Improved Domain Adaptation Network Based on Wasserstein Distance for Motor Imagery EEG Classification

Qingshan She, Tie Chen, Feng Fang, Jianhai Zhang, Yunyuan Gao, and Yingchun Zhang, *Senior Member, IEEE*

***Abstract*—Motor Imagery (MI) paradigm is critical in neural rehabilitation and gaming. Advances in brain-computer interface (BCI) technology have facilitated the detection of MI from electroencephalogram (EEG). Previous studies have proposed various EEG-based classification algorithms to identify the MI, however, the performance of prior models was limited due to the cross-subject heterogeneity in EEG data and the shortage of EEG data for training. Therefore, inspired by generative adversarial network (GAN), this study aims to propose an improved domain adaption network based on Wasserstein distance, which utilizes existing labeled data from multiple subjects (source domain) to improve the performance of MI classification on a single subject (target domain). Specifically, our proposed framework consists of three components, including a feature extractor, a domain discriminator, and a classifier. The feature extractor employs an attention mechanism and a variance layer to improve the discrimination of features extracted from different MI classes. Next, the domain discriminator adopts the Wasserstein matrix to measure the distance between source domain and target domain, and aligns the data distributions of source and target domain via adversarial learning strategy. Finally, the classifier uses the knowledge acquired from the source domain to predict the labels in the target domain. The proposed EEG-based MI classification framework was evaluated by two open-source datasets, the BCI Competition IV Datasets 2a and 2b. Our results demonstrated that the proposed framework could enhance the performance of EEG-based MI detection, achieving better classification results compared with several state-of-the-art algorithms. In conclusion, this study is promising in helping the neural rehabilitation of different neuropsychiatric diseases.**

***Index Terms*—Motor imagery (MI), deep neural network, electroencephalogram (EEG), adversarial learning, domain adaptation, machine learning.**

I. Introduction

BRAIN-COMPUTER interfaces (BCI) enable users to manipulate external devices by decoding their own neuronal activities into specific commands directly. Currently, BCI has been widely utilized in various areas such as exoskeleton rehabilitation robot, fatigue detection as well as intelligent furniture [1], [2], [3]. Electroencephalogram (EEG) is one of the most common neuroimaging technologies that acquire brain information as input for BCI systems. With greater portability, convenience and lower costs, EEG has several advantages over other neuroimaging modalities such as magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI), and positron emission tomography (PET) [4], [5], [6]. Recently, EEG-based BCI systems have been employed for the classification of motor imagery (MI) signals for neurological rehabilitation of various diseases such as stroke. For example, Lee et al. [7] proposed a novel network based on the weighted phase lag index (wPLI) and directed transfer function (DTF) to calculate the predictor of motor impairments in stroke rehabilitation. However, due to the high signal variance in temporal dimension, low signal-to-noise ratio (SNR) as well as the heterogeneity between different subjects, it is difficult to accurately distinguish features of different motor imagery tasks.

Machine learning approaches have been employed to decode MI-EEG signals. For example, previous study utilized common spatial pattern (CSP) strategy [8], which can compute the optimal spatial filter and maximize (or minimize) the ratio of filtered variance between different categories to extract the spatial features from EEG signals, then, different classifiers such as linear discrimination analysis (LDA), support vector machine (SVM) and nonlinear Bayesian were applied to classify different MI-EEG signals [9]. Afterwards, the CSP-based classification methods were developed into

Manuscript received 10 October 2022; revised 9 December 2022 and 13 January 2023; accepted 29 January 2023. Date of publication 1 February 2023; date of current version 7 February 2023. This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LZ22F010003, in part by the National Natural Science Foundation of China under Grant 61871427 and Grant 61971168, and in part by the National Key Research and Development Program under Grant 2017YFE0116800. *(Corresponding authors: Qingshan She; Yingchun Zhang.)*

Qingshan She, Tie Chen, and Yunyuan Gao are with the School of Automation, Hangzhou Dianzi University, Hangzhou, Zhejiang 310018, China (e-mail: qsshe@hdu.edu.cn).

Feng Fang and Yingchun Zhang are with the Department of Biomedical Engineering, University of Houston, Houston, TX 77204 USA (e-mail: yingchun.umn@gmail.com).

Jianhai Zhang is with the Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province, Hangzhou 310018, China.

Digital Object Identifier 10.1109/TNSRE.2023.3241846

filter bank CSP (FBCSP) [10], discriminative filter band CSP (DFBCSP) [11], and sub-band CSP (SBCSP) [12]. Taking FBCSP as an example, it decomposes the original frequency into several sub-bands without overlap, and employs CSP to extract features in each band.

Although the aforementioned machine learning methods have made certain progress in MI-EEG classification, owing to the non-stationary nature of EEG signals, these methods cannot precisely extract the complex non-linear features from signals. Therefore, deep learning (DL) strategies were then utilized to improve the extraction of the non-linear features in EEG signals. For instance, Tabar et al. [13] applied short time Fourier transform (STFT) to convert the MI-EEG time series into two-dimensional (2D) images, and then transferred the images into a convolutional neural network (CNN) or a stacked autoencoder (SAE) to classify MI-EEG signals. Schirrmeister et al. [14] proposed deep neural network structures, namely shallow CNN and deep CNN, which can be directly utilized for MI-EEG classification and substantially outperform traditional methods. Sakhavi et al. [15] developed a deep learning model, which can learn envelope representations for MI-EEG classification, and they also fine-tuned the model, enhancing the classification accuracy by 7% on BCI competition dataset IV 2a. In addition, Liu et al. [16] achieved better performance in identifying four-class MI-EEG signals by employing a parallel spatial-temporal self-attention-based CNN approach.

Nevertheless, the results obtained by these deep learning models were still limited because of the lack of annotated data and the heterogeneity of MI-EEG signals collected from different subjects. To tackle this issue, transfer learning, which utilizes the knowledge learned from the source domain to help target domain learning [17], has been employed by some BCI studies. Specifically, Raza et. al [18] adopted a novel covariate shift-detection and adaptation method, which can reduce the difference between two feature spaces. Zanini et al. [19] utilized a Riemannian alignment method to minimize the distance between different domains in Riemannian space, and achieved good MI-EEG classification performance in cross-subject transfer situation. Azab et al. [20] applied a new similarity measure based on the Kullback-Leibler divergence (KL) to the logistic regression classifier, measuring the similarity between two feature spaces and realizing weighted transfer learning. He and Wu [21] explored a Euclidean alignment (EA) approach which can transform and align MI-EEG signals of different trials in the Euclidean space. Apart from these traditional transfer learning methods, some studies have also combined transfer learning approach with deep learning models. Dose et al. [22] utilized transfer learning to adapt the global classifier to single individuals, thus improving the classification performance on specific subjects. Besides, based on GAN [23], Zhao et al. [24] utilized an end-to-end deep domain adaptation method for MI-EEG classification, which can reduce the calibration time for the use of BCI and increase the classification accuracy by 3% in the two BCI competition IV datasets. Phunruangsakao et al. [25] integrated few-shot learning strategy into the deep domain adaptation model to reduce cross-subject variance, thus leveraging the knowledge acquired from several source subjects to further enhance the classification performance on a single target subject.

Even though prior studies have combined the domain adaptation approach with deep neural networks and obtained good performance in MI-EEG classification, these strategies still shared some specific limitations. First, previous models merely considered improving the structure of neural network utilized to extract feature from raw MI-EEG signals, impairing its effectiveness in handling the spatial and temporal non-stationarity of the signals. Therefore, the features extracted by previous neural networks might omit some useful information, and the discrimination of features from different motor imagery tasks might be reduced, which would negatively affect the performance of the following domain adaptation and classification. Second, existing models adopted the adversarial loss to realize adversarial domain adaptation, which has problems such as gradient vanish or model collapse that will degrade the efficacy of domain adaptation.

