A Multi-Source Transfer Joint Matching Method for Inter-Subject Motor Imagery Decoding

Fulin Wei, Xueyuan Xu, Tianyuan Jia, Daoqiang Zhang, *Senior Member, IEEE*, and Xia Wu, *Senior Member, IEEE*

***Abstract*—Individual differences among different subjects pose a great challenge to motor imagery (MI) decoding. Multi-source transfer learning (MSTL) is one of the most promising ways to reduce individual differences, which can utilize rich information and align the data distribution among different subjects. However, most MSTL methods in MI-BCI combine all data in the source subjects into a single mixed domain, which will ignore the effect of important samples and the large differences in multiple source subjects. To address these issues, we introduce transfer joint matching and improve it to multi-source transfer joint matching (MSTJM) and weighted MSTJM (wMSTJM). Different from previous MSTL methods in MI, our methods align the data distribution for each pair of subjects, and then integrate the results by decision fusion. Besides that, we design an inter-subject MI decoding framework to verify the effectiveness of these two MSTL algorithms. It mainly consists of three modules: covariance matrix centroid alignment in the Riemannian space, source selection in the Euclidean space after tangent space mapping to reduce negative transfer and computation overhead, and further distribution alignment by MSTJM or wMSTJM. The superiority of this framework is verified on two common public MI datasets from BCI competition IV. The average classification accuracy of the MSTJM and wMSTJ methods outperformed other state-of-the-art methods by at least 4.24% and 2.62% respectively. It's promising to advance the practical applications of MI-BCI.**

***Index Terms*—Brain–computer interface, inter-subject variability, transfer joint matching, multi-source transfer learning.**

Manuscript received 26 August 2022; revised 5 December 2022; accepted 26 January 2023. Date of publication 8 February 2023; date of current version 15 February 2023. This work was supported in part by the Special Fund for Research on National Major Research Instruments of the Nature Science Foundation of China under Grant 62227801, in part by the Beijing Natural Science Foundation under Grant 4212037, and in part by the China Postdoctoral Science Foundation under Grant 2022M720332. *(Corresponding author: Xia Wu.)*

Fulin Wei, Tianyuan Jia, and Xia Wu are with the School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China, and also with the Engineering Research Center of Intelligent Technology and Educational Application, Ministry of Education, Beijing 100816, China (e-mail: weifulin@mail.bnu.edu.cn; 201921210018@mail.bnu.edu.cn; wuxia@bnu.edu.cn).

Xueyuan Xu is with the Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China (e-mail: xxy@bjut.edu.cn).

Daoqiang Zhang is with the MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: dqzhang@nuaa.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2023.3243257

I. Introduction

Brain-computer interface (BCI) is a direct interactive system between the brain and the outside world that can convert brain signals into control instructions to interact with external devices [1]. Due to the high temporal resolution and good portability, electroencephalogram (EEG) is the most commonly used control signals for BCIs [2]. Motor imagery (MI) is a popular active BCIs paradigm where users image the movements of their body parts but don't execute them. Since MI is a spontaneous brain activity and doesn't require external stimuli, it has been widely studied and applied in stroke rehabilitation and online BCI game fields [2].

Nevertheless, the large inter-subject variability in brain patterns will hinder the widespread use of BCI devices. Because the data distribution change caused by individual differences will significantly degrade the decoding performance of MI-BCI [3], [4], [5], [6]. Typically, a 20-30 minutes system calibration phase needs to be done at the beginning, aiming at acquiring sufficient labeled samples to train a subject-specific model for a new user [7]. It's time-consuming and fatiguing for users. Whereas if the calibration is reduced, the available labeled samples are limited, resulting in poor decoding performance [8]. Thus, developing a reliable inter-subject system that can shorten calibration time and maintain satisfactory performance is highly desirable in the practical application of MI-BCI [9].

One promising way to reduce individual differences and data requirement is transfer learning (TL) [4], [10]. It can transfer shared knowledge across different subjects, and use some existing data to alleviate the limitation of insufficient data of target subjects [4], [11]. For cross-subject transfer, the multiple existing subjects are usually called the source domains, and the current new user with few or without labeled samples is called as the target domain. Several related works have attempted to reduce individual differences by considering the information on different levels of instances, features, and models [5], [6], [12].

In recent years, some studies have proved that data of multi-source subjects can obtain better accuracy than single subject [6], [13], since it can expand the available data and increase data diversity, which would facilitate the transfer model to learn more generalized and robust representations [14], [15]. Thus, some researchers have adopted

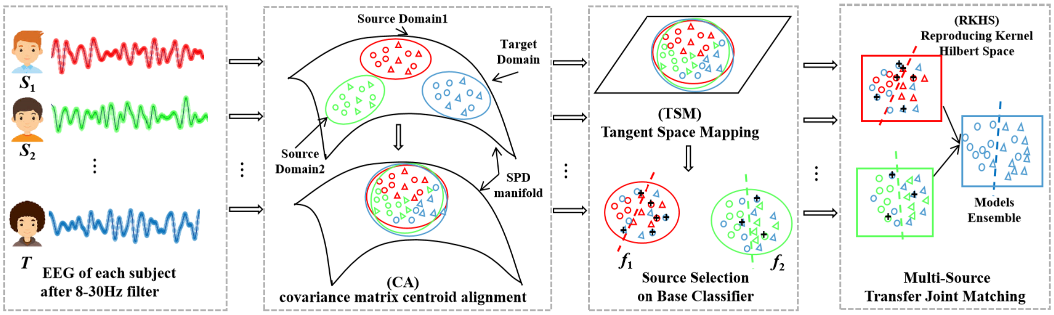


Fig. 1. Overview of inter-subject framework. Triangles and circles represent two classes of data distribution in different spaces, and different colors represent subjects from different domains, where data with plus symbol (+) are task-irrelevant samples that are difficult to classify.

multi-source transfer learning (MSTL) to reduce individual differences in MI-BCI field. Multi-source fusion adaptation regularization (MFAR) is based on Euclidean alignment via rest-state knowledge (EARK) [16]. Cai et al. integrated two Riemannian manifolds into a TL framework to align marginal distribution and joint distribution simultaneously, named manifold embedded TL (METL) [17]. Besides that, Liang et al. proposed a multi-source fusion TL (MFTL) algorithm [18]. Reference [7] presented two TL methods, where multi-source joint domain adaption (MJDA) worked in Riemannian space, while multi-source joint Riemannian adaption (MJRA) worked in Euclidean space.

To sum up, most of these inter-subject transfer works can be summarized into two categories: one is to select the most similar subject as a source domain to assist the target subject task, and the other is to directly combine multiple source domains of other existing subjects into a mixed domain [15]. The former does not make full use of available data and needs to find the optimal golden subject as the source domain [19]. The latter treats each instance and each source domain equally, which ignores the large difference in multiple subjects, a common nature in physiological signals [20], [21]. It would cause negative transfer and degrade the MI decoding accuracies, even result in worse results than using one appropriate subject as the source domain sometimes [13]. In fact, an effective transfer method should assign a large weight to vital instances in the source domain [22]. It is the same true for source domains with more similar distribution [23]. In general, distribution shift exists not only between each pair of source and target domains, but also exists on different source domains, so directly combining multiple domains may influence each other during knowledge transfer [14]. Thus, how to effectively use the rich information from multiple sources is important to MSTL.

To resolve the above problems, this paper propose an inter-subject MI-BCI framework to reduce individual differences, as shown in Fig. 1. The hypothesis is that, despite some differences, the stable and consistent patterns still exist across subjects [24], [25]. A simple yet effective TL method, transfer joint matching (TJM) [22], is introduced into the MI decoding task to address the problems of few labeled trials available for a new subject and inter-subject variability. The framework adopts a semi-supervised domain adaptation (SSDA) setting [26]. It assumes that individual differences lie in multiple levels, including instance and feature levels. TJM jointly performs two TL methods, instance reweighting and feature matching alignment, in a principled dimensionality reduction procedure. It's suitable for EEG data with noises and high-dimensional features [27], [28]. However, it works only in a single source to the target, so we extend it to a multiple sources condition. The main contributions of this paper are as follows.

