Transfer Learning With Optimal Transportation and Frequency Mixup for EEG-Based Motor Imagery Recognition

Peiyin Chen, He Wang, Xinlin Sun, Haoyu Li, Celso Grebogi, and Zhongke Gao, *Senior Member, IEEE*

***Abstract*— Electroencephalography-based Brain Computer Interfaces (BCIs) invariably have a degenerate performance due to the considerable individual variability. To address this problem, we develop a novel domain adaptation method with optimal transport and frequency mixup for cross-subject transfer learning in motor imagery BCIs. Specifically, the preprocessed EEG signals from source and target domain are mapped into latent space with an embedding module, where the representation distributions and label distributions across domains have a large discrepancy. We assume that there exists a nonlinear coupling matrix between both domains, which can be utilized to estimate the distance of joint distributions for different domains. Depending on the optimal transport, the Wasserstein distance between source and target domains is minimized, yielding the alignment of joint distributions. Moreover, a new mixup strategy is also introduced to generalize the model, where the inputs trials are mixed in frequency domain rather than in raw space. The extensive experiments on three evaluation benchmarks are conducted to validate the proposed framework. All the results demonstrate that our method achieves a superior performance than previous state-of-the-art domain adaptation approaches.**

***Index Terms*— Electroencephalogram (EEG), brain-computer interface (BCI), transfer learning, optimal transportation.**

I. Introduction

BRAIN-COMPUTER interface (BCI) based on electroencephalography (EEG) is capable of establishing an interactive pathway between human brains and electronic devices, which is promising for a wide range of applications, such as rehabilitation, smart house systems, entertainments, and so forth [1], [2], [3], [4], [5], [6]. In general, a complete and robust EEG-based BCI system requires a calibration phase and a testing phase [7]. In calibration phase, a decoding model is learned with some labelled EEG signals, aiming to predict each input with low risk. In testing phase, the learned model is applied to those unseen signals and outputs their predictions, which can be served as the control instructions for external devices [2], [3]. However, there are still remaining much challenges in the application of BCIs, which are majorly contributed in the characteristics of EEG signals. On one hand, EEG signals are non-stationary, non-Gaussian and have a low signal-to-noise ratio [8], leading to the difficulty of discriminative features extraction and signal analysis, despite the numerous available methods, like Common Spatial Pattern (CSP), Short-time Fourier Transform (STFT), Differential Entropy (DE) and complex network theory [9], [10], [11], [12], [13], [14], [15]. On the other hand, large individual difference of EEG signals makes it difficult to learn a robust model across subjects, which can bridge the data distributions shift between different subjects [16], [17]. The above dilemma calls for more powerful and effective approaches to be developed in further studies.

Transfer learning provides an effective solution to bridge the data shift in EEG-based BCIs. Over the past few years, numerous transfer learning approaches based on Unsupervised Domain Adaptation (UDA) have been investigated and applied to EEG classification [18], [19], [20], [21], [22], [23]. According to the standard setting of UDA, the domain with enough labeled samples refers to source domain, while the domain with only unlabeled samples is regarded as target domain [24]. A major issue of UDA is to adapt/align the distribution and reduce distribution discrepancy between different domains. To this effect, the earlier works aim at transforming the target domain onto source domain, like Euclidean space data Alignment (EA) [18], Transfer Component Analysis (TCA) [20], [25] and Joint distribution adaptation (JDA) [26], [27]. Recently, most studies majorly focus on the shared representations space construction on both domains, like Domain Adversarial Neural Network (DANN) [19], [28], Joint Adaptation Network (JAN) [29], Conditional Adversarial Domain Adaptation Neural Network (CDAN) [21], [30] and Dynamic

Manuscript received 9 June 2022; revised 24 September 2022; accepted 28 September 2022. Date of publication 4 October 2022; date of current version 20 October 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61873181, Grant 61922062, and Grant 61903270; in part by the Innovation of Science and Technology Forward 2030 Program "Brain Science and Brain-Inspired Intelligence Technology" under Grant 2021ZD0201600; and in part by the Natural Science Foundation of Tianjin, China, under Grant 21JCJQJC00130. *(Corresponding author: Zhongke Gao.)*

Peiyin Chen, He Wang, Xinlin Sun, Haoyu Li, and Zhongke Gao are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: pychen@tju.edu.cn; hewang@tju.edu.cn; xinlinsun@tju.edu.cn; lchaoyu@outlook.com; zhongkegao@tju.edu.cn).

Celso Grebogi is with the King's College, Institute for Complex Systems and Mathematical Biology, University of Aberdeen, AB24 3UE Aberdeen, U.K. (e-mail: grebogi@abdn.ac.uk).

Digital Object Identifier 10.1109/TNSRE.2022.3211881

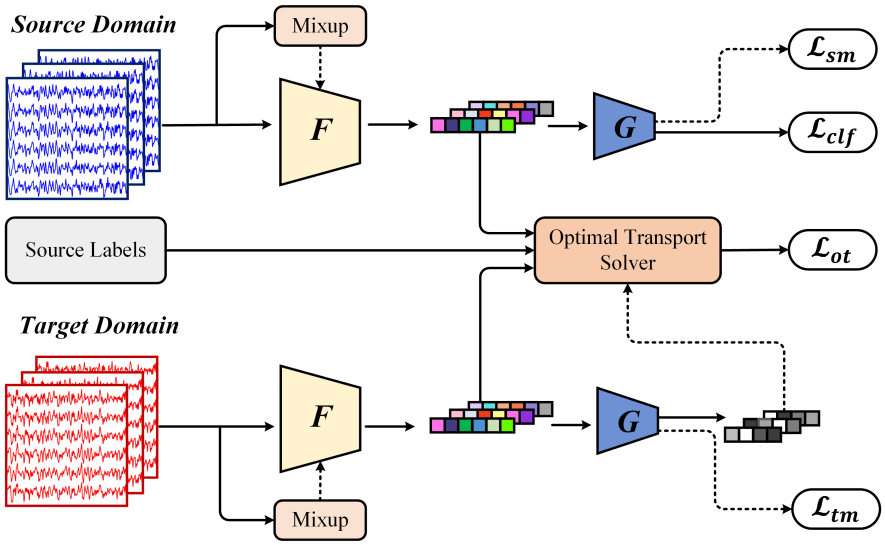


Fig. 1. The pipeline of the proposed domain adaptation framework. The samples from source and target domain are mapped into latent space with a shared embedding block *F*. Then both the representations and labels are employed to estimate the domain distance across domains, which helps guide the alignment of domain joint distributions. To further generalize the model, the signals from each domain are mixed in the frequency domain of temporal space, those new mixed samples are also fed to embedding block *F* and classifier *G* for prediction. Note that the predictions for target signals are utilized as their pseudo labels.

Joint Domain Adaptation Network (DJDAN) [22]. These works adapt the distribution across domains with either Maximum Mean Discrepancy (MMD) or domain discriminator, only a few concentrate on the domain adaptation with optimal transport [31], a more general framework with geometric interpretability. As a consequence, we develop the domain adaptation approach based on optimal transport for cross-subject transfer learning task in EEG-based BCIs.