To solve the issues abovementioned, in this study, we were motivated to propose an improved domain adaptation network based on the Wasserstein distance matrix by combining an improved feature extractor with an adversarial domain adaptation model. In the feature extractor, an attention mechanism called convolutional block attention model (CBAM) was integrated into the model to enhance the discrimination of spatial features. Meanwhile, a variance layer was employed to extract temporal features, replacing the convolutional layer which only has a fixed-size kernel. In the adversarial domain adaptation, instead of the original adversarial loss function, the Wasserstein distance matrix was utilized. Through the adversarial training between the feature extractor and the domain discriminator, the data distributions of the source and target domains can be aligned, and the domain invariant representations can be obtained. In this way, the cross-subject discrepancy was reduced, so that the labeled data from source subjects can be applied to enlarge the amount of data for training, which helped make the classifier more reliable in classifying the target data. This approach has provided a better solution for classifying new subjects' MI-EEG data when the labels of data are totally unknown during the training process, contributing to the deployment of BCI in more practical scenarios.

II. Material and Methods

A. Data Description

The Dataset 2a and 2b from BCI Competition IV were utilized in this study. In Dataset 2a, EEG signals were collected from 22 channels in two recording sessions from nine healthy participants (A01 to A09) with a sampling rate of 250 Hz. The participants were instructed to perform four motor imagery tasks, including the movement of left hand, right hand, feet, and tongue, respectively. Each recording session contains 288 trials of EEG data (72 trials for each task), the first session contains the class labels for all trials, whereas the second session are used to test the classifier and hence to evaluate the performance [26]. Each trial of EEG data is segmented

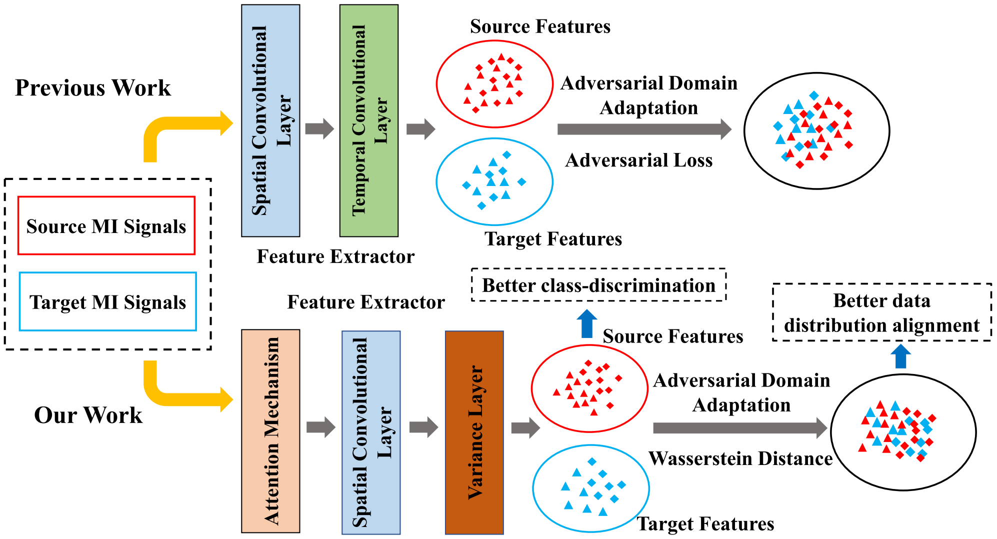


Fig. 1. Comparison between previous domain adaptation and our work.

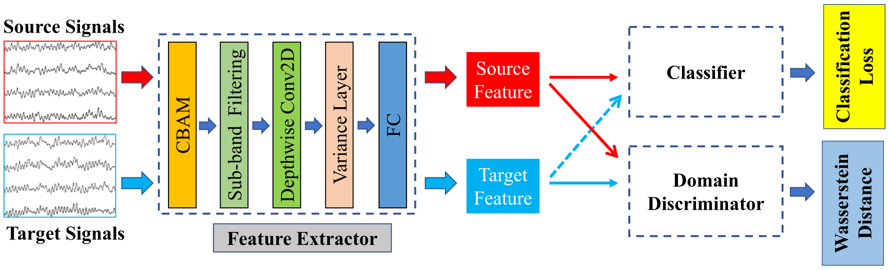


Fig. 2. Architecture of the proposed model.

from the 2nd second to 6th second. In this paper, we used all the trials in our experiments, regarding EEG data in the first session for training and that in the second session for the test.

In Dataset 2b, EEG signals were collected from 3 channels (C3, Cz, and C4) in five recording sessions from nine participants (B1 to B9) with a sampling rate of 250 Hz. The participants were instructed to perform two motor imagery tasks, including the movement of left hand and right hand, respectively. Each of the first two sessions contains 120 trials, and each of the other three sessions has 160 trials. Meanwhile, the first three sessions contain the class labels for all trials, whereas the remaining two sessions are used to test the classifier and hence to evaluate the performance [27]. Each trial of EEG data is segmented from the 3rd second to 7th second. In this paper, we utilized all the trials in our experiments, so each subject has a total of 400 trials and 320 trials for training and test, respectively.

B. Data Preprocessing

The EEG signals in the aforementioned datasets were first filtered using a Fourth-order Butterworth band-pass filter with the frequency ranging from 8 Hz to 32 Hz.

The exponential moving standardization method was then utilized to eliminate occasional noises, which helped to obtain motor imagery signals with high signal-to-noise ratio (SNR), thus enhancing the classification performance.

C. Domain Adaptation

The EEG data collected in a session is defined as $\{(x_i, y_i)\}_{i=1}^{n}$, where n is the total number of samples, $x_i \in \mathbb{R}^{C\times T}$ denotes an EEG trial with C electrodes and T sampling points, and $y_i \in \mathbb{R}^{N}$ is the corresponding label of N categories. Thereby, the labeled source domain can be expressed as $D_s = \{x_i^s, y_i^s\}_{i=1}^{n_s}$, while the unlabeled target domain can be expressed as $D_t = \{x_j^t\}_{j=1}^{n_t}$. In the context of EEG data, domain adaptation aims to make good use of the labeled data in the source domain, extracting important information and training the classifier adapted to the target domain, where the labels are not sufficient [28], [29]. In this way, the trained classifier can perform better in classifying the data in target domain.

D. Network Architecture

Different from previous domain adaptation network, the framework proposed in this study utilizes the attention mechanism and the variance layer in the feature extractor to increase the discrimination of motor imagery features. Then, based on the Wasserstein distance matrix instead of the adversarial loss function, the framework conducts domain adaptation to decrease the cross-subject discrepancy. Comparison between previous domain adaptation work and the framework in this study is presented in Fig. 1. In our framework, not only the features of different motor imagery tasks can be more discriminative, but also the distributions of different participants' MI-EEG data can be better aligned. In this way, the MI-EEG data from multiple participants can be utilized to help classify a single participant's data, solving the data shortage problem and improving the classification results.

As is shown in Fig. 2, there are three main modules in the proposed model, including a feature extractor, a classifier, and a domain discriminator. During the training process of our model, the source and target EEG signals were first sent to the feature extractor, where a sub-band filter and a convolutional layer combined with CBAM were utilized to extract spatial information. A variance layer was then employed to extract temporal information. Subsequently, the source and target features can be obtained, defined as source domain D_s and target domain D_t respectively. Through minimizing the Wasserstein distance between the two domains in an adversarial manner, the data distribution discrepancy between the two domains can be reduced, so that the data distributions of two domains were aligned and domain invariant feature representations were learned simultaneously. As such, the labeled data from multiple subjects (source domain) can be leveraged to help enhance the classification performance on the single subject (target domain).