- We improve TJM to multi-source methods, referred to as MSTJM and weighted MSTJM (wMSTJM). They can effectively utilize the information of multiple subjects to overcome the lack of new subject data and consider the large differences in multiple source subjects.
- We propose an inter-subject MI-BCI framework to reduce inter-subject variability based on MSTJM or wMSTJM. It could reduce the effects of individual differences and task-irrelevant instances. In addition, only ten calibration data are needed to select important source subjects whose data distribution is similar to that of target subjects.
- The superiority of this framework is verified on two public MI datasets of BCI Competition IV. The classification accuracies can achieve 85.53% and 82.69% respectively, superior to most state-of-the-art (SOTA) methods.

The rest of this paper is organized as follows: MI-BCI related works are briefly reviewed in Section II. Then we describe the inter-subject framework based on MSTJM and wMSTJM. Experimental settings and results are shown in Section IV. The following are discussions and conclusion.

II. Related Work

A. Methods on Inter-Subject Variability

To cope with the inter-subject variability and construct a generic MI decoding model, common spatial pattern (CSP) is a commonly used spatial filter method in MI-BCI, and there are many variants, such as regularized CSP [29], and filter bank CSP [30], since the spatial information and frequency band are important for decoding MI. When performing MI, it will typically elicit a decrease in mu and beta rhythms contralateral to the movement [24], [25]. Another commonly used method in recent years is based on deep neural networks, making use of good learning capacity [24], while it requires many labeled training samples. The performance of MI decoding is limited when only a few labeled calibration data can be obtained from the current user at the beginning. Besides that, a series of MI decoding algorithms have been proposed, such as selecting the optimal frequency band or channels [31], [32], and adding physiological information [15]. However, most of them ignore the distribution shift between training and test samples, resulting in a degradation of decoding performance [33]. Thus, it's highly necessary to find a way to achieve high decoding performance with limited calibration data.

B. TL on Inter-Subject Variability With Limited Data

TL can not only bridge the gap in the data distributions among different subjects, but also compensate for the lack of labeled data by leveraging data of other existing subjects [34]. Thus, in recent years, TL has been widely applied in BCI [4], [34]. According to the learning strategies, TL in MI-BCI can be divided into 3 categories, including instance-based, feature-based, and model-based transfer [11], [18].

The first type of instance-based TL usually selects important instances which have more impact on the parameter estimation or weights instances according to the distribution similarity, such as active transfer learning [12]. While it cannot handle task-irrelevant instances or noises when randomly selecting instances [17].

The second feature-based TL maps the features of both domains into a common latent space or feature distribution matching. Some widely used TL methods have been introduced into MI-BCI, such as transfer component analysis (TCA) [35], balanced distribution adaptation (BDA) and weighted BDA (WBDA) [36]. Another widely used method is based on Riemannian manifold, since the congruence invariance property of Riemannian metric can reduce the influence of brain's volume conduction effect [6], [37]. Typical works are the minimum distance to Riemannian mean (MDRM) classifier [37], Riemannian geometry alignment (RGA) [5], and manifold embedded knowledge transfer (MEKT) [6]. Moreover, Wu et al. proposed Euclidean space data alignment (EA) [38], and covariance matrix centroid alignment (CA) [6] to simplify the calculation of geodesic. However, most studies perform TL on a certain level. Long et al. proved that jointly performed feature matching and instance reweighting would be more effective when there are large differences [22].

The last model-based TL focus on sharing parameters. Pre-training model and fine-tuning it on a few labeled instances is one of the most popular methods [24], but [39] suggested that the effect of fine-tuning is not obvious when the distribution shift is large.

Thus, we adopt CA as a preprocessing step, then TJM integrates feature distribution matching and instance reweighting into a unified framework to reduce individual differences.

III. The Proposed Method

Our inter-subject MI-BCI framework based on MSTJM or wMSTJM will be presented in this section. It consists of 3 modules, as illustrated in Fig. 1. In the first modules, we adopt CA as preprocessing to align the centroid of the covariance matrix of source and target trials by taking Euclidean mean as a reference matrix, which can minimize marginal probability distribution shift and whiten EEG into an approximate identity matrix. The second module is source selection after features Tangent Space Mapping (TSM). The source subjects similar to the target subjects are selected by a few calibration data (ten trials). In the third module, MSTJM or wMSTJM is utilized to reduce the data distribution shift among different subjects and the influence of task-irrelevant instances or noise. Then the results of multi-source transfer learning models are fused in the decision-making stage, by majority voting rules or weighted voting rules. The details will be described below.

TABLE I

Notions Summary and Their Descriptions

Notation	Description	Notation	Description
E	electrodes	$\mathbf{X}_i$	EEG trial as input matrix
C	classes label	$\mathbf{P}_i$	sample covariance matrix
Ts	time sample points	$\mathbf{M}_E$	Euclidean mean
$\mathcal{D}_s$, $\mathcal{D}_t$	source/target domain	$\mathbf{A}$	mapping matrix
n_s,n_t	source/target samples	$\mathbf{H}$	the centering matrix
λ	regularization parameter	$\mathbf{K}$	kernel matrix
δ_R	Riemannian distance	$\mathbf{M}$	maximum mean discrepancy
h_i	classifier	$\mathbf{Z}$	new representation of input

A. Problem Description

1) Definition 1 (Domain): Given the collection of labeled source domains $\mathcal{D}_S = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N\}$, and the target domain $\mathcal{D}_T$, each subject from multiple existing subjects is regarded as a source domain, $\{(\mathbf{X}_i, y_i)\}_{i=1}^{n_s}$, where n_s is the number of labeled trials used. $\mathbf{X}_i \in \mathbb{R}^{E \times T_S}$ is *i-th* EEG trial, where E and T_s are the number of electrodes and time points, and $y_i \in \mathbb{R}^C$ is the corresponding label for $\mathbf{X}_i$ of C classes. The target domain is consists of a few labeled samples in $\mathcal{D}_{Tl}$ and many unlabeled samples in $\mathcal{D}_{Tu}$. These few labeled samples in the target domain are also called calibration data in MI-BCI from the new subject. A feature space $\mathcal{X}$ and a marginal probability distribution $P(\mathbf{X})$ form a domain $\mathcal{D}$, i.e., $\mathcal{D} = \{\mathcal{X}, P(\mathbf{X})\}$, where $\mathbf{X} \in \mathcal{X}$.

2) Definition 2 (Task): A task $\mathcal{T}$ consists of C label set $\mathcal{Y}$ and a modle $f(\mathbf{x})$ to learn a relationship after given domain $\mathcal{D}$, i.e., $\mathcal{T} = \{\mathcal{Y}, f(\mathbf{x})\}$, where $y \in \mathcal{Y}$. Note that $f(\mathbf{x}) = Q(y \mid \mathbf{x})$, it also can be explained from the perspective of conditional probability distribution. In our experiments, only binary classification problem is considered, i.e., $C \in \{-1, +1\}$. The goal of TL is to predict $y_t \in \mathcal{Y}_t$ using data in D_S and D_{Tl} by the constructed task-specific classifier $\mathcal{H} : \mathbf{X}_t \rightarrow y_{\mathrm{t}}$.

