table I. In this work, we focus on the last experiment setting and provide an effective solution to deal with cross-dataset channel heterogeneity and cross-subject data distribution heterogeneity simultaneously.

B. Methodology Overview

Our framework is built upon the architecture of the popular EEGNet [10]. The original EEGNet consists of temporal convolution, spatial convolution, as well as depth-wise&pointwise separable convolution to extract temporal feature, spatial feature, and spatial-temporal feature step by step. Our framework consists of two branches, a fixed architecture f_c and a dynamic architecture f_{dyn}, as illustrated in Figure 1. The fixed network f_c takes the data from the common channels (i.e. C3, C4, Cz) as input, with a fixed input dimensionality. An implicit domain generalization solution is applied to process the data distribution shifts across large number of subjects. The dynamic network f_{dyn} is built on adaptive components that can process input data of varied dimensionalities. Since these two networks master unique properties, online knowledge distillation is conducted between f_c and f_{dyn} to collaboratively transfer knowledge from each other in an online manner. Hence, coherent performance boost can be achieved for both.

C. Fixed network with homogeneous common channels - f_c

The network f_c takes the data from shared channels across all the datasets, with the dimension as $\mathbb{R}^{C_0 \times T}$.

1) Batch-Instance Normalization (BIN): In f_c, with the homogeneous data dimensionality, we would like to pay more attention to handling heterogeneity across subjects. This is inspired by recent works of domain generalization in the general computer vision field [25]. Existing domain generalization works either implicitly or explicitly align the distribution shifts across domains. Explicit methodologies mostly seek help from domain-specific architecture or modules during training, which significantly increases the computational cost when the domain number (in our case, namely subject number) is large [26], [36]. On the other hand, implicit based methods tend to have much less computational cost since they save the burden of aligning the distributions across domains [37], [29].

We adopted a simple yet effective batch-instance normalization strategy [29] to incorporate instance normalization into the batch normalization layers. Given the intermediate layer output in a minibatch as $\boldsymbol{X} \in \mathbb{R}^{b \times dim_f \times dim_C \times dim_T}$, where b refers to the minibatch size, dim_f refers to the intermediate feature number, and dim_C,dim_T indicates the channel and time dimension. The batch-instance normalization is formulated as below,

$$\begin{aligned} \hat{\boldsymbol{X}}^{(B)} &= \frac{\boldsymbol{X} - \boldsymbol{\mu}^{(B)}}{\sqrt{\boldsymbol{\sigma}^{(B)2} + \epsilon}}, \\ \hat{\boldsymbol{X}}^{(I)} &= \frac{\boldsymbol{X} - \boldsymbol{\mu}^{(I)}}{\sqrt{\boldsymbol{\sigma}_c^{(I)2} + \epsilon}}, \end{aligned} \tag{1}$$

$$\hat{\boldsymbol{X}} = \gamma(\rho \cdot \hat{\boldsymbol{X}}^{(B)} + (1 - \rho) \cdot \hat{\boldsymbol{X}}^{(I)}) + \beta, \tag{2}$$

where ρ,γ,β are learnable weights, $\boldsymbol{\mu}^{(B)}$ and $\boldsymbol{\sigma}^{(B)}$ refer to feature statistics tracked from minibatches (batch-norm), $\boldsymbol{\mu}^{(I)}$ and $\boldsymbol{\sigma}^{(I)}$ refer to features statistics calculated from each individual sample (instance-norm), and ϵ is the small added value for numerical stability.

For f_c, cross-entropy loss $\mathcal{L}_{ce}^{c} = \frac{1}{N}\sum_{i=1}^{N} CE(f_c(\boldsymbol{x}_i), y_i)$ is utilized for optimization.