1) Feature Extractor: To extract task-related features from raw EEG signals more effectively, the feature extractor contains a spatial convolutional module and a variance layer to extract spatial features and temporal features, respectively.

In the spatial convolutional module, given that when subjects performed different motor imagery tasks, their corresponding body parts stimulated different functional regions of their brains [30], if all the channels were equally treated, the channels having a stronger connection with motor imagery tasks would not be assigned with higher weights, which could negatively impact the quality of the extracted spatial features, eventually resulting in poor classification performance. Therefore, inspired by attention mechanism that has been successfully applied in computer vision field, we integrated CBAM [31], an effective attention mechanism for feed-forward convolutional neural networks, into the spatial convolutional module, which can help increase the discrimination of the extracted spatial features by assigning higher weights to MI-related channels and lower weights to MI-unrelated channels.

The attention mechanism CBAM consists of a channel attention module and a space channel attention module. The channel attention module performs average-pooling and max-pooling operations, mainly focusing on the inter-channel relationship of the input. The input is $\boldsymbol{f} \in \mathbb{R}^{1\times C\times T}$, the output is a channel attention map $\mathbf{M}_{\mathrm{c}}(\boldsymbol{f})$. The computation is:

$$\begin{aligned}\mathbf{M}_{\mathrm{c}}(\boldsymbol{f}) &= \sigma(MLP(\mathrm{AvgPool}(\boldsymbol{f})) + MLP(\mathrm{MaxPool}(\boldsymbol{f})))\\ &= \sigma(\mathbf{W_1}(\mathbf{W_0}(f^{\mathrm{c}}_{\mathrm{avg}})) + \mathbf{W_1}(\mathbf{W_0}(f^{\mathrm{c}}_{\mathrm{max}}))) \qquad (1)\end{aligned}$$

where σ represents the sigmoid function. *MLP* is a multi-layer perceptron (MLP) with a single hidden layer. $\mathbf{W_0}, \mathbf{W_1} \in \mathbb{R}^{1\times 1}$ are the shared MLP weights for inputs. $f^{\mathrm{c}}_{\mathrm{avg}}$ and $f^{\mathrm{c}}_{\mathrm{max}}$ denote average-pooled features and max-pooled features respectively.

The spatial attention module is complementary to the channel attention module, focusing on the locations of informative parts. In EEG-based BCI, this module is useful in recognizing which brain regions were active when the subject was performing motor imagery tasks. The module operates average-pooling and max-pooling along the channel axis, and then concatenates the pooled results to create a spatial attention map $\mathbf{M}_{\mathrm{s}}(\boldsymbol{f}) \in \mathbb{R}^{C\times T}$. The computation is given as follows:

$$\begin{aligned}\mathbf{M}_{\mathrm{s}}(\boldsymbol{f}) &= \sigma(f^{n\times n}([\mathrm{AvgPool}(\boldsymbol{f}); \mathrm{MaxPool}(\boldsymbol{f})]))\\ &= \sigma(f^{n\times n}([f^{\mathrm{s}}_{\mathrm{avg}}; f^{\mathrm{s}}_{\mathrm{max}}])) \qquad (2)\end{aligned}$$

where σ represents the sigmoid function, $\boldsymbol{f}^{n\times n}$ represents the filter size $n \times n$. $f^{\mathrm{s}}_{\mathrm{avg}}$ and $f^{\mathrm{s}}_{\mathrm{max}}$ denote average-pooled features and max-pooled features respectively. In this work, $n = 1$, which can not only reduce the computation parameters and simplify the model, but also can maintain the original size of the input signal, thus preserving the spatial and temporal information for the following feature extraction.

After the attention mechanism, EEG signals were filtered into multiple non-overlapping sub-bands. Since the majority of MI-related information exists in the mu (8-12 Hz) and beta (12-32 Hz) bands, this procedure aimed to localize the discriminative information of MI-EEG signals. According to the experimental results in [32], the 8-30 Hz frequency band achieved a better classification performance compared with 4-40 Hz frequency band. Therefore, in this work, the 8-30 Hz frequency band was adopted. With the bandwidth of 4 Hz, the frequency band was divided into six sub-bands, namely 8-12 Hz, $12-16$ Hz, $\ldots$, $28-32$ Hz, and the parameter m was equal to the number of non-overlapping frequency bands. Next, a depthwise convolutional layer [33] was utilized to extract spatial information, the kernel size was $(C,1)$, wherein the parameter C was the number of electrodes and the depth parameter d was set to control the number of spatial filters per frequency band. As such, the spatial information from all electrodes can be fused together to one single electrode.

In the extraction of temporal information, a variance layer [34] was employed. It obtains the features of a time series by computing the variance v, which can be expressed as:

$$v = Var(s(t)) = \frac{1}{L}\sum_{t=0}^{L-1}(s(t)-\mu)^2 \qquad (3)$$

where $s(t)$ is the input signal, L is the total number of time samplings and μ is the mean of $s(t)$.

The output of spatial convolutional module was sent to the variance layer, and in this procedure, the size of non-overlapping temporal window was set as w, thereby the computation can be expressed as follows:

$$x_V(k) = \frac{1}{w}\sum_{t=w*k}^{(k+1)*w-1}(x(t)-\mu(k))^2 \qquad (4)$$

where $\mu(k)$ is the temporal mean of $x(t)$ within the k^{th} window.

Since the window size is a crucial factor determining the quality of the extracted temporal features, a too big or too small size will eventually affect the classification results, the impact of its size on the classification performance will

TABLE I
PARAMETERS OF THE PROPOSED FEATURE EXTRACTOR

Module	Layer	Parameters	Output
Input	-	-	$1{\times}C{\times}T$
Feature Extractor	CBAM	$1{\times}1$	$1{\times}C{\times}T$
	Sub-band Filtering	m	$m{\times}C{\times}T$
	Depthwise Conv2D	$C{\times}1$, d	$d{\times}m{\times}1{\times}T$
	Variance Layer	$1{\times}w$	$d{\times}m{\times}1{\times}T/w$
	FC	64	-

be discussed in the following experiment section. TABLE I presents the parameters of the proposed feature extractor.

2) Domain Discriminator: According to WGAN [35], the adversarial training mechanism contains a feature extraction and a domain discriminator. In our adversarial training, the feature extractor learned the domain invariant feature representations from source and target domains in order to make the domain discriminator struggle to distinguish which domain the feature came from, while the domain discriminator measured the Wasserstein distance between data distributions of source domain and target domain, trying to figure out the domain to which the data belonged. Finally, the learned feature representations can fool the domain discriminator, which meant that the Wasserstein distance between two domains was minimized, in other words, the discrepancy between the two domains was reduced. As such, the marginal data distributions of two domains were aligned, making it possible to utilize the knowledge acquired from the source domain to help classify the data in the target domain.

For the domain discriminator, the Wasserstein distance between source and target domains can be evaluated by maximizing the domain discriminator loss $\mathcal{L}_{wd}$ with respect to parameter θ_d:

$$\mathcal{L}_{wd}(x^s, x^t) = \frac{1}{n^s} \sum_{x^s \in D^s} f_w(f_g(x^s)) - \frac{1}{n^t} \sum_{x^t \in D^t} f_w(f_g(x^t)) \tag{5}$$

where x^s and x^t are the data samples coming from source domain and target domain, respectively. f_g is a function learned by the feature extractor, which maps samples to a representation with the corresponding parameter θ_f. f_w is a function learned by domain discriminator, which maps the feature representation to a real number with parameter θ_d.