3) Problem Setting: Under SSDA setting, besides a labeled source domain $\mathcal{D}_i = \left\{(\mathbf{x}_1, y_1), \ldots, \left(\mathbf{x}_{n_s}, y_{n_s}\right)\right\}$, there are a small number of labeled calibration data and many unlabeled instances in the target domain, i.e., $\mathcal{D}_T = \mathcal{D}_{Tl} \cup \mathcal{D}_{Tu}$, where $\mathcal{D}_{Tl} = \left\{\mathbf{x}_{n_s+1}, \ldots, \mathbf{x}_{n_s+n_l}\right\}$, and $\mathcal{D}_{Tu} = \left\{\left(\mathbf{x}_{n_s+n_l+1}, y_{n_s+n_l+1}\right), \ldots, \left(\mathbf{x}_{n_s+n_l+n_u}, y_{n_s+n_l+n_u}\right)\right\}$. The aim is to reduce the inter-subject variability in a new feature representation space. All notions are summarized in Table I.

B. Covariance Matrix Centroid Alignment

In order to make use of the congruence invariance property of Riemannian metric while reducing computing overhead, we adopt CA as preprocessing and Euclidean mean as a reference matrix to minimize marginal probability distribution shift among different subjects [6].

For *i-th* EEG trial in the source or target domains, the sample covariance matrix (SCM) is

$$\mathbf{P}_i = \frac{1}{T_s - 1}\mathbf{X}_i\mathbf{X}_i^T, \tag{1}$$

where T_s is time points of each trial and $\mathbf{P}_i \in \mathbb{R}^{E \times E}$. Since the SCMs are SPD matrices, thus that each of them can be regarded as a point in Riemannian manifold space $\mathcal{M}$ [5], [37]. Then the Riemannian distance of two point P_1 and P_2 is: $\delta_R(\mathbf{P}_1, \mathbf{P}_2) = \left\| \log\left(\mathbf{P}_1^{-1}\mathbf{P}_2\right)\right\|_F$, where log is the logarithm for the eigenvalues of $\mathbf{P}_1^{-1}\mathbf{P}_2$.

Euclidean mean $\mathbf{M}_E$, i.e., the arithmetic mean, is used to calculate distribution distance. The Euclidean distance of two points is $\delta_E(\mathbf{P}_1, \mathbf{P}_2) = \|\mathbf{P}_1 - \mathbf{P}_2\|_F$ on $\mathcal{M}$. Then the Euclidean mean of SPD matrices can be defined by

$$\mathbf{M}_E = \underset{\mathbf{P}\in\mathbf{P}(n)}{\arg\min}\sum_{i=1}^{I}\delta_E^2(\mathbf{P}, \mathbf{P}_i) = \frac{1}{I}\sum_{i=1}^{I}\mathbf{P}_i. \tag{2}$$

With Euclidean mean $\mathbf{M}_E$, the SCMs are aligned by

$$\mathbf{P}_i' = \mathbf{M_E}^{-1/2}\mathbf{P}_i\mathbf{M_E}^{-1/2}. \tag{3}$$

It's the same to the target domain samples, so we can obtain aligned SCMs, $\{\mathbf{P}_{s_i}'\}_{i=1}^{n_s}$, $\{\mathbf{P}_{t_i}'\}_{i=1}^{n_l}$, and $\left\{\mathbf{P}_{t_j}'\right\}_{j=1}^{n_u}$.

Two desirable properties of CA can be used to align the data distribution [6]: 1) Minimization of marginal probability distribution shift. 2)EEG trial whitening.

According to the nature property of Riemannian manifold, including congruence invariance, we have

$$\delta_R\left(\mathbf{P}_1^{-1}, \mathbf{P}_2^{-1}\right) = \delta_R(\mathbf{P}_1, \mathbf{P}_2),$$
$$\delta_R\left(\mathbf{W}^T\mathbf{P}_1\mathbf{W}, \mathbf{W}^T\mathbf{P}_2\mathbf{W}\right) = \delta_R(\mathbf{P}_1, \mathbf{P}_2) \quad \forall\mathbf{W} \in Gl(n), \tag{4}$$

with $Gl(n) = \{\mathbf{W} \in \mathcal{M}\}$ the set of invertible matrices. These properties are very important in MI-BCI, because it means that the distance between two SPD matrices is invariant to a change of reference matrix [37]. So we can perform some operations, such as PCA, on this space without affecting the distance.

When the reference $\mathbf{M}_{ref} = \mathbf{M_E}^{-1/2}$ is adopted,

$$\delta_R\left(\mathbf{M}_{\text{ref}}^{\top}\mathbf{P}_1\mathbf{M}_{\text{ref}}, \ldots, \mathbf{M}_{\text{ref}}^{\top}\mathbf{P}_{n_s}\mathbf{M}_{\text{ref}}\right)$$
$$= \mathbf{M}_{\text{ref}}^{\top}\delta_R(\mathbf{P}_1, \ldots, \mathbf{P}_{n_s})\mathbf{M}_{\text{ref}} = \mathbf{M}_{\text{ref}}^{\top}\mathbf{M}_E\mathbf{M}_{\text{ref}} = I. \tag{5}$$

The arithmetic centers of different domains will be approximated as an identity matrix. Thus, the data distribution of different subjects are brought closer, and the aligned SCMs of EEG trials after CA is equivalent to whitening on the manifold.

C. Source Selection After Tangent Space Mapping

Following the CA is Tangent Space Mapping (TSM), which can transform the operation on Riemannian manifold into a Euclidean tangent space. A SPD matrix $\mathbf{P}_i$ is in manifold space, while by using TSM, it will be converted to a vector $\mathbf{x}_i \in \mathbb{R}^{d \times 1}$ in tangent space, where $d = \frac{E \times (E+1)}{2}$. TSM can be denoted as

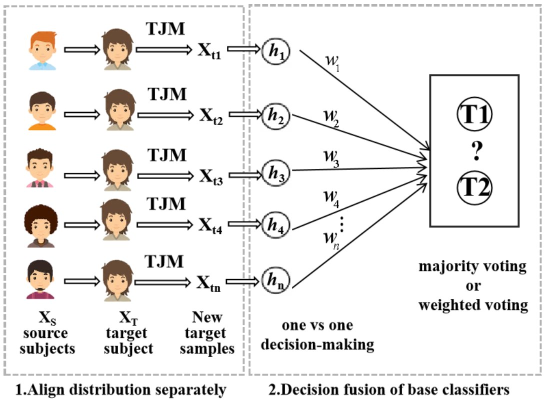


Fig. 2. The architecture of (weighted) multi-source transfer joint matching. $\{h_1, h_2, \ldots, h_n\}$ are base classifiers from different source subjects, and $\{w_1, w_2, \ldots, w_n\}$ are the corresponding weights.

$$\mathbf{x}_{S,i} = \text{upper}\left(\log_M\left(P_{S,i}'\right)\right), \quad i = 1, \ldots, n_s$$
$$\mathbf{x}_{T,i} = \text{upper}\left(\log_M\left(P_{T,i}'\right)\right), \quad i = 1, \ldots, n_l$$
$$\mathbf{x}_{T,j} = \text{upper}\left(\log_M\left(P_{T,j}'\right)\right), \quad j = 1, \ldots, n_u. \tag{6}$$

Here, upper(.) operator represents taking the upper triangular part of an SPD matrix while vectorizing it. The unity weights are applied to diagonal elements and $\sqrt{2}$ are assigned to off-diagonal elements [37].

The tangent space consists of a set of tangent vectors at a point $\mathbf{P}_i$ on Riemannian manifold. When two conditions are met: 1) $\mathbf{P}_i$ is embedded in the local space of manifold; 2) $\mathbf{P}$ is the mean of the $\mathbf{P}_i$, the distance in tangent space is approximately equal to that in Riemannian manifold space. Further details can be found in [37] and [40].