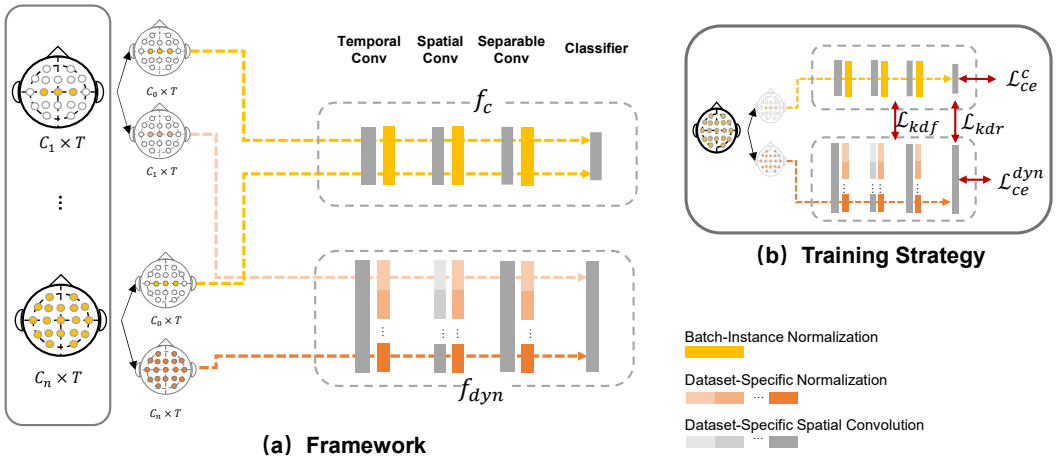


Fig. 2. **Illustration of our proposed frameworks for training with heterogeneous motor imagery EEG datasets.** (a) The input sample $\boldsymbol{x}_i \in \mathbb{R}^{C_n \times T}$ is decomposed into both the shared channel $\mathbb{R}^{C_0 \times T}$ and the original complete channel $\mathbb{R}^{C_n \times T}$. The data of selected channels is fed into a fixed network f_c with batch-instance normalization to handle inter-subject heterogeneity implicitly. Simultaneously, the original $\boldsymbol{x}_i$ is put to a dynamic network f_{dyn} with dataset-specific spatial convolution and normalization to deal with inter-dataset heterogeneity. (b) With these two networks, we performed online knowledge co-distillation to transfer knowledge from each other, with both feature-level $\mathcal{L}_{kdf}$ and response-level $\mathcal{L}_{kdr}$ distillation. Together with classification loss $\mathcal{L}_{ce}^{c}$ and $\mathcal{L}_{ce}^{dyn}$, the whole network is optimized end-to-end.

D. Dynamic network with heterogeneous channels - f_{dyn}

In order to adaptively process varying channel numbers, i.e., C_d of the input data, the following adaptations were made to the original EEGNet design, by introducing C_d-related dataset-specific modules.

1) Dataset-specific spatial convolution: In the original EEGNet [10], the kernel size of Spatial Conv is (C_d,1), which aims to aggregate features across channels and merge them into one single channel. To deal with the changes of C_d, we use dataset-specific spatial convolution in f_{dyn}. As shown in Figure 2, there are multiple dataset-specific spatial convolution components to process the data belonging to their corresponding datasets, separately.

2) Dataset-specific batch normalization (DSN): Meanwhile, since there exists heterogeneity of different datasets as caused by devices, experiment paradigms, etc., we employ dataset-specific batch normalization throughout the whole f_{dyn} to deal with such inter-dataset heterogeneity, by calculating dataset-specific statistics (i.e., $\boldsymbol{\mu}^{(B)}$ $\boldsymbol{\sigma}^{(B)}$ as in Equation 1) to perform normalization ($\hat{\boldsymbol{X}} = \gamma \cdot \hat{\boldsymbol{X}}^{(B)} + \beta$).

For the optimization of f_{dyn}, the cross-entropy loss $\mathcal{L}_{ce}^{dyn} = \frac{1}{N}\sum_{i=1}^{N} CE(f_{dyn}(\boldsymbol{x}_i), y_i)$ is applied.

E. Knowledge Distillation

As discussed above, the two networks are expected to master different properties, with f_c focused on inter-subject data shifts, and f_{dyn} focused on handling inter-dataset channel heterogeneity. To achieve coherent performance boost by transferring the knowledge from each other, we apply an online knowledge distillation framework to perform co-distillation of both networks, in terms of both the response and feature level.

1) Response-Level Distillation:

$$\mathcal{L}_{kdr} = \frac{1}{N}\sum_{i=1}^{N} KL(softmax(\boldsymbol{z}_i^c/T), softmax(\boldsymbol{z}_i^{dyn}/T)) + KL(softmax(\boldsymbol{z}_i^{dyn}/T), softmax(\boldsymbol{z}_i^c/T)), \quad (3)$$

where $\boldsymbol{z}_i^c$ and $\boldsymbol{z}_i^{dyn}$ represents the logit outputs of f_c and f_{dyn}, respectively, T is the temperature to soften the output logits, and KL represents KL divergence loss.