Another prevalent challenge in EEG-based BCIs is the problem of data scarcity [8], [32], leading to model overfitting and performance degradation. A direct and simple solution is to utilize the data from different subjects to train a subject-specific model, which is found to be effective in cross-session classification [33]. Nevertheless, this strategy does not address the above problem in cross-subject classification task. Data augmentation is a popular strategy to generalize the model in computer vision [34], [35], [36], [37], [38], [39], [40], [41]. To date, there are many augmentation methods have been developed and applied to image data, including geometric transformations [34], [35], random erasing [36], [37], GAN-based augmentations [35], [38], Mixup-based augmentations [39], [40], [41] and so on. Reassuringly, most of augmentation approaches in vision have been transferred to BCI [42], [43], [44] and achieved considerable performance improvement. However, few studies concentrate on the application of Mixup-based augmentations in EEG signals. Therefore, we introduce a novel mixup strategy for EEG augmentation, which would be detailed in Section III.

In this work, we develop an offline transfer learning framework, namely Joint Distribution Adaptation with Optimal Transportation and frequency Mixup (JDAOT-Mix), to address the above limitations in cross-subject classification of BCI. The overview of the proposed framework is presented in Figure 1. Specifically, the joint distributions across domain is adapted by employing the optimal transportation theory, where the optimal transport plan between source and target domains is calculated. In addition, a novel Mixup in Frequency Domain (MFD) strategy is introduced to generalize the model. Compared to other mixup strategies, new mixup samples are created in the frequency domain of input space, instead of raw space. Specially, a fixed mixup ratio is adopted in our mixup scheme, which is different from the random mixup ratio in previous methods. During training, the proposed mixup strategy is applied to both domains, utilizing the ground truths and pseudo labels from source and target domains.

Our major contributions can be highlighted as follows:

- A novel transfer learning framework called joint distribution adaptation with optimal transportation and frequency mixup (JDAOT-Mix) is developed for the cross-subject classification task in EEG-based motor imagery.
- A Mixup in Frequency Domain (MFD) method is introduced as a new data augmentation for EEG-based BCI, where the new samples is created in the frequency domain of input space by taking a pair of samples and their corresponding labels.

The remainder of this paper is organized as follows. Section II mainly review the previous works on domain adaptation, data augmentation and their applications for BCI. Section III details the components of the proposed framework. Section IV describes the experimental setting and section V presents their results on two evaluation datasets. Next, section VI discusses the related ablation studies. Finally, Section VII concludes this work.

II. Related Works

A. Unsupervised Domain Adaptation

Denote $\mathcal{D}^s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^m$ the set of EEG signals from source domain, where $\mathbf{x}_i^s \in \mathbb{R}^{C\times T}$ represents a EEG trial with

C electrodes and the length of T, y_i^s is the corresponding label, and m is the number of all EEG trials. Similarly, $\mathcal{D}^t = \{(\mathbf{x}_i^t)\}_{i=1}^n$ denotes the unlabeled target domain with n EEG trials. The ultimate goal of unsupervised domain adaptation is to generalize the model from source domain $\mathcal{D}^s$ to target domain $\mathcal{D}^t$, with the adequate labeled source samples and unlabeled target samples. In BCIs, knowledge transfer across different subjects is the promising research topic, yet with a great challenge. Recently, many interesting domain adaptation approaches have been proposed to address this issue. Earlier searchers concentrate on subspace-based methods, where the low-dimensional subspace preserves the data properties and intrinsic distributions across domains. For example, transfer component analysis (TCA) method [25] aims to learn the transfer components across domains in a reproducing kernel Hilbert space with maximum mean discrepancy. Joint distribution adaptation (JDA) method [26] tries to jointly align both the marginal distribution and conditional distribution in a principled dimensionality reduction procedure. As the efficacy of GAN networks have been proved in visual domain, most recent works gradually exploit the domain-shared space learning with adversarial-based frameworks, such as DANN [19], [23], [28], CDAN [21], [30] and DJDAN [22]. For domain-invariant representations learning, these approaches always play a minimax game between the shared feature extractor and domain discriminator, where the feature extractor is utilized to map source and target samples into a latent space, the domain discriminator tries to distinguish the representations from source domain to target domain. Most recently, a few methods based on optimal transportation are proposed to domain adaptation. In [45], a novel framework based on regularized discrete optimal transport is applied to cross-subject transfer learning for the P300-Speller paradigm, where a transport plan is estimated to map the target features onto source domain. In [7], a backward optimal transport for domain adaptation (BOTDA) is proposed for cross-session MI-based BCI, where the target samples are transformed by a transport mapping to modify the trained classifier. Despite these recent developments for domain adaptation and their successful applications in EEG-based BCIs, much challenges are still remained in cross-subject transfer learning, especially in the case of multiple categories classification. Therefore, we develop a novel framework, namely joint distribution adaptation with optimal transportation and frequency mixup (JDAOT-Mix), inspired by the previous works in computer vision [31], [46]. In the proposed JDAOT-Mix, the joint distributions across domains are aligned by optimal transport estimation, and a frequency mixup strategy is introduced to train a more robust classifier. Compared with BOTDA, JDAOT-Mix is a deep domain adaptation method using optimal transport loss to reduce domain shift, without transform target features to source domain by transport matrix.

B. Mixup-Based Methods for Data Augmentation

In the field of computer vision, the Mixup-based augmentation is a popular and effective strategy for generalizing models [39], [40], [41]. The original mixup method [39] takes a pair of samples and their labels to create a new sample and corresponding label, via the convex combination of training samples. Another mixup strategy, called CutMix [40], using a patch from input image to replace the removed regions of another image, their labels are mixed proportionally to their number of pixels. But the above methods may disregard the local saliency information embedded in the underlying data structure. To address this issue, a Puzzle Mix method, leveraging the saliency information and the underlying regional statistics of input samples, is proposed for vision tasks. Recently, many works on BCI have investigated the mixup augmentation and applied it to training procedure [47], [48], [49]. These works apply mixup to training samples in raw input space; however, few studies concentrate on creating new samples in frequency domain using mixup. Specially, we introduce a Mixup in Frequency Domain (MFD) augmentation to generalize the models. In our MFD, new samples are the linear interpolation of the FFT series from two randomly selected samples, with a fixed mixup ratio instead of random ratio. From the following empirical experiments, we find that the introduced MFD strategy is more suitable for EEG-based BCI systems.

III. Methods

In this section, we will present our proposition for cross-subject transfer learning in details.

A. Representations Learning With Neural Network

To extract representations from input EEG signals, we design a lightweight architecture as baseline network, which comprises the Embedding Block F and the Classifier G (as shown in Figure 2). The parameters of network is presented in Table I.