Then, source subjects similar to the target subject are selected to avoid negative TL and reduce the computation overhead. After TSM, the source subjects selection can be done in Euclidean space with accuracy. The SVM is trained on the TSM features $\mathbf{x}_S$ of each source subject to get base classifier $h : \mathbf{x}_S \rightarrow y_s$. The vectorized calibration data $\mathbf{x}_{T,j}$ are utilized to evaluate the similarity directly by the classification accuracies. We hold that if the model has higher accuracy, the corresponding subject in the source domain may be more similar to the subject in the target domain, so the data of them can be mixed as training samples to train the classification model of target domain when there are limited data.

D. Multi-Source Transfer Joint Matching

The third module is the MSTJM or wMSTJM approaches for domain adaptation, which combines both MSTL and decision fusion into a uniform framework, as shown in Fig. 2. It's composed of two main steps: 1) distribution alignment separately according to each selected subject in the source domain; 2) decision fusion to integrate multiple results of base classifiers. The difference between MSTJM and wMSTJM is whether the decision fusion process considers the different weights. The scheme for distribution alignment before fusion instead of integrating before alignment, can consider the difference between source subject and target subject, but also the difference in different subjects in the source domains.

1) Distribution Alignment Separately: TJM is introduced to handle the difference in two levels, including samples-based and feature-based differences, which transfers knowledge by jointly implementing instance reweighting and feature distribution matching to construct domain-invariant features in the subspace [22]. For adaptive instance reweighting, structured sparsity penalty, $\ell_{2,1}$-norm is performed on instances in the source domains. The features are mapped into a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ by "kernel trick", and then feature distribution is matched by minimizing the maximum mean discrepancy (MMD) between the source domain and target domain. Principal Component Analysis (PCA) is used to combine these two operations since the reconstruction error of the input data can be minimized to learn a new feature representation in this dimensionality reduction process.

Given the number of samples $n = n_s + n_l + n_u$, and the feature dimension m, the input data matrix can be denoted as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$. Through kernel mapping $\psi(\mathbf{X}) = [\psi(\mathbf{x}_1), \ldots, \psi(\mathbf{x}_n)]$, we can calculate the kernel matrix by $\mathbf{K} = \psi(\mathbf{X})^{\mathrm{T}}\psi(\mathbf{X}) \in \mathbb{R}^{n \times n}$. The kernelize PCA can be denoted as

$$\max_{\mathbf{A}^{\mathrm{T}}\mathbf{A}=\mathbf{I}} \operatorname{tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{K}\mathbf{H}\mathbf{K}^{\mathrm{T}}\mathbf{A}\right), \tag{7}$$

where $\mathbf{A} \in \mathbb{R}^{n \times k}$ is the mapping matrix to realize kernelize PCA. The new representation is embedded in $\mathbf{Z} = \mathbf{A}^{\mathrm{T}}\mathbf{K}$.

MMD is adopted to measure the distance between Kernel-PCA representations for comparing different distributions in the RKHS. MMD matrix $\mathbf{M}$ can be computed by

$$\mathbf{M}_{ij} = \begin{cases} \dfrac{1}{(n_s+n_l)(n_s+n_l)}, & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_s \cup \mathcal{D}_{Tl} \\ \dfrac{1}{n_u n_u}, & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_{Tu} \\ \dfrac{-1}{(n_s+n_l)\, n_u}, & \text{otherwise.} \end{cases} \tag{8}$$

Then MMD between source and target domain is

$$\left\| \frac{1}{n_s+n_l} \sum_{i=1}^{n_s+n_l} \mathbf{A}^{\mathrm{T}}\mathbf{k}_i - \frac{1}{n_u} \sum_{j=n_s+n_l+1}^{n_s+n_l+n_u} \mathbf{A}^{\mathrm{T}}\mathbf{k}_j \right\|_{\mathcal{H}}^2 = \operatorname{tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{K}\mathbf{M}\mathbf{K}^{\mathrm{T}}\mathbf{A}\right). \tag{9}$$

Eq. (7) is maximized by minimizing Eq. (9), such that statistics of feature distributions in the first- and high-order are matched under the new representation $\mathbf{Z} = \mathbf{A}^{\mathrm{T}}K$.

Although feature matching can partly minimize the distribution shift among subjects, there are also differences caused by non-stationary EEG and task-irrelevant noises. The instance reweighting method is used to further deal with these differences. The structured sparsity regularizer, $\ell_{2,1}$-norm is applied to the transformed matrix $\mathbf{A}$. By introducing row-sparsity to each instance, so the instances are reweighted as follows:

$$\|\mathbf{A}_s\|_{2,1} + \|\mathbf{A}_u\|_F^2 \tag{10}$$

where $\mathbf{A}_s := \mathbf{A}_{1:(n_s+n_l),:}$ is the transformed matrix for source instances, and $\mathbf{A}_u := \mathbf{A}_{n_s+n_l+1:n_s+n_l+n_u,:}$ is the transformed matrix for the target instances.

By integrating Eq. (7) and Eq. (10), the optimization problem can be expressed as:

$$\begin{aligned} &\min_{\mathbf{A}} \operatorname{tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{K}\mathbf{M}\mathbf{K}^{T}\mathbf{A}\right) + \lambda\left(\|\mathbf{A}_s\|_{2,1} + \|\mathbf{A}_u\|_F^2\right) \\ &\text{s.t. } \mathbf{A}^{T}\mathbf{K}\mathbf{H}\mathbf{K}^{T}\mathbf{A} = \mathbf{I}, \end{aligned} \tag{11}$$

where λ is the balance parameter to weigh the importance of feature matching and instance reweighting. By solving the transformation matrix in the new representation $\mathbf{Z} = \mathbf{A}^{\mathrm{T}}\mathbf{K}$, the discrepancy between different subjects can be reduced.

2) Decision Fusion of Base Classifiers:

After TJM, the test data of target subject is transformed into a new subspace, and the distribution of them is relatively consistent with the corresponding single subject in the source domains. Then the key issue is how to integrate the new represents of test data from multiple TL models. We adopt parallel architecture to construct base classifiers of each subject in the source domains, and then the final decisions of new test samples from these base learners are fused by majority voting or weighted voting, corresponding to MSTJM and wMSTJM.

For MSTJM, suppose labeled source domains $\mathcal{D}_S = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N\}$, and binary classification problem, i.e., $C \in \{-1, +1\}$. The goal is to learn a more robust and accurate classifier $\mathcal{H} = \{h_1, h_2, \ldots, h_N\}$ by integrating multiple base classifiers $h : \mathbf{z}_S \to y_s$ to predict y_t using data in D_S and D_{Tl}. The base classifiers are constructed by SVM classifiers, using data of each subject in the source domains. Given a sample x and base classifier h_i, the prediction of each classifier is $\left(h_i^1(x);\, h_i^2(x);\, \ldots h_i^N(x)\right)$. Here, $h_i^j(x)$ represents the output of h_i on class label C_j. Then the output label is obtained by the most voted class through the majority voting method by Eq. (12), which could achieve a good performance [41].

$$H(x) = C_{\operatorname{argmax}} \sum_{i=1}^{N} h_i^j(x) \tag{12}$$

For wMSTJM, weighted voting is utilized to ensemble multiple outputs by considering the unequal similarity of different subjects. After constructing base models, a few labeled calibration data $\mathbf{x}_{T,j}$ in the target domain D_{Tl} are used to measure differences between each pair of source subject and target subject. We intuitively believe that the higher the accuracy, the more similar the two subjects are. Thus, higher weights are endowed to the corresponding subjects, and the integration weights are learned adaptively for each target subject to cope with subject variations. In general, the voting weights should be normalized. The final output is the label with the highest voting class through the weighted voting method by Eq. (13). If the voting scores of two classes are the same, randomly select one as the final label.

$$H(x) = C_{\operatorname{argmax}} \sum_{i=1}^{N} w_i h_i^j(x), \qquad \text{s.t. } w_i \geqslant 0, \sum_{i=1}^{N} w_i = 1. \tag{13}$$

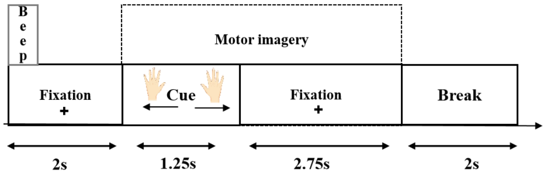


Fig. 3. Recording paradigms of motor imagery.