2) Feature-Level Distillation:

$$\mathcal{L}_{kdf} = \frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L} MMD(f_c^l(\boldsymbol{x}_i), f_{dyn}^l(\boldsymbol{x}_i)), \quad (4)$$

where f_c^l, f_{dyn}^l represent the intermediate feature from the l_{th} layer of f_c and f_{dyn}, MMD represents the Maximum Mean Discrepancy measurement between two feature maps [32]. In practice, we select the feature maps after spatial convolution and separable convolution blocks, since after these blocks, the outputs from both networks have the same feature dimensionality.

The whole network is trained end-to-end with weighted combination of the above four, $\mathcal{L}_{CE}^{c}$,$\mathcal{L}_{CE}^{dyn}$,$\mathcal{L}_{klf}$,$\mathcal{L}_{klr}$. The final prediction of our framework is an ensemble of both networks by $\frac{1}{2}(f_c(\boldsymbol{x}_i) + f_{dyn}(\boldsymbol{x}_i))$.

IV. Results and Discussion

A. Experimental Settings

1) Dataset Descriptions: In line with our motivations, we picked up four typical established left-right-hand motor imagery EEG datasets, with limited subject sizes and varied channel numbers. The details are given below.

- **BNCI2014001**. This dataset consists of 9 subjects with 22 channel EEG recordings with cue-based data collection paradigm [30]. Each trial was started with a short acoustic warning tone and an arrow pointing to the

TABLE II
Description of Datasets Used for Experiments.

Dataset	#Channel	#Subject	#Trial
BNCI2014001 [30]	22	9	2592
BNCI2014004 [31]	3	10	6519
Weibo2014 [14]	60	10	1580
Zhou2016 [13]	14	4	1199

TABLE III

RESULT OF LEAVE-ONE-SUBJECT-OUT VALIDATION. SUBJECT-MEAN ACC AND F1 SCORES OF EACH DATASET ARE REPORTED.

Method	Channel	Training Set	BNCI2014001		BNCI2014004		Weibo2014		Zhou2016	
			Acc	F1	Acc	F1	Acc	F1	Acc	F1
ShallowCNN [11]	Common	Uni	0.64±0.12	0.62±0.22	0.70±0.10	0.71±0.09	0.62±0.13	0.62±0.12	0.73±0.06	0.74±0.08
DeepCNN [11]			0.66±0.10	0.67±0.11	0.66±0.10	0.65±0.11	0.64±0.12	0.63±0.12	0.72±0.07	0.69±0.12
EEGNet [10]			0.65±0.08	0.62±0.14	0.71±0.09	0.71±0.09	0.57±0.08	0.61±0.06	0.70±0.04	0.67±0.10
f_c only			0.66±0.08	0.62±0.10	0.72±0.10	0.72±0.08	0.58±0.09	0.62±0.07	0.71±0.05	0.67±0.10
ShallowCNN [11]	All	Uni	0.73±0.11	0.71±0.16	0.71±0.09	0.71±0.09	0.66±0.14	**0.69**±0.10	0.74±0.08	0.73±0.13
DeepCNN [11]			0.70±0.13	0.72±0.15	0.67±0.09	0.67±0.08	0.68±0.10	0.67±0.11	0.76±0.04	0.74±0.06
EEGNet [10]			0.77±0.06	0.76±0.08	0.72±0.08	0.71±0.08	0.65±0.10	0.62±0.13	0.79±0.07	0.79±0.08
Ours-f_c			0.66±0.06	0.63±0.09	0.72±0.07	0.72±0.07	0.60±0.09	0.63±0.07	0.71±0.03	0.68±0.07
Ours-f_{dyn}			0.80±0.05	0.80±0.06	0.73±0.08	0.73±0.08	0.68±0.09	0.68±0.12	0.79±0.05	0.79±0.09
Ours-Ensemble			0.80±0.05	0.79±0.06	0.73±0.08	0.74±0.08	0.69±0.08	**0.69**±0.09	0.80±0.05	0.80±0.09
ShallowCNN [11]	Common	Multi	0.64±0.11	0.60±0.11	0.64±0.07	0.60±0.10	0.58±0.11	0.48±0.16	0.73±0.06	0.76±0.04
DeepCNN [11]			0.63±0.09	0.62±0.15	0.60±0.08	0.52±0.15	0.61±0.10	0.54±0.15	0.73±0.04	0.74±0.06
EEGNet [10]			0.61±0.09	0.62±0.13	0.63±0.07	0.61±0.10	0.60±0.11	0.55±0.15	0.73±0.06	0.76±0.04
f_c only			0.66±0.09	0.63±0.13	0.65±0.10	0.64±0.11	0.61±0.10	0.58±0.10	0.73±0.02	0.77±0.04
f_{dyn} only	All	Multi	0.79±0.07	0.78±0.08	0.74±0.10	0.72±0.10	0.66±0.10	0.67±0.11	0.79±0.08	0.78±0.05
Ours-f_c			0.67±0.07	0.67±0.09	0.73±0.09	0.72±0.10	0.61±0.08	0.63±0.06	0.75±0.05	0.75±0.01
Ours-f_{dyn}			0.80±0.07	**0.81**±0.07	**0.76**±0.10	0.74±0.08	0.69±0.10	**0.69**±0.10	**0.81**±0.07	0.80±0.09
Ours-Ensemble			**0.81**±0.06	**0.81**±0.06	**0.76**±0.11	**0.75**±0.08	**0.70**±0.08	**0.69**±0.11	**0.81**±0.06	**0.81**±0.03