Given an input signal $\mathbf{x} \in \mathbb{R}^{C\times T}$, the Embedding Block is responsible for mapping it into the latent space. Firstly, the input signal is fed into two cascaded convolutional blocks (denoted Temporal Conv and Spatial Conv in Figure 2) followed by a BatchNorm layer for temporal-spatial dependencies extraction. Then the output of Spatial Conv is transformed with a EPD block (combining ELU, Average Pooling and Dropout layers), which employs an Average Pooling layer for computation efficiency and a Dropout layer for alleviating the overfitting problem. Next, one Separable Conv [50] and another EPD block are sequentially applied to aggregate high-level representations. Finally, the representations $f \in \mathbb{R}^d$ output from the Embedding Block is taken by the classifier (a Dense layer with Softmax activation function) to predict final classification score $\hat{y} \in \mathbb{R}^c$, where d and c are the dimension of representation and the number of the categories for classification, respectively. To avoid the performance degradation in source domain, the objective of representations learning is defined as follows:

$$\mathcal{L}_{clf} = \frac{1}{m}\sum_{i}^{m} L\left(y_i^s, G\left(F\left(\mathbf{x}_i^s\right)\right)\right), \tag{1}$$

where $L(\cdot,\cdot)$ is the cross-entropy function.

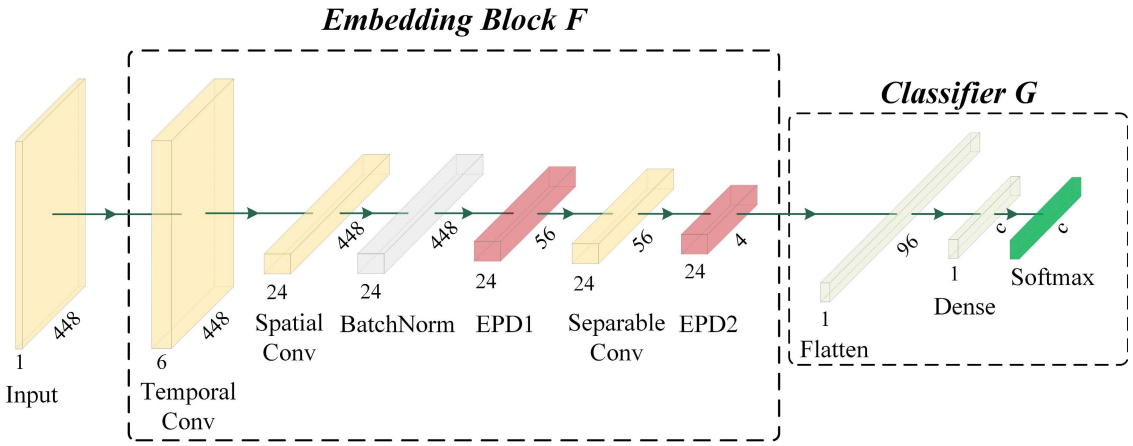


Fig. 2. The architecture of baseline model for given an input of 22×448, which consists of an Embedding Block *F* and a Classifier *G*. The embedding block *F* is utilized to map input signals to high-level feature space, and outputs discriminative representations to classifier *G* for predictions.

TABLE I
MODEL PARAMETERS OF BASELINE

Module	Layer	Kernel	Option
Embedding Block	Input	-	-
	Temporal Conv	$6 \times 1 \times 32$	SAME
	Spatial Conv	$24 \times C \times 1$	VALID
	Batch Norm	-	-
	ELU	-	-
	Average Pooling	1×8	-
	Dropout	-	0.25
	Separable Conv	$24 \times 1 \times 16$	SAME
	ELU	-	-
	Average Pooling	1×12	-
	Dropout	-	0.5
	Flatten	-	-
Classifier	Dense	$d \times c$	-
	Softmax	-	-

B. Representations Adaptation With Optimal Transportation

Denote μ_s and μ_t as the empirical probability measures for source and target domains, respectively. Optimal transportation provides us a geometrical solution to measure the distance between two different domains, by solving the discrete version of Monge-Kantorovich problem [51] as follows:

$$W(\mu_s, \mu_t) = \min_{\gamma \in \Pi(\mu_s, \mu_t)} \mathcal{L}_{ot} , \tag{2}$$

$$\mathcal{L}_{ot} = \sum_{i}^{m} \sum_{j}^{n} \gamma_{ij} c\left(\mathbf{x}_i^s, \mathbf{x}_j^t\right) , \tag{3}$$

where $\gamma \in \mathbb{R}^{m \times n}$ is a probabilistic coupling between source and target domains, where each element γ_{ij} represents the coupling coefficient between source sample $\mathbf{x}_i^s$ and target sample $\mathbf{x}_j^t$. $\Pi(\mu_s, \mu_t)$ denotes a collection of joint probability distributions with marginals μ_s and μ_t, and $c\left(\mathbf{x}_i^s, \mathbf{x}_j^t\right)$ is the cost function for measuring the dissimilarity between samples $\mathbf{x}_i^s$ and $\mathbf{x}_j^t$.

In this work, the underlying idea of our proposition is to jointly adapt the marginal and conditional distributions across domains in latent space. To this end, $c\left(\mathbf{x}_i^s, \mathbf{x}_j^t\right)$ of Eq. (3) is replaced by a generalized joint cost measure as follows, which is inspired by previous works [31], [46]:

$$c\left(F(\mathbf{x}_i^s), y_i^s; F(\mathbf{x}_j^t), \hat{y}_j^t\right) = \lambda_1 c\left(F(\mathbf{x}_i^s), F(\mathbf{x}_j^t)\right) + \lambda_2 L\left(y_i^s, \hat{y}_j^t\right), \tag{4}$$

where $F(\mathbf{x}_i^s)$ is the source representation of sample $\mathbf{x}_i^s$, while $F(\mathbf{x}_j^t)$ is the target representation. The hyperparameters λ_1 and λ_2 are two positive factors to scale the different distance terms. Considering the true label of target sample $\mathbf{x}_j^t$ is unavailable in UDA, its prediction $\hat{y}_j^t = G(F(\mathbf{x}_j^t))$ generated by classifier is utilized to replace the ground truth, as described in Eq. (4). Finally, the original problem can be depicted as follows:

$$W'(\mu_s, \mu_t) = \min_{\gamma \in \Pi(\mu_s, \mu_t)} \mathcal{L}'_{ot} , \tag{5}$$

$$\mathcal{L}'_{ot} = \sum_{i}^{m} \sum_{j}^{n} \gamma_{ij} c\left(F(\mathbf{x}_i^s), y_i^s; F(\mathbf{x}_j^t), \hat{y}_j^t\right) . \tag{6}$$

In our case, $c(\cdot, \cdot)$ is the ℓ_2^2 norm, the solution of the above optimization problem can be considered as the Wasserstein distance. Interestingly, the shared latent space and label space across domains could be matched by minimizing their Wasserstein distance, which is proven in [51].

C. Mixup in Frequency Domain

In this work, we introduce a new Mixup-based approach to transform the trials within domains, as a data augmentation technique. Different from the previous mixup-based augmentation methods [39], [40], [41], the proposed method mixes up the input signals in Frequency domain, with a fixed mixup ratio.