IV. Experiments and Results

A. Datasets and Preprocessing

Two common public MI datasets were used in the experiments, BCI competition IV[1] *dataset 1* and *dataset 2a*, as [6], [16]. Recording paradigms of these two datasets are similar, displayed in Fig. 3. The duration of MI execution is about 4 seconds starting from 2 seconds.

1) Dataset1: The Berlin BCI group [42] provided 59-channel EEGs, sampled at 1000Hz. Each subject randomly performed 2 classes of MI tasks (left hand, right hand, foot), a total of seven subjects. Only labeled calibration MI data from left and right hands were used in this paper. Each subject had 200 trials, 100 trials for left hand or right hand.

2) Dataset2a: The Graz University of Technology [43] provided 22-channel EEGs, sampled at 250Hz. Each subject performed 4 classes of MI tasks, including left hand, right hand, feet, tongue MI tasks. Nine subjects participated in the experiment, and each of them performed 288 trials in each session. For consistency as [6] and [16] and labeled samples, we only used training session trials to verify our MSTLs. Each subject had 144 trials, 72 trials for left hand or right hand.

For both datasets, a 8-30Hz bandpass filter was used to remove artifacts and help to analyze the characteristics of Mu and Beta rhythms [6], [24], [25]. The time interval of each trial EEG data was segmented between [2.5, 5.5] seconds.

B. Experimental Settings

To estimate our subject-independent MI-BCI framework based on MSTJM and wMSTJM, we implemented the leave one-subject-out cross validation (LOSOCV) paradigm on multi-to-single (MTS) transfer tasks as [17] and [6]. Here, we took each subject as the current target domain to test the model in turn, and other existing subjects as multiple source domains mixed a few calibration trials of a target subject to train the base models. For dataset1, there were seven MTS tasks, and dataset2a included nine MTS tasks. The mean value of the classification accuracy of each target subject was used as the evaluation metric.

$$\text{Accuracy} = \frac{|\mathbf{x} : \mathbf{x} \in \mathcal{D}_t \wedge \widehat{y}(\mathbf{x}) = y(\mathbf{x})|}{|\mathbf{x} : \mathbf{x} \in \mathcal{D}_t|} \tag{14}$$

where $\mathcal{D}_t$ is the test data in the target domain, $y(\mathbf{x})$ is the truth label of $\mathbf{x}$, and $\widehat{y}(\mathbf{x})$ is the predicted label.

C. Baseline Algorithms

The proposed MSTJM and wMSTJM algorithms were compared with many algorithms, including classical CSP, several SOTA Riemannian manifold relevant methods, and MSTL algorithms for MI decoding. Some common used TL methods were also considered into our experiments, such as TCA, WBDA, JDA, CORAL, JGSA and GFK, combined with different preprocessing steps, such as EA, EARK or CA.

- CSP-LDA, a typical decoding method in Euclidean space.
- EA-CSP-LDA. EA is Euclidean alignment by task-state knowledge as a preprocessing step [38].
- EARK-WBDA (balanced distribution adaptation) [36]. EARK is Euclidean alignment by rest-state knowledge.
- EARK-TCA (transfer component analysis), minimizing MMD in new RKHS [35].
- CA-CSP-LDA, centroid alignment in Tangent space [6].
- CA-CORAL (correlation alignment), covariance matrices matching by minimizing the Frobenius norm [44].
- CA-JGSA (joint geometrical and statistical alignment), considering shared and domain specific features [45].
- CA-GFK (geodesic flow kernel), domain shifts in a Grassmann manifold by integrating subspaces [46].
- CA-JDA (joint distribution adaptation), adapts the marginal distribution and conditional distribution [47].
- RGA-MDRM, a typical Riemannian space method [5].
- MFAR (multi-source fusion adaptation regularization), combined WBDA, source empirical risk, and manifold regularization [16].
- METL (manifold embedded transfer learning), using geometric properties in Riemannian manifold and JDA [17].
- S-STM (supervised style transfer mapping) [48], a similar MSTL method with two different prototypes, nearest prototype and Gaussian model.
- Semi-STM (semisupervised style transfer mapping) [48], using both labeled calibration data and unlabeled data in the target domain to learn STM.
- MEKT (manifold embedded knowledge transfer) [6]. The reference matrices of MEKT-R, MEKT-E, and MEKT-L are the Riemannian mean, Euclidean mean, and Log-Euclidean mean, respectively.

D. Parameters Details

For CSP, we used three pairs of CSP variances to form 6 features as [49]. The parameters were set as original papers for other methods. In our MSTJM and wMSTJM, the number of calibration data was set to 10 depending on the aim of our study and experience, which will be discussed later. The weights for wMSTJM directly came from the sorted accuracies of calibration data to adapt to the subject variants. The regularization parameter and subspace bases were $\lambda = 0.01$, $k = 300$ for dataset1, and $\lambda = 0.01$, $k = 200$ for dataset2a. In addition, the number of iterations and kernel type were fixed as $T = 10$ and 'RBF'. Note that the parameters were the same in both algorithms. In addition, since data distribution between the source domain and target domain were different for different subjects, tuning optimal parameters by cross-validation is not realistic, so we empirically searched the parameter space to obtain the optimal parameters as [13], [22]. λ searched in the range of $\lambda \in \{0.01, 0.1, 1\}$, and k searched in the range of $k \in \{20, 50, 100, 200, 300\}$. When the highest

[1] https://www.bbci.de/competition/

TABLE II
MEAN ACCURACY AND STANDARD DEVIATION (%) OF DATASET1 AND DATASET2A FOR OUR wMSTJM AND MSTJM, COMPARED WITH OTHER SOTA METHODS

Index	Algorithm	dataset1	dataset2a	Avg
1	CSP-LDA	58.71 ± 12.93	67.75 ± 12.92	63.73
2	EA-CSP-LDA [38]	79.79 ± 6.57	73.53 ± 15.96	76.66
3	EARK-WBDA [36]	65.29 ± 7.87	63.43 ± 11.14	64.51
4	EARK-TCA [35]	64.93 ± 7.61	64.89 ± 11.40	64.91
5	CA-CSP-LDA [6]	76.29 ± 9.66	71.84 ± 13.89	74.07
6	CA-CORAL [44]	78.86 ± 8.73	72.38 ± 13.38	75.62
7	CA-JGSA [45]	76.79 ± 12.35	73.07 ± 16.33	74.93
8	CA-GFK [46]	76.79 ± 12.57	72.99 ± 15.82	74.89
9	CA-JDA [47]	81.07 ± 11.19	74.15 ± 15.77	77.61
10	RGA-MDRM [5]	73.29 ± 9.25	72.07 ± 9.88	72.68
11	MFAR [16]	78.57 ± 7.07	75.08 ± 13.35	76.83
12	METL [17]	83.14 ± 7.29	76.00 ± 16.14	79.57
13	S-STM-proto [48]	81.67 ± 7.81	77.45 ± 11.89	79.56
14	S-STM-Gauss [48]	81.29 ± 7.67	77.53 ± 12.00	79.41
15	Semi-STM-proto [48]	79.37 ± 15.42	79.95 ± 11.17	79.66
16	Semi-STM-Gauss [48]	79.62 ± 10.99	79.86 ± 11.61	79.74
17	MEKT-E [6]	81.29 ± 10.18	76.00 ± 17.61	78.65
18	MEKT-L [6]	83.07 ± 9.30	76.54 ± 16.72	79.81
19	MEKT-R [6]	83.42 ± 9.55	76.31 ± 16.76	79.87
20	wMSTJM (Ours)	84.03 ± 11.69	80.95 ± 11.17	82.49
21	MSTJM (Ours)	**85.53 ± 10.80**	**82.69 ± 10.03**	**84.11**

TABLE III
THE MEAN ACCURACY (%) OF DATASET1 WHEN EACH SUBJECT IS TAKEN AS THE TARGET DOMAIN IN TURN

Subject	S1	S2	S3	S4	S5	S6	S7	Mean ± Std
wMSTJM	87.43	73.16	81.58	98.95	68.59	98.42	80.10	**84.03 ± 11.69**
MSTJM	89.53	78.95	83.16	98.42	68.59	98.42	81.68	**85.53 ± 10.80**

average accuracy achieved on each dataset, the optimal parameters were fix up, then the experiment repeated ten times to avoid randomness.