Meanwhile, extra attention should be paid to satisfying the Lipschitz constraint condition of Wasserstein distance, otherwise problems such as capacity underuse and gradient vanishing or exploding could occur, degrading the domain adaptation performance. Gulrajani et al. [36] proposed a method to realize the constraint by using a penalty $\mathcal{L}_{grad}$ on the gradient norm for the domain discriminator parameter θ_d, the expression is:

$$\mathcal{L}_{grad}(\tilde{h}) = (||\nabla_{\tilde{h}} f_w(\tilde{h})||_2 - 1)^2 \tag{6}$$

where $\tilde{h}$ is feature representations extracted from source domain and target domain.

Due to the better differentiation and continuum of the Wasserstein distance, the domain discriminator was trained to optimality first, maximizing the Wasserstein distance between source and target domains. Then, the parameter of domain discriminator was fixed and at the same time, the Wasserstein distance was minimized through training the feature extractor regarding to the parameter θ_f. The whole process can be expressed as follows:

$$\underset{\theta_f}{min}\, \underset{\theta_d}{max} \{\mathcal{L}_{wd} - \lambda \mathcal{L}_{grad}\} \tag{7}$$

where λ is the balancing coefficient, which is equal to 0 when optimizing the minimum operation so that the influence of the gradient penalty on the representation learning process can be avoided. Finally, through adversarial learning, the Wasserstein distance converges to zero, and the domain invariant representations can be learned.

3) Classifier: The classifier was designed to predict the labels of representations learned from the feature extractor. It is composed of two fully connected (FC) layers and a softmax function which can transform the network predictions into class labels. In this work, labels of target features were not used to train the classifier. Instead, the classifier was trained only with labeled MI-EEG data from the source domain. Then, the trained classifier was directly applied to target domain data prediction. The classifier used a cross-entropy loss, which is calculated as:

$$\mathcal{L}_{cls} = -\mathbb{E}_{x \sim D} \sum_{k=1}^{cls} \mathbb{I}_{(y==k)} \log(\mathcal{M}(x)) \tag{8}$$

where $\mathbb{I}$ is the indicator function, if y is equal to k, its result is 1, if not, its result is 0; $\mathcal{M}$ is the proposed model.

Now the final objective function can be expressed as follows:

$$\underset{\theta_f, \theta_c}{min} \left\{ \mathcal{L}_{cls} + \mu \underset{\theta_d}{max} \left[\mathcal{L}_{wd} - \lambda \mathcal{L}_{grad} \right] \right\} \tag{9}$$

where θ_c is the parameter of softmax prediction. μ is the parameter maintaining the balance between the discrimination and transferability of features. λ is zero when optimizing the minimum operation.

The pseudocode of the proposed framework is shown in Algorithm 1. The algorithm of the proposed method was trained by the standard back-propagation. All the parameters were optimized through updating gradient. Firstly, the feature extractor was fixed, with a minibatch containing labeled source data and unlabeled target data. The domain discriminator with the parameter θ_d was updated via gradient ascent, maximizing the empirical Wasserstein distance in (7). Secondly, after optimizing the parameters θ_c and θ_f simultaneously via gradient descent, the classification loss and the maximum Wasserstein distance in (9) were both minimized, so that the feature extractor can be updated. Finally, after the three parameters aforementioned converged, the training process ended and the domain invariant features can be obtained.

E. Comparison Between Our Proposed Model and State-of-the-Art Models

To illustrate the advantages of the proposed method, several state-of-the-art algorithms were chosen for comparison, including traditional method (FBCSP [10]), traditional transfer learning methods (TLCSD [18], RA-MDRM [19], WTLT [20], EA-CSP-LDA [21]), deep learning models (EEGNet [33], ConvNet [14], C2CM [15]), and deep transfer learning models (MI-CNN [22], DRDA [24], DAFS [25]). These algorithms are introduced as follows.

1) FBCSP: A traditional method that performs autonomous selection of frequency bands and corresponding CSP feature.

2) TLCSD: An adaptive transfer learning method which initiates an adaptation based on covariate shift-detection of each subject to update the classifier.

3) RA-MDRM: A transfer learning approach based on Riemannian geometry to affine transform the spatial covariance matrices of different subjects' EEG signals, making the data more comparable.

4) WTLT: A weighted transfer learning method employing KL-divergence to measure similarity between different subjects' feature spaces and assigning weights to the classifier according to the similarity.

5) EA-CSP-LDA: A transfer learning approach aligning EEG trials from different subjects in the Euclidean space.

6) EEGNet: A compact CNN framework designed for EEG signals decoding, containing a 2D convolutional filter, a depthwise convolutional spatial filter, and a second block with a separable convolutional operation.

7) ConvNet: A shallow convolutional neural network designed to decode different band-power features, which contains a temporal convolutional layer, a spatial convolutional layer, a max pooling layer, and three blocks, each of which has a convolutional layer and a max-pooling layer.

8) C2CM: A classification framework for MI-EEG data which employs a temporal representation of the data and a convolutional neural network (CNN) architecture.

9) MI-CNN: A deep transfer learning approach based on CNN, which introduces subject-specific adaptation to improve the performance of a single subject.

10) DRDA: An end-to-end deep domain adaptation method which learns deep feature representations by reducing the marginal and conditional data distribution discrepancy between source and target domains.

11) DAFS: A model integrating deep domain adaptation with few-shot learning to leverage the knowledge from multiple source subjects to improve the classification performance of a single target subject's MI-EEG data.

For both datasets, EEG signals from all electrodes were utilized for classification and the three electrooculography (EOG) channels were discarded without any artifact removing operation. The model was trained with Adam optimizer, the learning rate α was set to 0.0005, and the batch size was set to 64. The programming language adopted in this paper was Python. All the methods were implemented based on the TensorFlow framework.

Algorithm 1 Domain Adaptation Based on Wasserstein Distance

```
Require: source data $D_s$, target data $D_t$, minibatch size $m$, domain discriminator training step $n$, learning rate of domain discriminator $\alpha_1$, learning rate of feature extractor and classifier $\alpha_2$, balancing coefficient $\mu$ and $\lambda$.
Initialize feature extractor, domain discriminator, classifier with random weights $\theta_f$, $\theta_d$ and $\theta_c$.
while $\theta_f$, $\theta_d$ and $\theta_c$ have not converged do
Sample $\{x_i^s, y_i^s\}_{i=1}^{n_s}$, a batch from source data $D_s$
Sample $\{x_j^t\}_{j=1}^{n_t}$, a batch from source data $D_t$
for $i = 1 \ldots n$ do
$h^s \leftarrow f_g(x^s), h^t \leftarrow f_g(x^t)$
Sample $h$ is the random points between $h^s$ and $h^t$ pairs
$\overset{\sim}{h} \leftarrow \{h^s, h^t, h\}$
$\theta_d \leftarrow \theta_d + \alpha_1 \nabla_{\theta_d}[\mathcal{L}_{wd}(x^s, x^t) - \lambda \mathcal{L}_{grad}(\overset{\sim}{h})]$
end for
$\theta_c \leftarrow \theta_c - \alpha_2 \nabla_{\theta_c} \mathcal{L}_{cls}(x^s, y^s)$
$\theta_f \leftarrow \theta_f - \alpha_2 \nabla_{\theta_f}[\mathcal{L}_{cls}(x^s, y^s) + \mathcal{L}_{wd}(x^s, x^t)]$
end while
```

In our experiments, leave-one-out validation was conudcted on each dataset. Specifically, one subject was chosen as the target subject while the remaining subjects were selected as the source subjects. For instance, in Dataset 2a, when we tested our model on Subject A01, the EEG data in A01's second session was set as the target domain, and the data in the remaining eight subjects' first session was merged as the source domain. In Dataset 2b, when we tested our model on Subject B01, the EEG data of B01's last two sessions was set as the target domain, and the data in the remaining eight subjects' first three sessions was merged as the source domain. This was the same for other subjects in BCI Competition IV Dataset 2a and 2b, respectively.