The best results are in **bold** and the second best results are underlined.

left/right appeared to indicate corresponding left/right hand movement imagination.

- **BNCI2014004**. It contains 9 subjects from 3 channels (C3,C4,Cz), under a cue-based paradigm [31]. Each trial was started with a fixation cross and an acoustic warning tone. The visual cue was shown subsequently to present the motor imagery type.
- **Weibo2014**. It was collected with 10 subjects and 60 electrodes [14]. A white circle appearing on the screen indicated the start of each trial, followed by a red circle as a preparation cue, and text showing the imagery type.
- **Zhou2016**. This dataset contains 4 subjects with 14 channels [13]. Each trail started with a short beep, followed by a red arrow that indicated the imagination task by its arrow direction.

The details are given in Table II for intuitive comparison. We segmented out the first 2s of each trial across all the datasets to construct the whole dataset, with a bandpass filter of 3-40 Hz and z-normalization as preprocessing. All the data is loaded by MOABB [1]. The common channels across all datasets are {C3, C4, Cz}.

2) Evaluation Paradigms: To evaluate the generalization capability of different methods, we applied the **Leave-One-Subject-Out** (**LOSO**) cross-validation evaluation strategy. Under this protocol, we selected one subject as the testing subject and the remaining data results in the training split. We randomly split 20% from the training split as validation to select the best model, and to report the performance on the testing split. The Accuracy (Acc) and the F1 scores were utilized as performance metrics.

To further evaluate the effectiveness of our proposed method under different channel numbers and with/without the help of additional datasets, the comparison was conducted under different settings of the channels and datasets.

- **Common-Channel or All-Channel.** Two options were available in channel settings. "Common-Channel" refers to utilizing the shared channels across all the datasets, namely {C3, C4, Cz}, whilst the other option "All-Channel" employed all channels in each dataset.
- **Uni-Dataset or Multi-Dataset.** Similarly, two options were provided in terms of the training data compositions. In the "Uni-Dataset" settings, all the remaining subjects from the same dataset as held-out testing subjects contributed to the training/validation data. In the "Multi-Dataset" settings, we also added all the data from the other datasets to form the final training split.

For comparison, we also implemented three motor imagery classification methodologies EEGNet [10], DeepCNN [11], ShallowCNN [11]. They shared the same experimental settings as our method for a fair comparison.

3) Implementation Details: We implemented all the networks by Pytorch with Titan Xp. For training, we applied AdamW as the optimizer, with learning rate as 5e-2, betas as (0.9,0.999), and weight decay as 1e-4. The batch size of each dataset was set as 32.

B. Quantitative Results

We present the quantitative results of LOSO cross-subject validation in Table III. Overall our proposed method (marked in) outperformed other compared methods. This was enabled by incorporating additional datasets, as well as leveraging our proposed architecture design and training strategy. Below we give detailed discussions of the results.