1) Mixup in Source Domain: Given a pair of inputs $(\mathbf{x}_A^s, y_A^s)$ and $(\mathbf{x}_B^s, y_B^s)$ from source domain, each of them is converted into frequency domain with Fast Fourier Transform (FFT) as following:

$$F_k^s = \sum_{j=0}^{T-1} e^{-i\frac{2\pi}{n}jk}\mathbf{x}_j^s, k = 0, 1, \ldots, T-1, \tag{7}$$

where F_k^s is the k-th FFT coefficient of signal $\mathbf{x}^s$. Next, the fixed mixup method is applied to two FFT coefficients series $\{F_{A,k}^s\}_{k=0}^{T-1}$ and $\{F_{B,k}^s\}_{k=0}^{T-1}$:

$$\{\tilde{F}_k^s\}_{k=0}^{T-1} = \lambda \cdot \{F_{A,k}^s\}_{k=0}^{T-1} + (1-\lambda) \cdot \{F_{B,k}^s\}_{k=0}^{T-1}, \tag{8}$$

$$\tilde{y}^s = \lambda \cdot y_A^s + (1-\lambda) \cdot y_B^s, \tag{9}$$

where λ denotes the fixed mixup ratio. Finally, the derived mixup coefficients series $\{\tilde{F}_k^s\}_{k=0}^{T-1}$ is transformed with inverse FFT (iFFT) to obtain a new mixed signals $\tilde{\mathbf{x}}^s$, which is taken as input by embedding block. Formally, the optimization objective of our mixup approach in source domain can be defined as:

$$\mathcal{L}_{sm} = \frac{1}{\bar{m}} \sum_{j=1}^{\bar{m}} \tilde{y}_j^s \log\left(p\left(y \mid \tilde{\mathbf{x}}_j^s\right)\right), \tag{10}$$

where $p(y \mid \tilde{\mathbf{x}})$ is the prediction generated from classifier for an input mixup trial $\tilde{\mathbf{x}}$, $\bar{m}$ is the number of mixed trials in source domain.

2) Mixup in Target Domain: Similarly, the aforementioned mixup pipeline in source domain can be applied to target domain, with only a few modifications. Since the scarcity of labels for target domain, we utilize the pseudo labels of target samples generated from classifier as their supervision. For more semantic mixup representations from target domain, only those instances with high certainty would be performed our mixup strategy. To this end, we randomly pick two samples $(\mathbf{x}_A^t, \hat{y}_A^t)$ and $(\mathbf{x}_B^t, \hat{y}_B^t)$ from target domain, each of them should satisfy the following criterion:

$$\hat{y}_i^t = \arg\max_c p(y|\mathbf{x}_i^t), \ max(p(y|\mathbf{x}_i^t)) > \sigma, \tag{11}$$

where $\hat{y}_i^t$ denotes the pseudo label of $\mathbf{x}_i^t$, σ is a confidence threshold to control the number of target samples performing mixup strategy. In the following, a new mixup trial $\tilde{\mathbf{x}}^t$ and its label $\tilde{y}^t$ are created by two selected target samples with Eq. (7) $\sim$ Eq. (9). Finally, the objective of our mixup strategy in target domain is formulated as:

$$\mathcal{L}_{tm} = \frac{1}{\bar{n}} \sum_{k=1}^{\bar{n}} \tilde{y}_k^t \log\left(p\left(y \mid \tilde{\mathbf{x}}_k^t\right)\right), \tag{12}$$

where $\bar{n}$ is the number of mixed trials in target domain.

D. Optimization of Network

The baseline network is jointly optimized by the final objective as follows:

$$\min_{\Theta} \ (1-\alpha_1)\mathcal{L}_{clf} + \alpha_1\mathcal{L}_{sm} + \alpha_2\mathcal{L}_{tm} + \mathcal{L}'_{ot}, \tag{13}$$

where Θ denotes the parameters of network, α_1 and α_2 are the hyperparameters to balance different loss items.

IV. Experiments

We evaluate our proposed framework on three public EEG-based datasets, including BCI IV dataset IIb, BCI IV dataset IIa and CLA MI. And compare their classification performance with previous state-of-the-art transfer learning methods.

A. Datasets

1) BCI IV IIb [52]: This dataset records the EEG signals from nine different subjects with 3 electrodes, at the sampling rate of 250 Hz. It contains five sessions for each subject, each session contains two categories of EEG signals, including imagery movements of the left hand and right hand. Note that the segment between 3.5 $\sim$ 7 seconds of each trial is utilized in our experiments.

2) BCI IV IIa [53]: This dataset is more challenging than BCI IV IIb. It records the 22-channel EEG signals from nine different subjects at the sampling rate of 250 Hz. Each subject performed four imagery movement experiments at two session, and each session contains 288 trials of four categories, including the left hand, the right hand, the feet and the tongue. Note that the segment between 2.5 $\sim$ 6 seconds of each trial is utilized in our experiments.

3) CLA MI [54]: In the work of [54], Kaya *et al* published a large set of EEG data collected in four experiments of motor imagery (MI), where CLA MI is one of interaction paradigms. In CLA MI paradigm, three imageries from seven subjects are recorded by a standard 10-20 EEG cap (with 21 channels) at sampling rate of 200 Hz, including left-hand movement, right-hand movement and passive mental imagery. Each imagery signal has the length of one second and thus contains 200 sample points. In the following experiments, the segment between 0.15 $\sim$ 1 seconds of each trial after stimulus onset time is extracted for classification.

B. Comparison Algorithms

In the following experiments, the previous state-of-the-art algorithms are employed for comparison, including EA [18], transfer component analysis (TCA) [25], joint distribution adaptation (JDA) [26], BOTDA [7], DANN [28], CDAN [30] and DJDAN [22] methods. For traditional transfer learning methods, like EA, TCA, JDA and BOTDA, Common Spatial Pattern (CSP) method [11] is utilized to extract features. On binary classification task, two components decomposed by CSP is selected, and Linear Discriminant Analysis (LDA) [18] is employed as binary classifier. On multi-category classification task, we select 12 components and 6 components extracted from CSP for datasets IIa and CLA MI, respectively. Then a multi-class SVM classifier is trained to achieve robust performance. For fair comparison, the introduced baseline network in this work is selected as the backbone for all deep adaptation methods.