The evaluation results were presented in terms of classification accuracy and Cohen's kappa value, which are two of the most common evaluation matrices. The kappa value can estimate the possibility of generating accidental results, it is calculated as follows:

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{10}$$

where p_0 is the classification accuracy and p_e is the random classification accuracy.

III. Results

A. Comparison of Classification Performances Using the Proposed Method With Baseline Models

The method proposed in this paper was evaluated and compared with other state-of-the-art algorithms on two datasets. In BCI Competition IV dataset 2a, the classification accuracy of each subject (in percentage %), average accuracy (Average Acc), standard deviation (Std), and kappa value were presented in Table II. The highest accuracy and kappa value were highlighted. Our results showed that the performance of our method was superior than other algorithms. Compared with traditional transfer learning methods such as RA-MDRM, EA-CSP-LDA, and WTLT, our method obtained

TABLE II
CLASSIFICATION PERFORMANCE OF DIFFERENT ALGORITHMS ON BCI COMPETITION IV DATASET 2A

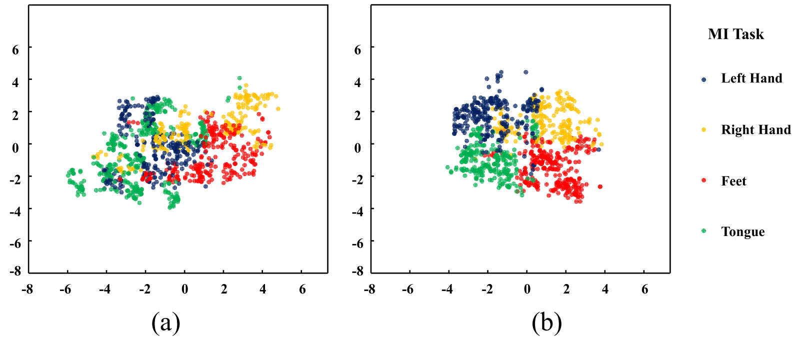
Methods	Subjects									Average Acc ± Std (kappa)	*p*-value
	A01	A02	A03	A04	A05	A06	A07	A08	A09		
FBCSP [10]	76.00	56.50	81.25	61.00	55.00	45.25	82.75	81.25	70.75	67.75 ± 13.73 (0.5700)	0.0003
RA-MDRM [19]	60.20	42.19	66.63	50.97	48.22	51.32	46.75	48.28	66.14	53.49 ± 8.67 (0.4281)	0.0001
EA-CSP-LDA [21]	69.50	40.25	83.01	51.61	38.20	46.58	53.25	68.88	56.12	56.37 ± 14.82 (0.4702)	0.0002
WTLT [20]	76.00	55.20	83.04	60.11	65.79	60.00	73.12	70.83	71.33	68.38 ± 8.87 (0.5471)	0.0010
EEGNet [33]	79.86	58.68	89.93	64.93	63.19	58.68	64.24	73.61	77.08	70.22 ± 10.72 (0.6633)	0.0159
ConvNet [14]	76.39	55.21	89.24	74.65	56.94	54.17	**92.71**	77.08	76.39	72.53 ± 14.24 (0.6338)	0.0107
C2CM [15]	**87.50**	**65.28**	90.28	66.67	62.50	45.49	89.58	83.33	79.51	74.46 ± 15.33 (0.6596)	0.1675
MI-CNN [22]	73.26	28.82	89.58	68.06	26.39	28.82	75.35	78.82	77.08	60.69 ± 25.17 (0.4758)	0.0110
DRDA [24]	83.19	55.14	87.43	75.28	62.29	57.15	86.18	**83.61**	82.00	74.75 ± 12.96 (0.6633)	0.0273
DAFS [25]	81.94	64.58	88.89	73.61	**70.49**	56.60	85.42	79.51	81.60	75.85 ± 10.47 (0.6780)	0.0192
Ours	83.29	63.97	**90.30**	**76.94**	69.34	**60.08**	89.31	82.35	**82.81**	**77.60 ± 10.85** **(0.6951)**	—

TABLE III
CLASSIFICATION PERFORMANCE OF DIFFERENT ALGORITHMS ON BCI COMPETITION IV DATASET 2B

Methods	Subjects									Average Acc ± Std (kappa)	*p*-value
	B01	B02	B03	B04	B05	B06	B07	B08	B09		
FBCSP [10]	70.00	60.36	60.94	**97.50**	93.12	80.63	78.13	92.50	86.88	80.00 ± 13.85 (0.6000)	0.0162
RA-MDRM [19]	73.33	59.17	47.50	85.00	60.00	57.50	54.17	59.17	65.83	62.41 ± 11.08 (0.5102)	0.0002
EA-CSP-LDA [21]	72.50	60.83	53.33	86.67	56.67	57.50	51.67	57.50	65.83	62.50 ± 11.09 (0.5217)	0.0006
TLCSD [18]	70.31	50.63	50.81	93.75	63.75	74.06	61.88	83.13	77.19	69.72 ± 14.37 (0.3944)	0.0006
ConvNet [14]	76.56	50.00	51.56	96.88	93.13	85.31	83.75	91.56	85.62	79.37 ± 17.26 (0.5875)	0.0381
EEGNet [33]	70.31	70.36	78.44	95.33	93.44	82.18	91.88	87.19	71.65	82.37 ± 10.15 (0.6507)	0.3911
MI-CNN [22]	75.31	57.50	56.56	96.88	92.19	83.44	84.06	92.81	86.25	80.56 ± 14.75 (0.6111)	0.0306
DRDA [24]	81.37	62.86	63.63	95.94	93.56	**88.19**	85.00	**95.25**	90.00	83.98 ± 12.67 (0.6796)	0.3321
DAFS [25]	70.31	**73.57**	**80.31**	94.69	**95.00**	83.75	**93.73**	95.00	75.31	84.63 ± 10.20 (0.7325)	0.8246
Ours	**84.66**	66.57	68.04	96.78	94.32	82.61	88.47	93.96	**90.10**	**85.06 ± 11.05** **(0.7403)**	—

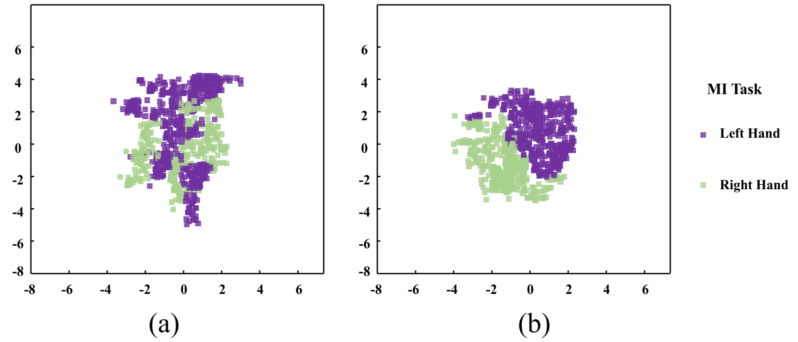
significantly higher average accuracy and kappa value. Compared with deep transfer learning methods such as DRDA and DAFS, our model can improve by around 2% and 0.02 in average accuracy and kappa value, respectively. Meanwhile, a paired t-test showed a significant difference between our method and most compared methods ($p < 0.05$), indicating the superiority of our method. Although the statistical difference between C2CM and our method was not significant ($p = 0.1675$), it should be noted that C2CM did not overcome cross-subject dicrepancy. Besides, in terms of standard deviation, our method was relatively lower than other methods, demonstrating the stability of our model in MI-EEG classification.