Effectiveness of Batch-Instance Normalization. As shown in the Table III, the adaptation to the original EEGNet with our proposed architecture (f_c only as marked in) can achieve better generalization performance, compared to the original EEGNet in most cases.

Effectiveness of Dynamic Architecture. Furthermore, in the training of All-Channel + Multi-Dataset, we also applied f_{dyn} only (marked in) to validate the efficacy of the dynamic architecture. Our proposed method can bring in better performance with additional datasets, compared to directly

[1] http://moabb.neurotechx.com/

selecting common channels. It is demonstrated to be a better solution to handle inter-dataset channel heterogeneity.

Knowledge Distillation between Models Trained with Different Channels. We also implemented our framework within All Channel + Uni-Dataset settings. In this way, the dynamic architecture can be viewed as a fixed architecture with the original data of all channels as input. In this way, the knowledge co-distillation is more focused on the representations under different channel numbers, and/or different normalization strategies. Performance gains can be noted in Table III, as marked in ■, compared to other methods under the same settings.

All Channels vs Common Channels. Comparing the same algorithm that is trained using only common channels and using complete channels, it can be easily found that only selecting partial electrodes can significantly drop out useful information, thus deteriorating the generalization performance. This validates that when there are large differences in electrode settings between datasets, simply selecting only the common shared channels might not be an ideal solution.

Multi-Dataset vs Uni-Dataset. An interesting finding is that, directly collecting data from multiple datasets by picking up common channels mostly is not functional. It can be observed that in Common-Channel settings (Common-Channel, Multi-Dataset VS Uni-Dataset), in most cases, directly utilizing multi-dataset to train would not result in a performance boost. This might be caused by the heterogeneity across datasets caused by devices, and experiment setups. Adding other datasets into training would bias the model by such heterogeneity. Instead, our data-specific modules can handle this issue by calculating dataset-specific statistics to perform dataset-specific normalization.

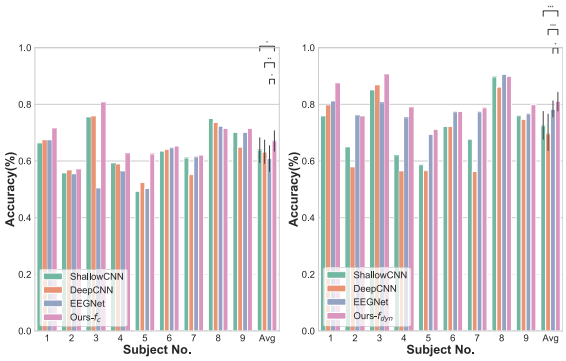
Performance across Each Subject. We also present the Acc results of each subject of BNCI2014001 in Figure 3, and performed paired t-test to show the significance of the performance gains enabled by our method.

C. Qualitative Results

The spatial convolution in EEGNet [10] functions as a spatial filter to perform a weighted average across all the channels. We visualized one representative kernel of the spatial convolutions from f_c, f_{dyn} in Figure 4. For more intuitive visualization, we performed extrapolation to the whole head to deal with channel-heterogeneity. It can be observed that they shared a similar pattern (larger weights on the right hemisphere area), whereas, for each dataset-specific filter, they demonstrate some unique characteristics.

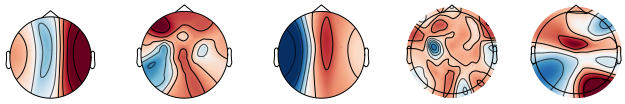
D. Ablation Studies

We further perform ablation studies to validate the effectiveness of each loss function of our proposed method, with the results shown in Table IV. Furthermore, we also tried to directly replace dataset-specific normalization with batch-instance normalization in f_{dyn} and test whether the modified f_{dyn} only is enough to handle both inter-subject and inter-dataset heterogeneity. However, it shows inferior performance compared to our method.



(a) C3,C4,Cz (b) All Channels

Fig. 3. **Acc results of each subject of LOSO validation on BCNI2014001.** On the left, ShallowCNN, DeepCNN, EEGNet were trained under Common-Channel + Multi-Dataset settings, whilst on the right, they were trained under All-Channel + Uni-Dataset settings. Our f_c and f_{dyn} is learned under All-Channel + Uni-Dataset settings. The rightmost column presents subject-average results, with paired t-test to perform significant test. (*: 0.01<p≤0.05, **: 0.001<p≤0.01, ***: 0.0001<p≤0.001)



(a) f_c (b) f_{dyn}-1 (c) f_{dyn}-2 (d) f_{dyn}-3 (e) f_{dyn}-4

Fig. 4. **Visualization of one representative spatial convolution filters from both f_c and four domains-specific modules of f_{dyn}.** They are visualized under the same range (red-larger; blue-smaller).