E. Results of MSTJM and wMSTJM

The average accuracies and standard deviation of different target subject in turn are illustrated in Table II. Bold indicates the optimal result and underlined indicates the suboptimal result. The results of MSTJM and wMSTJM achieved 85.53 ± 10.80% and 84.03 ± 11.69% in dataset1, and those of dataset2a can achieve 82.69 ± 10.03% and 80.95 ± 11.17% respectively. It's clear that our algorithms outperform all the other SOTA algorithms in both datasets. The average accuracy of the MSTJM and wMSTJ methods in two data sets were 4.24% and 2.62% higher than the suboptimal result. Compared with MEKT [6], which has similar preprocessing steps, our MSTJM and wMSTJM could obtain better results, especially in dataset2a. Meanwhile, CA-JDA [47], S-STM [48] and METL with JDA [17] showed good results on dataset1. S-STM and semi-STM had suboptimal results in dataset2a, which also regarded each subject as a source domain to reduce the difference in the source domains. The main difference between them lies in how to transfer knowledge in different subjects. All results demonstrated the necessity of considering the differences in the source subjects, and the effectiveness of MSTJM on EEG data with large differences.

To further interpret the results, Table III and Table IV list the details of each subject when they were taken as the target domain. 'S1-S7' represents 7 subjects in dataset1. The results are consistent with [5] and [16]. There are obvious individual differences in dataset1 and dataset2a, where the results of S2, S5 in dataset1 and those of S2, S4, S5 in dataset2a are relatively worse than other subjects. These subjects who are not proficient in MI are called 'Bad subject' [5] or 'BCI illiteracy' [3], since they are unable to control the BCI equipment well. Although many other subjects' data were used for knowledge transfer, the results of our MSTL methods were relatively limited. Thus, it is necessary to develop more advanced TL methods to overcome individual differences.

TABLE IV
THE MEAN ACCURACY (%) OF DATASET2A WHEN EACH SUBJECT IS TAKEN AS THE TARGET DOMAIN IN TURN

Subject	S1	S2	S3	S4	S5	S6	S7	S8	S9	Mean ± Std
wMSTJM	91.79	74.07	96.27	76.12	63.43	79.10	72.39	94.78	80.60	**80.95 ± 11.17**
MSTJM	92.54	74.81	97.76	78.36	67.91	79.10	78.36	94.78	80.60	**82.69 ± 10.03**

TABLE V
SIX EXPERIMENTAL SETTINGS AND MEAN ACCURACIES(%) OF ABLATION EXPERIMENTS

scheme	source domains	weighting	TL	Dataset1	Dataset2a
C1	source-combined	×	×	70.57	67.21
C2	multi-source	×	×	71.16	70.36
C3	multi-source	✓	×	72.49	71.55
C4	source-combined	×	✓	73.44	76.26
C5	multi-source	×	✓	**85.53**	**82.69**
C6	multi-source	✓	✓	84.03	80.95

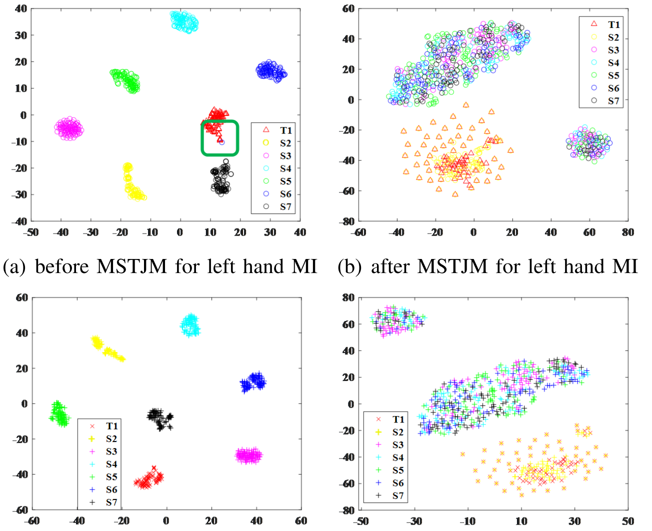
V. DISCUSSIONS

A. Ablation Experiments

To figure out which part of the framework works, we did ablation experiments. The six experimental settings are shown in Table V, abbreviated as **C1-C6**. The source-combined scheme means that all data in the multiple source domains were combined into a single mixed source. While the multi-source scheme considers each existing subject as an independent source domain to transfer knowledge individually. Note that the scheme **C5** and **C6** correspond to MSTJM and wMSTJM.

The mean accuracies of each target in ablation experiment are shown in Table V. Obviously, the results of **C5** and **C6** were much better than those in other cases, especially those without TL. Compared with **C1** and **C2**, **C4** and **C5**, transferring each source domain individually gave better results than combining all source domains. This is consistent with our hypothesis that the data distribution of source subjects is different, thus each source need to process separately.

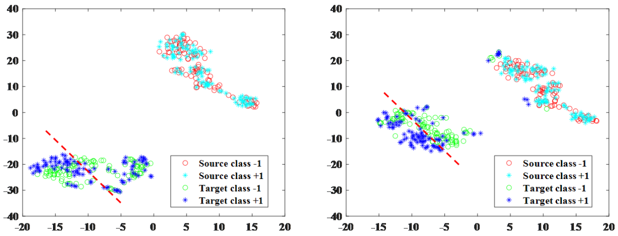
In addition, compared with **C2** and **C3** in the setting without TL, weighted voting could improve the decoding performance than majority voting. However, compared with **C5** and **C6** in the setting with TL, the mean accuracy of the weighted scheme in both datasets were worse than those of majority voting. Further analyzing the results of Table III in dataset1, the result of 'S5' in the wMSTJM was far worse than MSTJM, about 5%. It's the same for dataset2a in Table IV, especially for 'S5' and 'S7'. The weighted scheme is not always effective against those 'Bad subjects'. This is because the weights are directly from the standardized accuracy of 10 calibration



(a) before MSTJM for left hand MI (b) after MSTJM for left hand MI

(c) before MSTJM for right hand MI (d) after MSTJM for right hand MI

Fig. 4. The feature distribution visualization of dataset1 for different MI tasks when subject1 was used as target domain and others as source domains. Different colors represent features from different subjects, while different shapes represent subjects in different domains. The blue circle of S6 is a noise point, marked with a green box.



(a) before MSTJM (b) after MSTJM

Fig. 5. The feature distribution visualization of dataset1 when data of subject 1 were used as target domain and subject 2 as source domains.

data of target subject, while the data distribution of them is significantly different. Moreover, compared with **C1** and **C2** or **C3**, the multi-source scheme was superior to the source-combined scheme. It's the same to **C4**, **C5** and **C6**.