C. Implementation Details

Following the traditional protocol of unsupervised domain adaptation for BCI, we adopt a Leave-One-Out principle to delineate the source and target domains. Concretely, only one

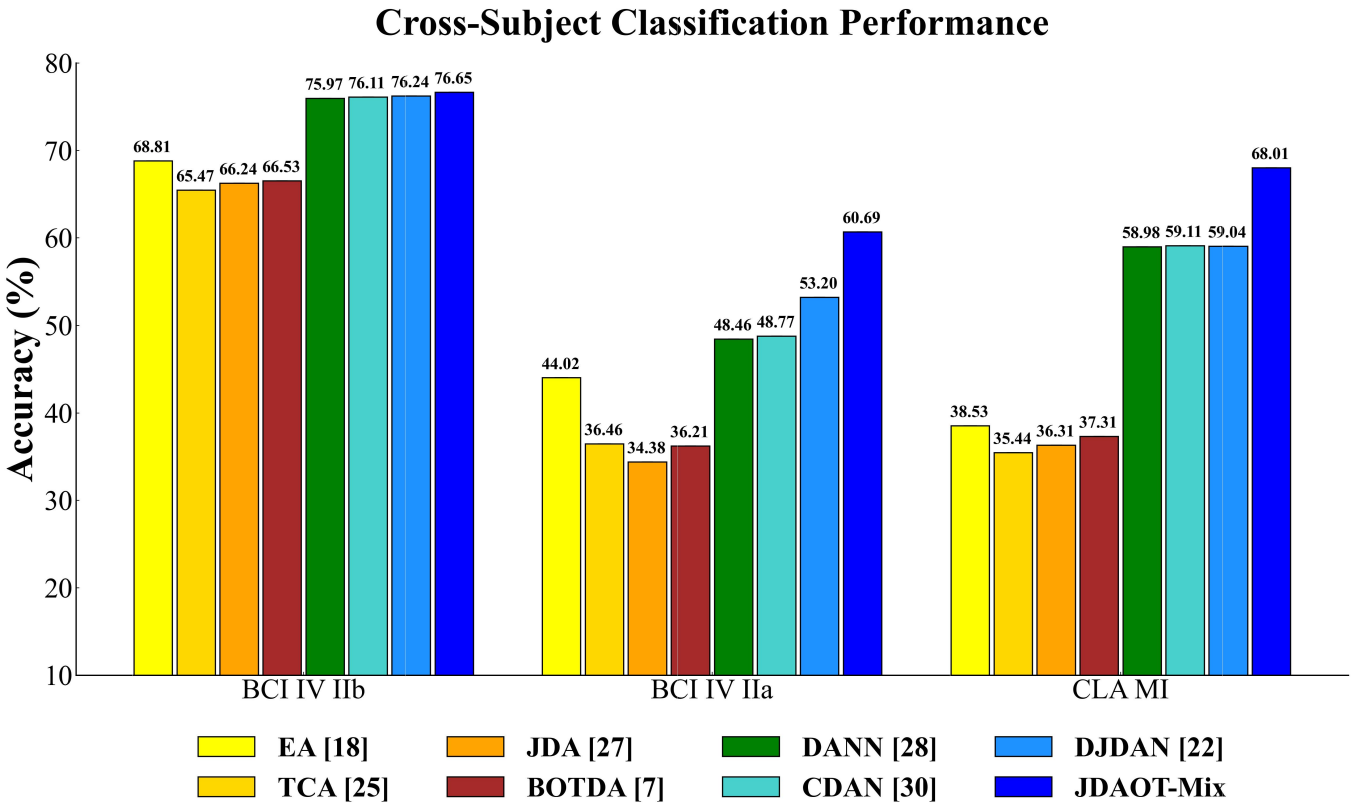


Fig. 3. The average performance (%) of different algorithms on three evaluation datasets, including BCI IV IIb, IIa and CLA MI datasets.

TABLE II
THE SETTING OF HYPERPARAMETERS FOR EVALUATION DATASETS

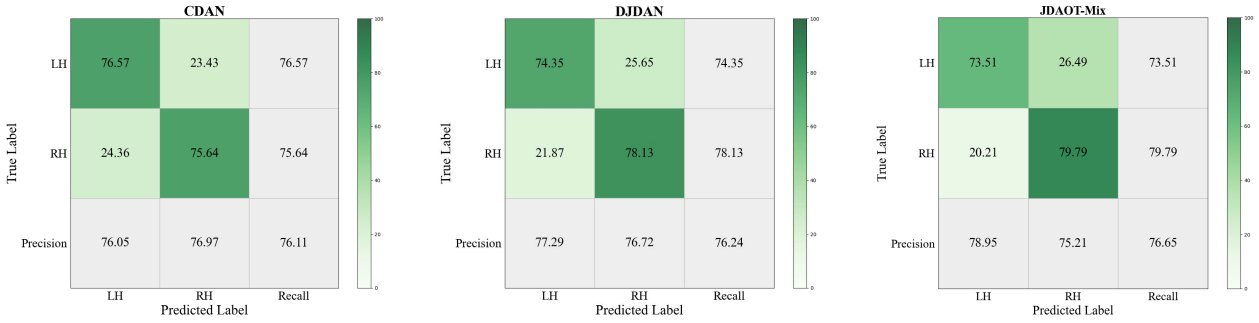
Dataset	λ_1	λ_2	λ	σ	α_1	α_2
IIb	4e-4	0.4	0.8	0.6	0.2	0.2
IIa	4e-4	0.4	0.4	0.6	0.2	0.2
CLA MI	4e-4	0.4	0.6	0.6	0.2	0.2

subject is selected as target domain, while other subjects are source domain. Specially, we only utilize the training set for each subject during training. For all evaluation datasets, we apply a third-order Butterworth filter to pre-process the raw EEG signals, where the filter band is set as 4-38 Hz as previous works [55]. For computation efficiency, all the signals are resampled to 128Hz. We train the model for 300 epochs, using Adam optimizer with momentum and weight decay set to 0.9 and 0.001, respectively. The learning rate for all layers is set to 0.0015. CSP algorithm is implemented by the MNE library [56], and deep methods are implemented by PyTorch [57]. All the hyperparameters of Eq. (4) $\sim$ (13) are empirically set as in Table II.

