Test-Time Adaptation for Cross-Subject Motor Imagery EEG Classification Using Information-Aggregation and Source-Guided Weighting

Yiheng Peng[1], *Jingjing Luo*[1*], *Hongbo Wang*[1], *Shijie Guo*[1], *Yuzhu Guo*[2], *Dongsheng Xu*[3], *Yang Li*[2]

[1]Academy for Engineering and Technology, Fudan University
[2]Department of Automation Sciences and Electrical Engineering, Beihang University
[3]Shanghai University of Traditional Chinese Medicine

2025 International Joint Conference on Neural Networks (IJCNN) | 979-8-3315-1042-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/IJCNN64981.2025.11228030

***Abstract*—Individual-specific calibration is a major bottleneck in motor imagery (MI) electroencephalogram (EEG) decoding, limiting real-world neural-feedback rehabilitation. Transfer learning, particularly Test-Time Adaptation (TTA), offers a promising solution for direct online cross-subject adaptation, handling sequentially arriving unlabeled MI-EEG data. However, existing TTA methods, primarily designed for domains such as computer vision, face challenges when applied to MI-EEG data due to its scarcity and non-stationary nature. To address the challenges in direct online MI-EEG decoding, this paper proposes MI-IASW, a novel framework combining Information-Aggregation (IA) and Source-Guided Pseudo-Label Weighting (SW). IA leverages Mixed and Adaptive Batch Normalization (MABN) to ensure effective aggregation of statistical and gradient information. Additionally, IA adopts a Weight Aggregation (WA) strategy to improve generalization under limited data. Meanwhile, SW first evaluates the overconfident pseudo-labels with the guidance of source centers and then employs Class-Aware Weighting (CAW) to adjust sample contributions to the loss function. Experimental evaluations on two public MI-EEG datasets demonstrate that our proposed framework outperforms various competitive baselines, achieving an average performance gain of 3.17% over the baseline TTA methods and 6.80% over the source model. By eliminating the need for individual-specific offline calibration, MI-IASW enables practical deployment in real-world rehabilitation and improves cross-subject decoding.**

***Index Terms*—Brain-Computer Interfaces, Motor Imagery, Test-Time Adaptation.**

I. Introduction

Brain-Computer Interfaces (BCIs) establish seamless and efficient interaction between the human brain and devices [1], [2]. BCIs often leverage electroencephalography (EEG) to decode movement intentions, particularly in motor imagery (MI) paradigms, where users imagine specific movements to trigger corresponding neural activities. MI-EEG decoding can be applied to various healthcare scenarios, including wheelchair control, prosthetic limb manipulation, and post-stroke rehabilitation [3], [4]. However, MI-EEG decoding is hindered by inter-subject variability, leading to the network performing well on one individual but poorly on others. Consequently, additional offline trials for model calibration are necessary for each new subject before model deployment. This process can be both time-consuming and burdensome.

Although Unsupervised Domain Adaptation [5], [6], [7] and Source-Free Domain Adaptation [8], [9], [10] can reduce the impact of inter-subject variability, they still require a substantial amount of offline target data for adaptation. This necessitates extensive calibration trials, which can be especially laborious for individuals with physical disabilities. To facilitate direct online rehabilitation without offline calibration, we delve into the challenging online MI-EEG Test-Time Adaptation (TTA) task [11], [12], where a model trained with labeled source subjects is gradually adapted to the target (test) subject using online unlabeled data from the subject, while making inferences concurrently.

TTA algorithms for Computer Vision (CV) have gained remarkable success in recent years. TBN substitutes the test-time Batch-Normalization (BN) [13] statistics for those from the source domain while PBN [14] mixes both source and target statistics. [15] updates the BN affine-parameters with entropy minimization loss. PL usually refers to a variant of [15] that adopts hard pseudo-labels proposed in [16]. Eta [17] filters samples based on their prediction entropy. CoTTA [18] tackles the continually changing target domain with the guidance from a teacher model. RMT [19] extends CoTTA with strategies like symmetric entropy loss [20]. ROID [21] proposes a source-target weight-averaging approach, along with sample selection and prior correction strategies to enhance diversity. Versa [22] adopts pseudo-label calibration and weight regularization methods to facilitate continual adaptation. [23] performs periodic restoration to enable long-term adaptation. [24]–[29] address the challenge of time-correlated data, where inner-batch class imbalance occurs.