To sum up, the effectiveness of MSTJM and wMSTJM were verified by these ablation experiments. The performance improvement stems not only from considering the differences between the source and target domains, but also from considering the differences between different source domains. However, when the data distribution is quite significant, this weighting method will fail. It is worth noting that Eq. (13) holds only when the outputs of the base models are independent of each other [41]. But in our experiments, the base models established for each subject seem to be relatively independent, but all of them are used to decode the same MI decoding problem, which means that there is a strong correlation between the outputs of these base models, so they do not meet the independence hypothesis. Therefore, in practical MI application, the results from weighted voting cannot be guaranteed to be superior to that of majority voting method. A more adaptive and better weighting method is needed for MI-BCI.

B. Visualization of Feature Distribution

To further explain the effect of MSTJM and wMSTJM on distribution, we did feature distribution visualization based on t-SNE (t-distributed stochastic neighbor embedding) [50]. The data distribution of dataset1 is intuitively illustrated Fig. 4, when data of subject1 were used as a target. The SCM features were vectorized by TSM to form a 1 × 1770 feature, then it was reduced to 300 after MSTJM. The dimension of all features was further reduced to two by t-SNE.

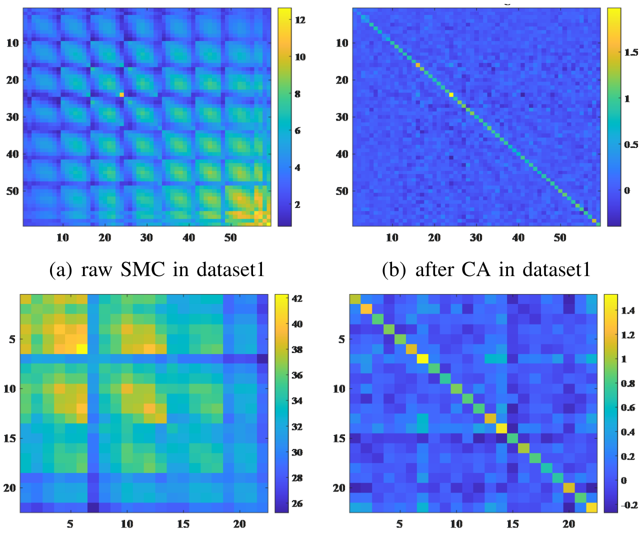
In Fig.4(a), the data spatial distribution of left hand MI for each subject is completely independent. There is an obvious noise point, marked with a green box, which is away from the cluster center for target subject1, abbreviated as T1. After MSTJM, the effect of noise point is not obvious, and the distribution of different subjects is relatively reduced. The data of T1 and S2 are almost overlapped, which indicates that after MSTJM, the distribution of the two subjects' data is relatively close. We also checked the corresponding weights of calibration data for S2, and it showed the highest accuracy. Three clusters after MSTJM illustrate that only some data are similar to the target data, so it's necessary to select source domains and data. It's the same for the MI tasks of right hand.

Moreover, the data distribution of two classes for only two subjects is also exhibited in Fig. 5, when data of subject1 were used as target domain and subject2 as source domain. The data distribution of the subjects in the source domain and the target domain is reduced after MSTJM, which means the individual difference is reduced. In addition, despite some mismatching data points, the data from different classes in the target domain can be separated more easily which is conducive to classification. For a better classification performance, it is not only necessary to reduce the intra-class distance, but also to increase the inter-class distance [51]. Thus, MSTJM and wMSTJM can further advance classification.

C. Effect of Centroid Alignment

We also used the *imagesc* to visualize the SCM before and after CA, as illustrated in Fig. 6. As explained in Section III, when using Euclidean mean as the reference matrix, the aligned SCM after CA is approximate to the identity matrix. The results in Fig. 6 verified this property of CA. Here, we took the first EEG trial of subject2 in dataset1 and dataset2a as examples. The left and right columns represent the raw SCM and the aligned SCM, respectively. It is obvious that the diagonal elements of SCM are close to 1, while off-diagonal values are around 0, thus that EEG whitening was approximately achieved by CA. This property of CA has been proven before. The marginal probability distribution shift of EEG trials will be minimized simultaneously for each subject.

Note that CA is used as a preprocessing step, which is similar to other Riemannian manifold methods, such as RGA [5], EA [38], EARK [16] and so on. Moreover, we can also directly match the covariance by calculating the distance between two matrices [52]. The effectiveness of Riemannian manifold in the practice of BCI has been verified in the above researches and many other works. It may become a standard paradigm for EEG data preprocessing in the future.



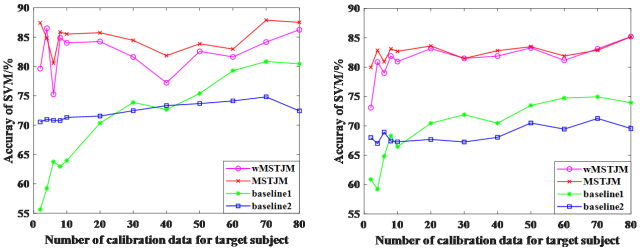
(a) raw SMC in dataset1

(b) after CA in dataset1

(c) raw SMC in dataset2a

(d) after CA in dataset2a

Fig. 6. The SCM before and after CA when taking trial 1 of subject 2 as an example, where Euclidean mean is reference matrix.



(a) Transfer in dataset1

(b) Transfer in dataset2a

Fig. 7. Accuracy of two baseline methods without TL and MSTMJ, wMSTMJ with TL in dataset1 and datase2a.

D. Effect of Knowledge Transfer With Limited Data

To further analyze the effect of subject transfer, and confirm our hypothesis that TL can facilitate MI decoding with limited data of the target subject, we compared our MSTLs with two basic methods without TL.

- Baseline 1, SVM was trained using only the labeled calibration data of target subject, and then the model was tested using the unlabeled data of target subject.
- Baseline 2, the training samples were mixed data from calibration data of target subject and labeled data of all existing subjects in the source domains.

To simulate real-world scenarios when a new subject uses BCI equipment initially, the number of calibration data was set to 2, 4, 6, 8 and 10 in sequence. We also verified the effectiveness of TL simultaneously, when the number of calibration data increased from 20 to 80 and the step length increased by 10.

Fig. 7 shows that our MSTLs on both datasets are better than those without TL, no matter how much labeled data of the target subjects were taken. Thus, in order to obtain better decoding performance with a few calibration data, we only used 10 instances samples of the target subject for calibration. Moreover, the result of baseline2 is better than those of baseline 1 when the labeled calibration data is insufficient. However, with labeled data increase, the results of baseline1 will exceed. This phenomenon is consistent with the results of [18] and [16].

This is because of the large differences among subjects. With the increase of target subject's data, a reliable model can be trained directly by his own data. Most traditional machine learning algorithms, such as SVM, hold the assumption of independent identically distributed (i.i.d.) [11]. When it comes to the non-i.i.d. data from different subjects, the individual variability should be reduced by TL methods. It indicates that there is inter-subject variability among different subjects, and our MSTLs can not only reduce individual differences, but also can maintain a relatively stable result with limited data.

VI. Conclusion and Future Work

This paper proposed two MSTL algorithms, MSTJM and wMSTJM, to minimize individual differences under the SSDA setting, which reduces the differences at two levels of instance and feature. They first align the data distribution of each pair of subjects in the source domain and target domain, and then fuse the results of multiple TL models in the decision-making stage. They not only consider the differences between subjects in the source domains and the target domain, but also the differences among subjects in the source domains. We also construct an inter-subject BCI framework for MI decoding. This framework consists of three modules: CA, source selection, and data distribution alignment by MSTJM or wMSTJM. The results of two public MI datasets demonstrate the superiority of MSTJM and wMSTJM over other methods.

In the future, we will consider further optimizing our MSLTs from the following aspects: (1) The limited MI decoding performance may be improved by other deep TL methods [53], due to the stronger representation ability when using a larger MI dataset. (2) Domain generalization techniques could be considered in MI decoding to avoid using labeled data from the target subject. (3) In order to effectively utilize the rich information of multi-source domains, an adaptive and efficient source domain selection method should be adopted for different target subjects in practice.