The same experiment was conducted on BCI Competition IV dataset 2b, the classification accuracy of each subject (in percentage %), average accuracy (Average Acc), standard deviation (Std), and kappa value were reported in Table III. Compared with other deep learning models and deep transfer learning methods, our method achieved a greater performance with higher average accuracy and kappa value. Although no statistical difference was observed between our proposed method and EEGNet ($p = 0.3911$), DRDA ($p = 0.3321$) as well as DAFS ($p = 0.8246$), compared with the three models, our method still had higher average accuracy and kappa value. Meanwhile, DRDA and DAFS methods required the source MI-EEG data to have



(a) (b)

Fig. 3. t-SNE feature distribution visualization of subject A07. (a) features extracted by the feature extractor used in DRDA. (b) features extracted by the proposed feature extractor.



(a) (b)

Fig. 4. t-SNE feature distribution visualization of subject B08. (a) features extracted by the feature extractor used in DRDA. (b) features extracted by the proposed feature extractor.

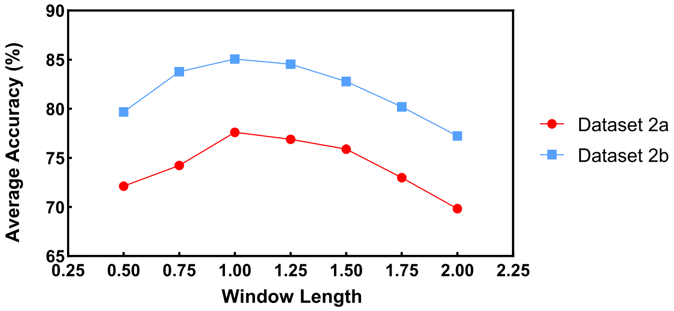


Fig. 5. The average accuracy under different window lengths.

higher signal-to-noise ratio (SNR), which would be more time-consuming and unpractical for real time online BCI application.

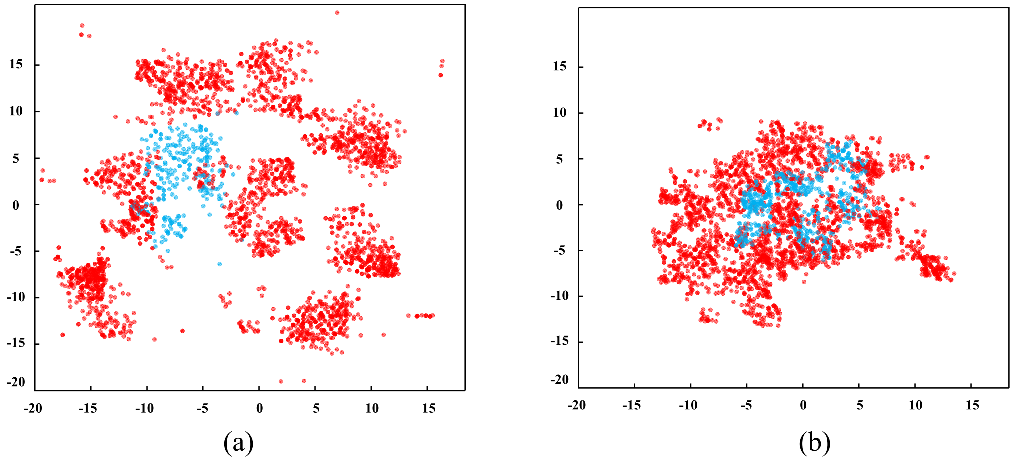
B. Comparison of Feature Distributions Achieved by the Proposed Feature Extractor With Baseline Feature Extractor

The proposed feature extractor aimed at increasing the discrimination of the extracted features to improve the classification results. To demonstrate its effectiveness, the t-SNE method [37] was employed to visualize the extracted features of different motor imagery tasks. Subject A07 and Subject B08 were taken as the examples. The proposed feature extractor was compared with that applied in DRDA, which mainly consisted of a simple spatial convolution layer and a temporal convolution layer. The visualization results of the features extracted by the two feature extractors were presented in Fig. 3 and Fig. 4, respectively. The meaning of x-axis and y-axis is the corresponding values when high-dimensional MI-EEG features are projected to the two-dimensional space through t-SNE. The details of this process were depicted in [37]. In Fig. 3 and Fig. 4, it can be seen that the conventional feature extraction approach used in DRDA cannot clearly group the features of different motor imagery tasks into discriminative clusters (Fig. 3(a) and Fig. 4(a)), whereas the feature extractor proposed in this study can achieve the expected the performance (Fig. 3(b) and Fig. 4(b)), as the boundaries between different motor imagery tasks are much clearer.

C. Effect of Window Size

In the variance layer, the window size plays a crucial role as it affects the extraction of temporal features, determining the quantity and quality of the extracted features. To illustrate its efficacy, the classification accuracies under different window lengths were presented in Fig. 5.

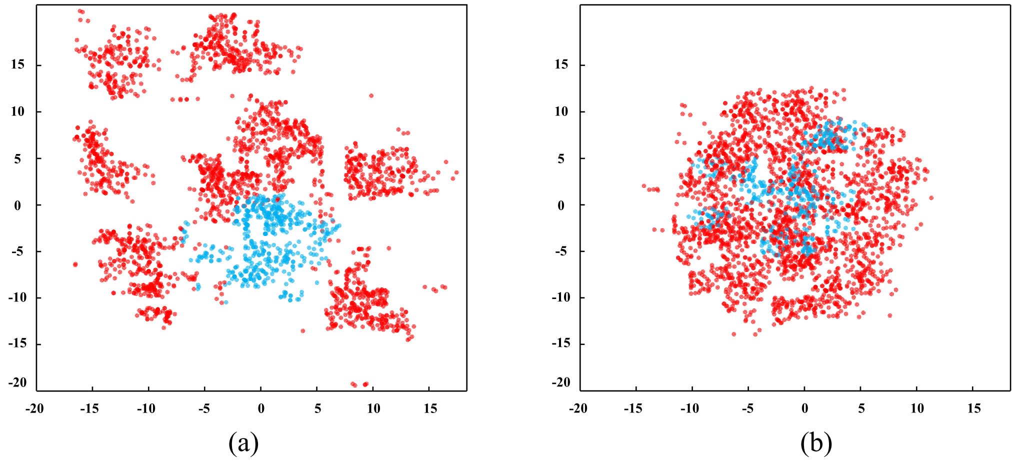
In both datasets, a too large or too small window size would degrade the classification performance. When the window size was set to 1, the classification accuracies were the highest in both datasets (77.60% and 85.06% in BCI IV dataset 2a and 2b, respectively), making it the most suitable value for the window size. However, with the window size increasing, the average accuracies in both datasets decreased. When the window length was set to 2.0, the accuracies for dataset 2a and 2b were only 69.83% and 77.24% respectively, almost 10% lower than the best ones. The same decreasing trend could also be observed when the window length was becoming smaller, as the average accuracies were approximately 5% lower than the best ones in both datasets when the window length equaled to 0.5.



(a)

(b)

Fig. 6. t-SNE feature distribution visualization of subject A03 and other eight subjects in BCI Competition IV dataset 2a. (a) before domain adaptation. (b) after domain adaptation.



(a)

(b)

Fig. 7. t-SNE feature distribution visualization of subject B04 and other eight subjects in BCI Competition IV dataset 2b. (a) before domain adaptation. (b) after domain adaptation.

D. Visualization of Domain Adaptation Based on Wasserstein Distance

To illustrate the effectiveness of Wasserstein distance in domain adaptation, t-SNE method [37] was used to visualize the feature distributions. A03 and other eight subjects in Dataset 2a, as well as B04 and other eight subjects in Dataset 2b before and after domain adaptation were set as examples. The results were presented in Fig. 6 and Fig. 7. In the two images, the red points represent features from source subjects while the blue points represent features from the target subject. It can be observed that before domain adaptation, the features of different subjects were sparse in the space. However, after domain adaptation, the features were grouped into clusters and the distance between different clusters became much shorter.