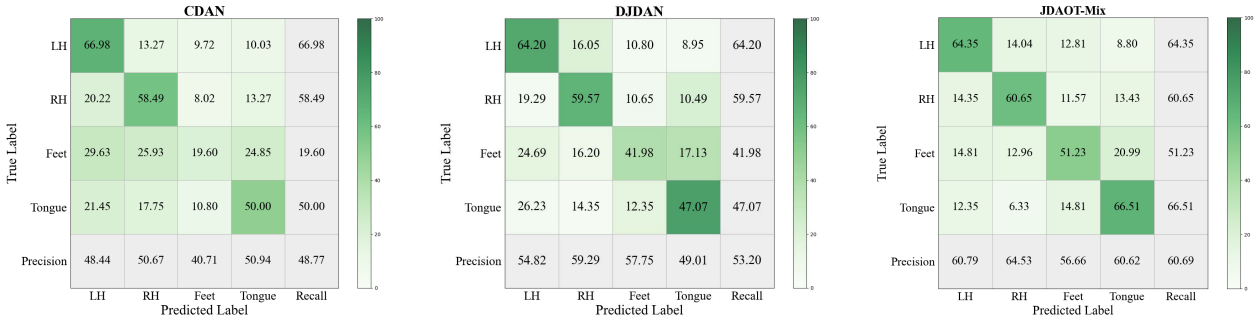
V. RESULTS

The results of all evaluation datasets are exhibited in Figure 3. On binary-category classification task, all the transfer learning methods based on deep network show a significant superior performance than traditional methods, like EA, TCA and JDA. It is contributed to a powerful expressivity of the neural network. Compared to adversarial domain adaptation methods, our approach only shows the slightly improved performance. This indicates that both JDAOT-Mix and those adversarial-based approaches achieve the comparable performance on simple motor imagery recognition task. However, it can be observed that transfer knowledge across different subjects is more difficult on the multi-class classification task. On dataset IIa, traditional approaches only achieve the average performance between 34.38% and 44.02%, while adversarial-based methods achieve the accuracy below 55%. In comparison of previous methods, JDAOT-Mix obtains an average accuracy of 60.69%, which outperforms other methods by a large margin between 7.49% and 26.31%. The classification performance is also greatly improved on CLA MI dataset. It demonstrates that adapt joint distributions across different domains with optimal transport are more effective than subspace alignment and adversarial training, resulting in improving the performance of cross-subject transfer learning for EEG-based motor imagery. Moreover, the introduced MFD strategy also helps to generalize the model.

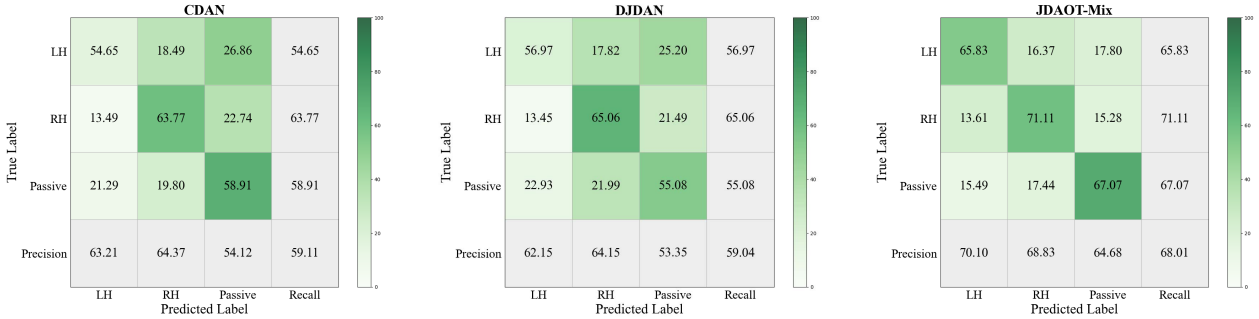
We also investigate the average performance of three condition-based domain adaptation methods (CDAN, DJDAN and JDAOT-Mix) on each category, as shown in Figure 4. The value in the diagonal of each confusion matrix indicates the average accuracy of each task among on all subjects. The precision and recall evaluation metrics are also attached to



(a) BCI IV IIb



(b) BCI IV IIa



(c) CLA MI

Fig. 4. Confusion matrices for deep methods on different datasets. (a) BCI IV IIb; (b) BCI IV IIa. (c) CLA MI. The **L**eft **H**and and **R**ight **H**and imagery are denoted as '**LH**' and '**RH**', respectively.

the last row and last column of each confusion matrix, respectively. On BCI IV IIb dataset, all deep methods perform well on both motor imagery tasks, with the accuracy more than 74%. Specially, our methods performs better on the right hand (RH) imagery task, while CDAN and DJDAN do the opposite. On the more challenging dataset IIa with four motor imagery tasks, both adversarial-based approaches only have a relative precise recognition ability on left hand, right hand and tongue imagery tasks, but totally fail to differ feet imagery from others. By contrast, the proposed method still remains the better recognition performance on all imagery tasks, despite it only achieves the average precision of 56.66% and the recall rate of 51.23% on feet imagery task. It can be explained by the same activate brain region of tongue and feet imagery task, which making the decision to distinguish them from each other become more difficult. On additional multi-categories dataset CLA MI, JDAOT-Mix outperforms the other comparison algorithms by significant margins in all motor imagery tasks. In summary, the experimental results on all datasets demonstrate the efficacy of the proposed framework.

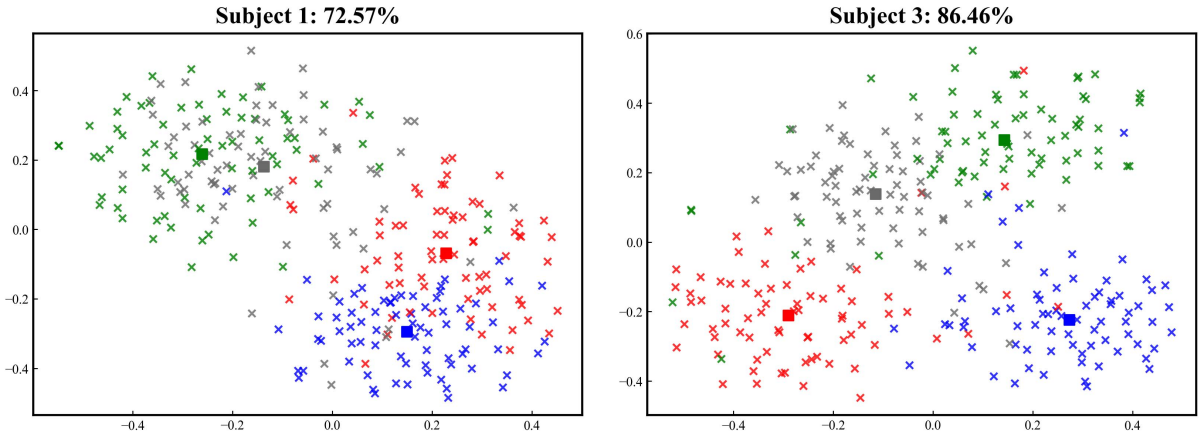
TABLE III
THE RESULTS (%) OF INVESTIGATING THE EFFECTS OF OUR COMPONENTS ON EVALUATION DATASETS

$\mathcal{L}_c$	$\mathcal{L}_{sm}$	$\mathcal{L}_{tm}$	$\mathcal{L}_{ot}$	IIa	IIb	CLA MI
✓				51.70	74.06	58.74
✓	✓			52.74	74.76	60.77
✓	✓	✓		53.05	75.00	60.86
✓			✓	58.68	75.86	66.05
✓	✓	✓	✓	**60.69**	**76.65**	**68.01**