TABLE IV
RESULT OF ABLATION STUDY.

	Ablation				BNCI2014001	
					Acc	F1
	$\mathcal{L}_{ce}^{c}$	$\mathcal{L}_{ce}^{a}$	$\mathcal{L}_{kdr}$	$\mathcal{L}_{kdf}$		
a	✓	✓			0.79±0.07	0.77±0.08
b	✓	✓	✓		0.79±0.06	0.79±0.08
c	✓	✓	✓	✓	0.81±0.06	0.81±0.06
d	f_{dyn} only, with BIN				0.76±0.12	0.74±0.10

V. CONCLUSION

Generalizing human movement recognition on unforeseen novel subjects is important for real-world human-robot interaction. Existing motor imagery EEG training could suffer from relatively small size of subjects due to ethics and tedious data calibration, thus resulting in poor generalization. In this work, we seek the help from multiple small-scale open-source datasets to increase the diversity and scale of training data. However, such operation is not trivial, rather associated with several issues including the inter-subject data distribution heterogeneity as well as the inter-dataset channel heterogeneity. To tackle these issues, we built a fixed network with batch-instance normalization and a dynamic network with dataset-specific modules, separately. Based on these two networks, we further proposed a novel collaborative online knowledge distillation framework to achieve coherent performance boosts. Experimental results show that our proposed method achieved superior performance against other baseline motor imagery algorithms. In this regard, a novel cross-subject generalization solution was developed by learning from multiple heterogeneous EEG datasets.

REFERENCES

[1] G.-Z. Yang, R. Riener, and P. Dario, "To integrate and to empower: Robots for rehabilitation and assistance," *Science Robotics*, vol. 2, no. 6, p. eaan5593, 2017.

[2] P. Beckerle, G. Salvietti, R. Unal, D. Prattichizzo, S. Rossi, C. Castellini, S. Hirche, S. Endo, H. B. Amor, M. Ciocarlie *et al.*, "A human–robot interaction perspective on assistive and rehabilitation robotics," *Frontiers in neurorobotics*, vol. 11, p. 24, 2017.

[3] Y. Guo, X. Gu, and G.-Z. Yang, *Human–Robot Interaction for Rehabilitation Robotics*. Cham: Springer International Publishing, 2021, pp. 269–295. [Online]. Available: https://doi.org/10.1007/978-3-030-65896-0_23

[4] S. An, S. Kim, P. Chikontwe, and S. H. Park, "Few-shot relation learning with attention for eeg-based motor imagery classification," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 933–10 938.

[5] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," *Science robotics*, vol. 3, no. 21, p. eaat5954, 2018.

[6] T. Smith, Y. Chen, N. Hewitt, B. Hu, and Y. Gu, "Socially aware robot obstacle avoidance considering human intention and preferences," *International journal of social robotics*, pp. 1–18, 2021.

[7] X. Chen, C. Li, A. Liu, M. J. McKeown, R. Qian, and Z. J. Wang, "Toward open-world electroencephalogram decoding via deep learning: a comprehensive survey," *IEEE Signal Processing Magazine*, vol. 39, no. 2, pp. 117–134, 2022.

[8] D. Wu, Y. Xu, and B.-L. Lu, "Transfer learning for eeg-based brain-computer interfaces: A review of progress made since 2016," *IEEE Transactions on Cognitive and Developmental Systems*, 2020.

[9] V. Jayaram and A. Barachant, "Moabb: trustworthy algorithm benchmarking for bcis," *Journal of neural engineering*, vol. 15, no. 6, p. 066011, 2018.

[10] 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.

[11] 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.

[12] 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, 2020.

[13] B. Zhou, X. Wu, Z. Lv, L. Zhang, and X. Guo, "A fully automated trial selection method for optimization of motor imagery based brain-computer interface," *PloS one*, vol. 11, no. 9, p. e0162657, 2016.