The features from source domain and target domain overlapped each other, which meant that different subjects shared a similar data distribution in the space.

E. Comparison Between Wasserstein Distance and Adversarial Loss

In this study, Wasserstein distance was utilized to conduct the adversarial training instead of adversarial loss [23]. To show the superiority of Wasserstein distance, we made a comparison between Wasserstein distance and adversarial loss under the same feature extractor and classifier proposed in this work. The results in BCI IV Competition dataset 2a and 2b were exhibited in Fig. 8 and Fig. 9, respectively.

It can be observed from the two figures that under the same condition, in terms of classification accuracy, Wasserstein distance outperformed adversarial loss in most subjects, achieving an average improvement of 2.9% and 2.68% on BCI IV Competition dataset 2a and 2b respectively. For subject A05, the improving margin was 5.46%, the highest in Dataset 2a, as opposed to 5.11% for subject B04, the highest in Dataset 2b.

IV. Discussion

In this study, we developed an improved domain adaptation network based on the Wasserstein distance matrix. The proposed model can increase the discrimination of the extracted features, and reduce the cross-subject discrepancy to make it possible to utilize labeled MI-EEG data of multiple subjects to help classify the data of one target subject, thus solving the EEG data shortage problem and enhancing the MI-EEG data classification results. Specifically, the CBAM and the variance layer were utilized to improve the extraction of spatial and temporal features, respectively. The Wasserstein distance matrix was then applied to implement adversarial training. Our results based on the two public datasets showed that the proposed method was feasible in achieving better classification results under both four-class and two-class situations.

Recently, machine learning and deep learning have been extensively employed to decode EEG-based MI signals, making promising progress in the field of BCI-MI classification. However, most of these previous studies trained and tested their models on the data from the same subject, scarcely considering the situation where the training data and testing data comes from different subjects. In practical scenarios, it is challenging to own the labeled MI data of a new subject for training in advance, only the labeled data of some existing subjects is available. Therefore, due to the lack of new subjects' labeled MI-EEG data and the discrepancy between different subjects' data, the performances of previous models were limited. Even though some latest studies, such as DRDA [24] and DAFS [25], have tried to reduce cross-subject discrepancy via domain adaptation, most of these strategies neglected increasing the discrimination of extracted spatial and temporal features, and focused on minimizing the adversarial loss to realize the domain adaptation, which has some problems such as gradient vanish or model collapse that will degrade the performance of domain adaptation. Therefore, our model was proposed to solve the above limitations.

A. The Role of Feature Extractor

In Fig. 3 and Fig. 4, the features extracted by the proposed feature extractor are more task-discriminative than those extracted by the feature extractor in DRDA. This phenomenon can be explained by the utilization of the attention mechanism CBAM and the variance layer, the former assigning a much more suitable weight to each channel and the latter extracting temporal information based on computing the variance in the given time series, which increases the discrimination of the extracted spatial and temporal features, respectively.

When a person is performing different motor imagery tasks, the power of the mu (8-12 Hz) and beta (16-26 Hz) rhythms varies in the sensorimotor region of the contralateral and ipsilateral hemispheres [38]. According to this principle, among all the electrode channels used in the collection of MI-EEG signals, some are more task-related, but some are less. Therefore, it is of necessity to select channels located in the brain regions that have stronger links to motor imagery tasks and assign them with higher weights. In our work, the attention mechanism CBAM employed the channel attention module to focus on channels having a strong connection with motor imagery tasks, and the spatial attention module to recognize which brain region was active when the subject was performing motor imagery tasks. In this way, different channels were assigned with different weights according to the contribution they made to the subjects' performance of motor imagery tasks. We also drew the brain active correlation map of two subjects from BCI IV dataset 2a to illustrate the importance of CBAM. In Fig. 10, the red color represents positive correlation of the event-related synchronization (ERS), and the blue color indicates negative correlation of the event-related desynchronization (ERD). It is noticeable that when the subject was performing different motor imagery tasks, the active regions of brain varied, some showing ERS whereas some showing ERD. Therefore, it made sense for us to employ CBAM to assign different weights to channels located in different brain regions, which was conducive to increasing the discrimination of spatial features.

As time series, raw EEG signals contain a large quantity of features in temporal domain, along with huge intra-class variance and high noise content. Previous studies have employed various methods to extract temporal information from EEG signals such as the max or average pooling strategy, a fixed-size temporal convolution kernel [15], and the temporal attention mechanism [16]. However, when a subject is performing motor imagery tasks, due to the delay of human brain's reaction and the subject's fatigue or distraction during the collection of MI-EEG signals, the real duration of motor imagery process may last longer or shorter than the experimental requirements. Therefore, the time slices during which the subject performs motor imagery tasks are likely to be irregular. Compared with a fixed convolution kernel or a temporal attention mechanism, a variance layer, which reflects the spectral power in a fixed time series, can extract MI-relevant temporal features in MI-EEG signals more flexibly without missing task-related time slices, thus making the temporal features more discriminative.

Meanwhile, the results in [34] proved that it is necessary to choose the best window size to make each window cover the whole duration of a motor imagery task, a too big or too small window size would degrade the classification performance. The results in Fig. 5 also showed the same trend. This phenomenon can be attributed to the factor that when the window size was too large, each window unit contained too much information and some of it was irrelevant to motor imagery tasks, so that the calculated variance cannot reflect the feature of each task precisely. Conversely, when the window size was too small, each window size contained too little information, which cannot cover the complete duration of a task, thus affecting the integrality and discrimination of the extracted temporal features. According to Fig. 5, the window size equaling to 1 suited the MI-EEG signal best as it contributed to the highest classification accuracy.

With the application of CBAM and the variance layer, in Fig. 3 and Fig. 4, features of different motor imagery tasks were more discriminative than those extracted by the feature extractor in DRDA, which indicated the superiority of the method proposed in this study.

B. The Role of Domain Adaptation Based on Wasserstein Distance

In Fig. 8 and Fig. 9, the model based on Wasserstein distance outperformed that based on adversarial loss in terms of classification accuracy. This phenomenon can be explained by the gradient superiority, better differentiation and continuum of Wasserstein distance [36], which can better adapt

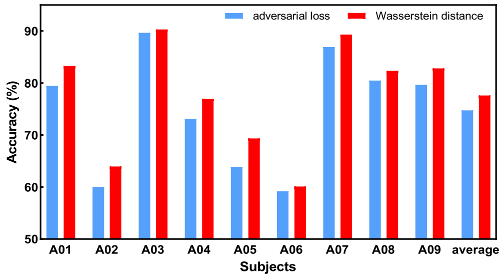


Fig. 8. Comparison between domain discriminator with Wasserstein distance and adversarial loss on Dataset 2a.

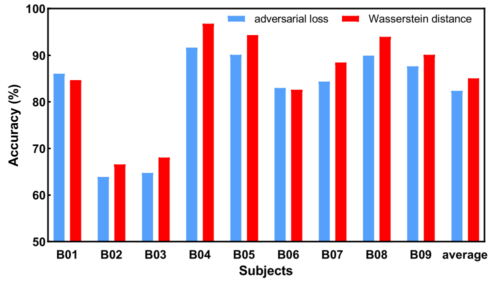


Fig. 9. Comparison between domain discriminator with Wasserstein distance and adversarial loss on Dataset 2b.