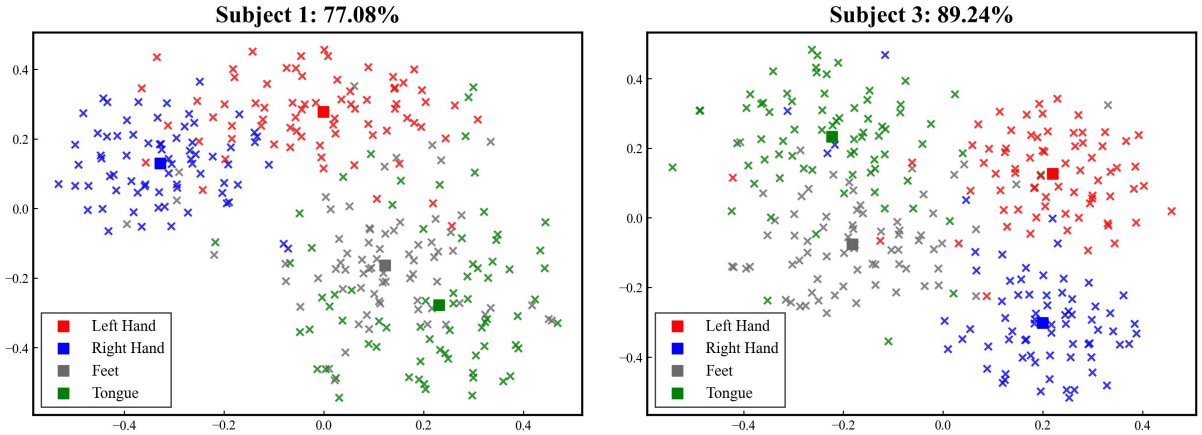
VI. ABLATION STUDIES

A. Relative Contribution of Each Component

We conduct ablation studies to investigate the relative contribution of each component of the proposed approach. As can be seen in Table III, our MFD strategy in source domain slightly improve the baseline on average by 1.04% on dataset IIa, while the MFD in both domains greatly improve the baseline by 1.35%. In addition, the joint distribution alignment



(a) JDAOT with vanilla mixup



(b) JDAOT with MFD

Fig. 5. Features visualization of JDAOT combined with different mixup strategies on dataset IIa. (a) JDAOT with vanilla mixup; (b) JDAOT with MFD. Each cross symbol represents a sample, while each square represents the centroid of the corresponding category.

based on optimal transportation also achieves the performance improvement by 6.98%. By integrating all the components together, our framework improves the baseline by average accuracy of 8.99% on dataset IIa, outperforming other variants and state-of-the-art methods. The similar conclusion can be also drawn from datasets IIb and CLA MI. All these results indicate that the components of our framework are effective for performance improvement.

B. Comparison of Different Mixup Strategies

In this experiment, we show the difference of feature spaces learned with vanilla mixup method [39] and proposed MFD strategy. Specifically, we train JDAOT with different mixup methods and fix other experimental conditions to be same, then the learned features are visualized with t-SNE [58]. For simplicity, the results of two subjects (subject 1 and 3) are randomly picked from dataset IIa for visualization, which are shown in Figure 5. The first row of Figure 5 exhibits the results of JDAOT with vanilla mixup method, while the second row exhibits JDAOT with MFD strategy. It can be observed that the features learned with MFD strategy is more discriminative than those learned with vanilla mixup method. Compared with vanilla mixup method, the inter-class distances of features learned with MFD strategy are larger (i.e., feet vs tongue features of subject 1), while the intra-class distances are smaller (i.e., left vs right hand features of subject 3). This phenomenon explains the better classification performance of JDAOT with MFD, indicating the effectiveness of MFD strategy.

TABLE IV
THE RESULTS (%) OF DIFFERENT MIXUP RATIOS ON EVALUATION DATASETS

λ	random	0.2	0.4	0.6	0.8
IIb	76.07	76.32	76.18	76.54	**76.65**
IIa	60.10	58.68	**60.69**	60.22	59.76
CLA MI	67.23	67.24	67.87	**68.01**	66.96

C. Comparison of Different Mixup Ratios

We further explore the effect of different mixup ratio settings on classification performance, which is exhibited in Table IV. The "random" represents the introduced mixup method with random ratio sampled from the β distribution. In comparison of mixing the input trials in frequency domain with random ratio, fixed mixup ratios might reduce the randomness the mixed samples and generate more distinguishable representations. Consequently, the improved performance is

observed on evaluation datasets, such as $\lambda = 0.8$ for BCI IV IIb, $\lambda = 0.4$ for BCI IV IIa and $\lambda = 0.6$ for CLA MI.

VII. Conclusion

In this work, we develop a joint distribution adaptation with optimal transportation and frequency mixup (JDAOT-Mix) framework, aiming to improve the performance of cross-subject transfer learning task in motor imagery BCIs. Our proposition jointly adapt the marginal distributions and conditional distributions across domains with performing the optimal transportation between source and target representations. In addition, a novel mixup-based augmentation, namely mixup in frequency domain (MFD), is also designed for EEG-based BCI systems. Empirical experiments on the public EEG datasets demonstrate that our JDAOT-Mix exhibits a competitive performance when compared to previous domain adaptation methods. This indicates that the proposed JDAOT-Mix provides a promising solution for cross-subject transfer learning in BCI systems.

References

[1] P. Arpaia, L. Duraccio, N. Moccaldi, and S. Rossi, "Wearable brain–computer interface instrumentation for robot-based rehabilitation by augmented reality," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 9, pp. 6362–6371, Sep. 2020.

[2] V. K. Benzy, A. P. Vinod, R. Subasree, S. Alladi, and K. Raghavendra, "Motor imagery hand movement direction decoding using brain computer interface to aid stroke recovery and rehabilitation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 3051–3062, Dec. 2020.

[3] M. Shi, C. Yang, and D. Zhang, "A novel human-machine collaboration model of an ankle joint rehabilitation robot driven by EEG signals," *Math. Problems Eng.*, vol. 2021, pp. 1–8, Mar. 2021.

[4] G. Edlinger, R. Prueckl, G. Krausz, C. Holzner, and C. Guger, "P4–24 P300 and SSVEP based brain-computer interface for control of a smart home virtual environment," *Clin. Neurophysiol.*, vol. 121, p. S126, Oct. 2010.

[5] T. Hafeez, S. M. U. Saeed, A. Arsalan, S. M. Anwar, M. U. Ashraf, and K. Alsubhi, "EEG in game user analysis: A framework for expertise classification during gameplay," *PLoS ONE*, vol. 16, no. 6, Jun. 2021, Art. no. e0246913.

[6] J. Lee, D. Lee, I. Jeong, and J. Cho, "A study on the content of mental and physical stability game in virtual reality through EEG detection," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2021, pp. 693–696.

[7] V. Peterson et al., "Transfer learning based on optimal transport for motor imagery brain-computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 2, pp. 807–817, Feb. 2022.

[8] Z. Wan, R. Yang, M. Huang, N. Zeng, and X. Liu, "A review on transfer learning in EEG signal analysis," *Neurocomputing*, vol. 421, pp. 1–14, Jan. 2021.

[9] B. Graimann, J. Huggins, S. Levine, and G. Pfurtscheller, "Visualization of significant ERD/ERS patterns in multichannel EEG and ECoG data," *Clin. Neurophysiol.*, vol. 113, no. 1, pp. 43–47, 2002.