[14] W. Yi, S. Qiu, K. Wang, H. Qi, L. Zhang, P. Zhou, F. He, and D. Ming, "Evaluation of eeg oscillatory patterns and cognitive process during simple and compound limb motor imagery," *PloS one*, vol. 9, no. 12, p. e114853, 2014.

[15] S. R. Tibor, S. J. Tobias, F. L. D. Josef, G. Martin, E. Katharina, T. Michael, H. Frank, B. Wolfram, and B. Tonio, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.23730

[16] 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 Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 5, pp. 1226–1238, 2020.

[17] P. Zhang, X. Wang, W. Zhang, and J. Chen, "Learning spatial–spectral–temporal eeg features with recurrent 3d convolutional neural networks for cross-task mental workload assessment," *IEEE Transactions on neural systems and rehabilitation engineering*, vol. 27, no. 1, pp. 31–42, 2018.

[18] Z. Jia, Y. Lin, J. Wang, R. Zhou, X. Ning, Y. He, and Y. Zhao, "Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI*, 2020, pp. 1324–1330.

[19] D. Zhang, K. Chen, D. Jian, and L. Yao, "Motor imagery classification via temporal attention cues of graph embedded eeg signals," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[20] J. Han, X. Gu, and B. Lo, "Semi-supervised contrastive learning for generalizable motor imagery eeg classification," in *2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 2021, pp. 1–4.

[21] K. Zhang, N. Robinson, S.-W. Lee, and C. Guan, "Adaptive transfer learning for eeg motor imagery classification with deep convolutional neural network," *Neural Networks*, vol. 136, pp. 1–10, 2021.

[22] X. Wei, A. A. Faisal, M. Grosse-Wentrup, A. Gramfort, S. Chevallier, V. Jayaram, C. Jeunet, S. Bakas, S. Ludwig, K. Barmpas, M. Bahri, Y. Panagakis, N. Laskaris, D. A. Adamos, S. Zafeiriou, W. C. Duong, S. M. Gordon, V. J. Lawhern, M. Śliwowski, V. Rouanne, and P. Tempczyk, "2021 beetl competition: Advancing transfer learning for subject independence & heterogenous eeg data sets," in *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, ser. Proceedings of Machine Learning Research, D. Kiela, M. Ciccone, and B. Caputo, Eds., vol. 176. PMLR, 06–14 Dec 2022, pp. 205–219.

[23] F. Fahimi, S. Dosen, K. K. Ang, N. Mrachacz-Kersting, and C. Guan, "Generative adversarial networks-based data augmentation for brain–computer interface," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 9, pp. 4039–4051, 2020.

[24] D. Wu, X. Jiang, and R. Peng, "Transfer learning for motor imagery based brain-computer interfaces: A tutorial," *Neural Networks*, 2022.

[25] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[26] Q. Dou, D. Coelho de Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[27] X. Gu, Y. Guo, Z. Li, J. Qiu, Q. Dou, Y. Liu, B. Lo, and G.-Z. Yang, "Tackling long-tailed category distribution under domain shifts," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII*. Springer, 2022, pp. 727–743.

[28] P. Autthasan, R. Chaisaen, T. Sudhawiyangkul, P. Rangpong, S. Kiatthaveephong, N. Dilokthanakul, G. Bhakdisongkhram, H. Phan, C. Guan, and T. Wilaiprasitporn, "Min2net: End-to-end multi-task learning for subject-independent motor imagery eeg classification," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 6, pp. 2105–2118, 2021.

[29] H. Nam and H.-E. Kim, "Batch-instance normalization for adaptively style-invariant neural networks," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[30] 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.

[31] R. Leeb, F. Lee, C. Keinrath, R. Scherer, H. Bischof, and G. Pfurtscheller, "Brain–computer communication: motivation, aim, and impact of exploring a virtual apartment," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 15, no. 4, pp. 473–482, 2007.

[32] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[33] D. Wu, J. Yang, and M. Sawan, "Bridging the gap between patient-specific and patient-independent seizure prediction via knowledge distillation," *Journal of Neural Engineering*, 2022.

[34] S. Zhang, C. Tang, and C. Guan, "Visual-to-eeg cross-modal knowledge distillation for continuous emotion recognition," *Pattern Recognition*, p. 108833, 2022.

[35] 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.

[36] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5400–5409.

[37] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," in *International Conference on Learning Representations*, 2020.