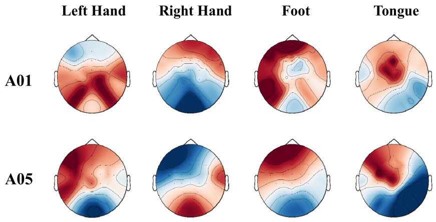


Fig. 10. Brain active correlation maps.

to the nonlinear traits of the extracted MI-EEG features than the adversarial loss, avoiding gradient vanish or model collapse that would degrade the performance of domain adaptation. Therefore, Wasserstein distance has a better ability to reduce data distribution discrepancy across different domains, thus more effectively leveraging knowledge learned from source domain to improve the classification performance on target domain.

C. Limitations and Future Works

Even though the proposed study has provided a novel and improved framework that can outperform other previous methods in classifying EEG-based MI signals, it still has some specific limitations. First, the source subjects were selected without considering the quality of their MI-EEG signals, such as the signal-to-noise ratio (SNR). However, due to some factors such as subjects' distractions, brain sensitivity, or external interferences, the MI-EEG data of some subjects might contain too much noise, which could make the information gained from it useless in improving the classification performance of the target subject's data. Therefore, in our future work, it is necessary to explore some effective methods to measure the quality of each subjects' MI-EEG data and evaluate the transferability, which would be helpful in improving the data alignment and classification performance.

Furthermore, recently, domain generalization (DG) has achieved great success in computer vision field, the goal of which is to learn a model from one or several different but related domains that can generalize well on unseen testing domains [39]. Therefore, in our future work, DG approaches could be utilized in MI-EEG classification, so the model trained on existing annotated data can directly be applied to test new data, omitting the process of aligning data distributions of source and target domain, which would be less time-consuming and more practical in BCI applications.

V. Conclusion

In this study, an improved domain adaptation network based on Wasserstein distance has been proposed to improve the performance of classifying EEG-based MI signals. The proposed framework enhances the performance by increasing the discrimination of MI features and reducing the cross-subject discrepancy, which can solve the problem of EEG data shortage. The CBAM and the variance layer were combined to improve the performance of spatial and temporal feature extraction, and the Wasserstein distance was then applied to implement the adversarial training. Our results demonstrated that the framework proposed in this study was capable of enhancing the classification performance of EEG-based MI signals. This study has provided a novel algorithm to detect the EEG-based MI signals for helping the neural rehabilitation of different neuropsychiatric diseases based on BCI systems.

References

[1] D. Wu, Y. Xu, and B.-L. Lu, "Transfer learning for EEG-based brain–computer interfaces: A review of progress made since 2016," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 1, pp. 4–19, Mar. 2022.

[2] V. Jayaram, M. Alamgir, Y. Altun, B. Schölkopf, and M. Grosse-Wentrup, "Transfer learning in brain–computer interfaces," *IEEE Comput. Intell. Mag.*, vol. 11, no. 1, pp. 20–31, Feb. 2015.

[3] L. Bi, J. Zhang, and J. Lian, "EEG-based adaptive driver-vehicle interface using variational autoencoder and PI-TSVM," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 10, pp. 2025–2033, Oct. 2019.

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

[5] Y. Yang, Z. Gao, Y. Li, Q. Cai, and J. Kurths, "A complex network-based broad learning system for detecting driver fatigue from EEG signals," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 51, no. 9, pp. 5800–5808, Sep. 2021.

[6] Y. Cai, Q. She, J. Ji, Y. Ma, J. Zhang, and Y. Zhang, "Motor imagery EEG decoding using manifold embedded transfer learning," *J. Neurosci. Methods*, vol. 370, Mar. 2022, Art. no. 109489.

[7] M. Lee, Y.-H. Kim, and S.-W. Lee, "Motor impairment in stroke patients is associated with network properties during consecutive motor imagery," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 8, pp. 2604–2615, Aug. 2022.

[8] J. Jin, R. Xiao, I. Daly, Y. Miao, X. Wang, and A. Cichocki, "Internal feature selection method of CSP based on L1-norm and Dempster–Shafer theory," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4814–4825, Nov. 2021.

[9] V. Mishuhina and X. Jiang, "Feature weighting and regularization of common spatial patterns in EEG-based motor imagery BCI," *IEEE Signal Process. Lett.*, vol. 25, no. 6, Jun. 2018, Art. no. 7830787.

[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. IEEE World Congr. Comput. Intell.*, Jun. 2008, pp. 2390–2397.

[11] K. P. Thomas, C. Guan, C. T. Lau, A. P. Vinod, and K. K. Ang, "A new discriminative common spatial pattern method for motor imagery brain–computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 11, pp. 2730–2733, Nov. 2009.

[12] Q. Novi, C. Guan, T. H. Dat, and P. Xue, "Sub-band common spatial pattern (SBCSP) for brain–computer interface," in *Proc. 3rd Int. IEEE/EMBS Conf. Neural Eng.*, May 2007, pp. 204–207.

[13] Y. R. Tabar and U. Halici, "A novel deep learning approach for classification of EEG motor imagery signals," *J. Neural Eng.*, vol. 14, no. 1, Feb. 2017, Art. no. 016003.

[14] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, pp. 5391–5420, Nov. 2017.

[15] S. Sakhavi, C. Guan, and S. Yan, "Learning temporal information for brain–computer interface using convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5619–5629, Nov. 2018.

[16] X. Liu, Y. Shen, J. Liu, J. Yang, P. Xiong, and F. Lin, "Parallel spatial–temporal self-attention CNN-based motor imagery classification for BCI," *Frontiers Neurosci.*, vol. 14, Dec. 2020, Art. no. 587520.

[17] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[18] H. Raza, H. Cecotti, Y. Li, and G. Prasad, "Adaptive learning with covariate shift-detection for motor imagery-based brain–computer interface," *Soft Comput.*, vol. 20, pp. 3085–3096, Aug. 2016.

[19] P. Zanini, M. Congedo, C. Jutten, S. Said, and Y. Berthoumieu, "Transfer learning: A Riemannian geometry framework with applications to brain–computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 5, pp. 1107–1116, May 2018.

[20] A. M. Azab, L. Mihaylova, K. K. Ang, and M. Arvaneh, "Weighted transfer learning for improving motor imagery-based brain–computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 7, pp. 1352–1359, Jul. 2019.

[21] 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.

[22] H. Dose, J. S. Møller, H. K. Iversen, and S. Puthusserypady, "An end-to-end deep learning approach to MI-EEG signal classification for BCIs," *Exp. Syst. Appl.*, vol. 114, pp. 532–542, Dec. 2018.

[23] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2014, pp. 2672–2680.

[24] 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.

[25] C. Phunruangsakao, D. Achanccaray, and M. Hayashibe, "Deep adversarial domain adaptation with few-shot learning for motor-imagery brain–computer interface," *IEEE Access*, vol. 10, pp. 57255–57265, 2022.

[26] 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.

[27] 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.

[28] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 5, pp. 1–46, Sep. 2020.

[29] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 801–814, Apr. 2019.

[30] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 2014, pp. 2204–2212.

[31] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 8–14.

[32] W. Zhang and D. Wu, "Manifold embedded knowledge transfer for brain–computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 5, pp. 1117–1127, May 2020.

[33] 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, Oct. 2018, Art. no. 056013.

[34] 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 *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 2950–2953.

[35] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[36] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. NIPS*, 2017, pp. 5767–5777.

[37] L. Van Der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[38] J.-H. Jeong, N.-S. Kwak, C. Guan, and S.-W. Lee, "Decoding movement-related cortical potentials based on subject-dependent and section-wise spectral filtering," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 3, pp. 687–698, Mar. 2020.

[39] J. Wang et al., "Generalizing to unseen domains: A survey on domain generalization," *IEEE Trans. Knowl. Data Eng.*, early access, May 26, 2022, doi: 10.1109/TKDE.2022.3178128.