[10] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (FBCSP) in brain-computer interface," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jun. 2008, pp. 2390–2397.

[11] M. Grosse-Wentrup and M. Buss, "Multiclass common spatial patterns and information theoretic feature extraction," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 8, pp. 1991–2000, Aug. 2008.

[12] Y. Li, W. Zheng, Z. Cui, T. Zhang, and Y. Zong, "A novel neural network model based on cerebral hemispheric asymmetry for EEG emotion recognition," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1561–1567.

[13] H. Fang, J. Jin, I. Daly, and X. Wang, "Feature extraction method based on filter banks and Riemannian tangent space in motor-imagery BCI," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 6, pp. 2504–2514, Jun. 2022.

[14] Z. Gao et al., "Complex networks and deep learning for EEG signal analysis," *Cognit. Neurodyn.*, vol. 15, pp. 369–388, Aug. 2021.

[15] M. Gosak, M. Milojević, M. Duh, K. Skok, and M. Perc, "Networks behind the morphology and structural design of living systems," *Phys. Life Rev.*, vol. 41, pp. 1–21, Jul. 2022.

[16] D. Zhang, K. Chen, D. Jian, and L. Yao, "Motor imagery classification via temporal attention cues of graph embedded EEG signals," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 9, pp. 2570–2579, Sep. 2020.

[17] Z. Ni, J. Xu, Y. Wu, M. Li, G. Xu, and B. Xu, "Improving cross-state and cross-subject visual ERP-based BCI with temporal modeling and adversarial training," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 369–379, 2022.

[18] H. He and D. Wu, "Transfer learning for brain–computer interfaces: A Euclidean space data alignment approach," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 2, pp. 399–410, Feb. 2020.

[19] W. Mu and B.-L. Lu, "Examining four experimental paradigms for EEG-based sleep quality evaluation with domain adaptation," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 5913–5916.

[20] X. Jiang, K. Xu, and W. Chen, "Transfer component analysis to reduce individual difference of EEG characteristics for automated seizure detection," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Oct. 2019, pp. 1–4.

[21] X. Tang and X. Zhang, "Conditional adversarial domain adaptation neural network for motor imagery EEG decoding," *Entropy*, vol. 22, no. 1, p. 96, Jan. 2020.

[22] X. Hong et al., "Dynamic joint domain adaptation network for motor imagery classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 556–565, 2021.

[23] P. Chen, Z. Gao, M. Yin, J. Wu, K. Ma, and C. Grebogi, "Multiattention adaptation network for motor imagery recognition," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 52, no. 8, pp. 5127–5139, Aug. 2022.

[24] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2009.

[25] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[26] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2200–2207.

[27] B. Wang, W. Li, W. Fan, X. Chen, and D. Wu, "Alzheimer's disease brain network classification using improved transfer feature learning with joint distribution adaptation," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 2959–2963.

[28] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, Jan. 2016.

[29] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2208–2217.

[30] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1640–1650.

[31] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy, "Joint distribution optimal transportation for domain adaptation," 2017, *arXiv:1705.08848*.

[32] J. C. Shaw, "Correlation and coherence analysis of the EEG: A selective tutorial review," *Int. J. Psychophysiol.*, vol. 1, no. 3, pp. 255–266, Mar. 1984.

[33] H. Zhao, Q. Zheng, K. Ma, H. Li, and Y. Zheng, "Deep representation-based domain adaptation for nonstationary EEG classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 535–545, Feb. 2021.

[34] L. Taylor and G. Nitschke, "Improving deep learning with generic data augmentation," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2018, pp. 1542–1547.

[35] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.

[36] A. Mikolajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *Proc. Int. Interdiscipl. PhD Workshop (IIPhDW)*, May 2018, pp. 117–122.

[37] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 13001–13008.

[38] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101552.

[39] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–13.

[40] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6023–6032.

[41] J.-H. Kim, W. Choo, and H. O. Song, "Puzzle Mix: Exploiting saliency and local statistics for optimal mixup," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5275–5285.

[42] Y. Luo and B.-L. Lu, "EEG data augmentation for emotion recognition using a conditional Wasserstein GAN," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 2535–2538.

[43] M. M. Krell and S. K. Kim, "Rotational data augmentation for electroencephalographic data," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2017, pp. 471–474.

[44] E. Lashgari, D. Liang, and U. Maoz, "Data augmentation for deep-learning-based electroencephalography," *J. Neurosci. Methods*, vol. 346, Dec. 2020, Art. no. 108885.

[45] N. T. Gayraud, A. Rakotomamonjy, and M. Clerc, "Optimal transport applied to transfer learning for P300 detection," in *Proc. BCI 7th Graz Brain-Comput. Interface Conf.*, 2017, p. 6.

[46] B. B. Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty, "DeepJDOT: Deep joint distribution optimal transport for unsupervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 447–463.

[47] D. Kostas and F. Rudzicz, "Thinker invariance: Enabling deep neural networks for BCI across more people," *J. Neural Eng.*, vol. 17, no. 5, Oct. 2020, Art. no. 056008.

[48] W. Ko, E. Jeon, S. Jeong, J. Phyo, and H.-I. Suk, "A survey on deep learning-based short/zero-calibration approaches for EEG-based brain–computer interfaces," *Frontiers Hum. Neurosci.*, vol. 15, May 2021, Art. no. 643386.

[49] Y. Lei, A. N. Belkacem, X. Wang, S. Sha, C. Wang, and C. Chen, "A convolutional neural network-based diagnostic method using resting-state electroencephalograph signals for major depressive and bipolar disorders," *Biomed. Signal Process. Control*, vol. 72, Feb. 2022, Art. no. 103370.

[50] 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," *J. Neural Eng.*, vol. 15, no. 5, pp. 056013.1–056013.17, Oct. 2018.

[51] I. Redko, A. Habrard, and M. Sebban, "Theoretical analysis of domain adaptation with optimal transport," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2017, pp. 737–753.

[52] R. Leeb, C. Brunner, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "BCI competition 2008—Graz data set B," Graz Univ. Technol., Graz, Austria, Tech. Rep., 2008, pp. 1–6.

[53] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "BCI competition 2008—Graz data set A," Inst. Knowl. Discovery, Lab. Brain-Comput. Interfaces, Graz Univ. Technol., Graz, Austria, Tech. Rep., 2008, pp. 136–142.

[54] M. Kaya, M. K. Binli, E. Ozbay, H. Yanar, and Y. Mishchenko, "A large electroencephalographic motor imagery dataset for electroencephalographic brain computer interfaces," *Sci. Data*, vol. 5, no. 1, pp. 1–16, Dec. 2018.

[55] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[56] A. Gramfort, "MEG and EEG data analysis with MNE-Python," *Frontiers Neurosci.*, vol. 7, p. 267, Dec. 2013.

[57] A. Paszke et al., "Automatic differentiation in PyTorch," in *Proc. NIPS Workshop*, 2017, pp. 1–4.

[58] J. Donahue et al., "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 647–655.