References

[1] X. Gao, Y. Wang, X. Chen, and S. Gao, "Interface, interaction, and intelligence in generalized brain–computer interfaces," *Trends Cogn. Sci.*, vol. 25, no. 8, pp. 671–684, 2021.

[2] M. A. Khan, R. Das, H. K. Iversen, and S. Puthusserypady, "Review on motor imagery based BCI systems for upper limb post-stroke neurorehabilitation: From designing to application," *Comput. Biol. Med.*, vol. 123, Aug. 2020, Art. no. 103843.

[3] M.-H. Lee et al., "EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy," *GigaScience*, vol. 8, no. 5, pp. 1–16, May 2019.

[4] 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.

[5] 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 2017.

[6] 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.

[7] F. Wang, J. Ping, Z. Xu, and J. Bi, "Classification of motor imagery using multisource joint transfer learning," *Rev. Sci. Instrum.*, vol. 92, no. 9, Sep. 2021, Art. no. 094106.

[8] 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.

[9] P. Gaur, R. B. Pachori, H. Wang, and G. Prasad, "A multi-class EEG-based BCI classification using multivariate empirical mode decomposition based filtering and Riemannian geometry," *Exp. Syst. Appl.*, vol. 95, pp. 201–211, Nov. 2018.

[10] A. M. Azab, J. Toth, L. S. Mihaylova, and M. Arvaneh, "A review on transfer learning approaches in brain–computer interface," in *Signal Processing and Machine Learning for Brain-Machine Interfaces*. London, U.K.: Institution of Engineering and Technology, Sep. 2018, pp. 81–98, doi: 10.1049/PBCE114E.

[11] 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.

[12] I. Hossain, A. Khosravi, and S. Nahavandhi, "Active transfer learning and selective instance transfer with active learning for motor imagery based BCI," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 4048–4055.

[13] Y. Zhou et al., "Cross-task cognitive workload recognition based on EEG and domain adaptation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 50–60, 2022.

[14] S. Zhao et al., "Multi-source distilling domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 12975–12983.

[15] F. Lotte, "Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces," *Proc. IEEE*, vol. 103, no. 6, pp. 871–890, Jun. 2015.

[16] L. Zhu et al., "Multi-source fusion domain adaptation using resting-state knowledge for motor imagery classification tasks," *IEEE Sensors J.*, vol. 21, no. 19, pp. 21772–21781, Oct. 2021.

[17] 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.

[18] Y. Liang and Y. Ma, "Calibrating EEG features in motor imagery classification tasks with a small amount of current data using multisource fusion transfer learning," *Biomed. Signal Process. Control*, vol. 62, Sep. 2020, Art. no. 102101.

[19] B. Sun, Z. Wu, Y. Hu, and T. Li, "Golden subject is everyone: A subject transfer neural network for motor imagery-based brain computer interfaces," *Neural Netw.*, vol. 151, pp. 111–120, Jul. 2022.

[20] J. Cheng, F. Wei, C. Li, Y. Liu, A. Liu, and X. Chen, "Position-independent gesture recognition using sEMG signals via canonical correlation analysis," *Comput. Biol. Med.*, vol. 103, pp. 44–54, Dec. 2018.

[21] X. Li et al., "Digital health: Tracking physiomes and activity using wearable biosensors reveals useful health-related information," *PLOS Biol.*, vol. 15, no. 1, Jan. 2017, Art. no. e2001402.

[22] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer joint matching for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2014, pp. 1410–1417.

[23] H. Zhao, S. Zhang, G. Wu, J. M. Moura, J. P. Costeira, and G. J. Gordon, "Adversarial multiple source domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–12.

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

[25] G. Pfurtscheller and F. L. Da Silva, "Event-related EEG/MEG synchronization and desynchronization: Basic principles," *Clin. Neurophysiol.*, vol. 110, no. 11, pp. 1842–1857, 1999.

[26] A. Singh et al., "Improving semi-supervised domain adaptation using effective target selection and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2709–2718.

[27] X. Xu, X. Wu, F. Wei, W. Zhong, and F. Nie, "A general framework for feature selection under orthogonal regression with global redundancy minimization," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 11, pp. 5056–5069, Nov. 2022.

[28] X. Xu, T. Jia, Q. Li, F. Wei, L. Ye, and X. Wu, "EEG feature selection via global redundancy minimization for emotion recognition," *IEEE Trans. Affect. Comput.*, early access, Mar. 24, 2021, doi: 10.1109/TAFFC.2021.3068496.

[29] S.-H. Park, D. Lee, and S.-G. Lee, "Filter bank regularized common spatial pattern ensemble for small sample motor imagery classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 498–505, Feb. 2018.

[30] K. Keng Ang, Z. Yang 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.

[31] Y. Yang, S. Chevallier, J. Wiart, and I. Bloch, "Subject-specific time-frequency selection for multi-class motor imagery-based BCIs using few Laplacian EEG channels," *Biomed. Signal Process. Control*, vol. 38, pp. 302–311, Sep. 2017.

[32] J. Fumanal-Idocin, Y.-K. Wang, C.-T. Lin, J. Fernández, J. A. Sanz, and H. Bustince, "Motor-imagery-based brain–computer interface using signal derivation and aggregation functions," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 1–12, May 2021.

[33] 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, no. 8, pp. 3085–3096, Aug. 2016.

[34] 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.

[35] 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.

[36] J. Wang, Y. Chen, S. Hao, W. Feng, and Z. Shen, "Balanced distribution adaptation for transfer learning," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 1129–1134.

[37] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain–computer interface classification by Riemannian geometry," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 4, pp. 920–928, Apr. 2012.

[38] 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.

[39] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," in *Proc. 10th Int. Conf. Learn. Represent. (ICLR)*, Apr. 2022, pp. 1–54.

[40] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Classification of covariance matrices using a Riemannian-based kernel for BCI applications," *Neurocomputing*, vol. 112, pp. 172–178, Jul. 2013.

[41] Z.-H. Zhou, "Ensemble learning," in *Machine Learning*. Singapore: Springer, 2021, pp. 181–210.

[42] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio, "The non-invasive Berlin brain–computer interface: Fast acquisition of effective performance in untrained subjects," *NeuroImage*, vol. 37, no. 2, pp. 539–550, 2007.

[43] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "BCI competition 2008—Graz data set A," *Institute for Knowledge Discovery* (Laboratory of Brain-Computer Interfaces), vol. 16. Graz, Austria: Graz University of Technology, 2008, pp. 1–6.

[44] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 30, no. 1, 2016, pp. 1–8.

[45] J. Zhang, W. Li, and P. Ogunbona, "Joint geometrical and statistical alignment for visual domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1859–1867.

[46] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2066–2073.

[47] 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.

[48] J. Li, S. Qiu, Y.-Y. Shen, C.-L. Liu, and H. He, "Multisource transfer learning for cross-subject EEG emotion recognition," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3281–3293, Mar. 2019.

[49] C. Brunner, M. Naeem, R. Leeb, B. Graimann, and G. Pfurtscheller, "Spatial filtering and selection of optimized components in four class motor imagery EEG data using independent components analysis," *Pattern Recognit. Lett.*, vol. 28, no. 8, pp. 957–964, 2007.

[50] L. Van Der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, p. 2579—2605, 2008.

[51] S. Li, S. Song, G. Huang, Z. Ding, and C. Wu, "Domain invariant and class discriminative feature learning for visual domain adaptation," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4260–4273, Sep. 2018.

[52] L. Li and Z. Zhang, "Semi-supervised domain adaptation by covariance matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2724–2739, Nov. 2018.

[53] Y. Wang, S. Qiu, X. Ma, and H. He, "A prototype-based SPD matrix network for domain adaptation EEG emotion recognition," *Pattern Recognit.*, vol. 110, Feb. 2021, Art. no. 107626.