However, research on MI-EEG TTA, such as [11], is still limited. We argue that several potential challenges in MI-EEG TTA need to be addressed: 1) **Limited Observable Data**: This limitation first stems from the inherent difficulty of MI-EEG data collection. As performing motor imagery requires subjects

*Corresponding author.

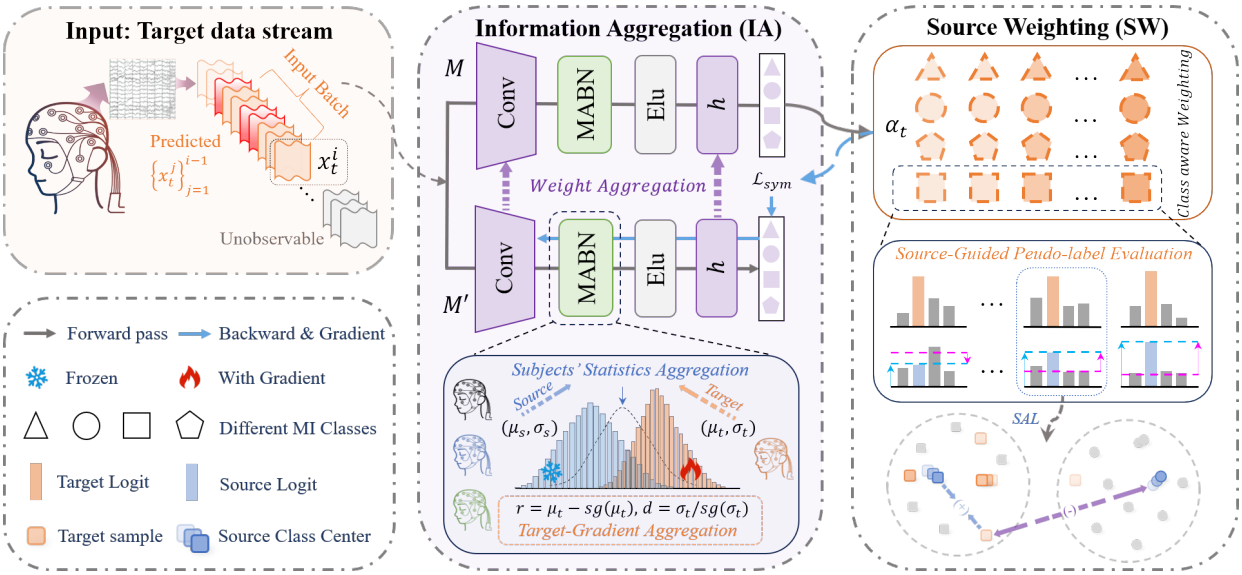

Fig. 1. Illustration of our proposed TTA framework. MI-EEG signals from the target subject arrive sequentially and we update the model when a new observable batch accumulates. The student model $\mathcal{M}'$ is guided by the teacher $\mathcal{M}$ and SAL based CAW. Both models are equipped with MABN.

to sustain intense concentration, the number of MI trials is usually smaller than that of CV datasets. Furthermore, the need for direct online inference in real-world MI paradigms requires immediate adaptation and inference based on only a limited amount of MI-EEG data from the target subject. Accordingly, fully leveraging accessible information becomes essential to improve the performance of MI-EEG TTA; 2) **Unreliable Pseudo-Labels**: In MI-EEG scenarios, inter-subject and intra-subject variations result in noisy pseudo-labels, making the careful evaluation and selection of these pseudo-labels essential for effective network adaptation. Existing CV-TTA methods frequently depend on the target model's own predictions, such as prediction entropy or maximum softmax probability, to evaluate the quality of its pseudo-labels [17], [21], [22]. However, this approach is less effective in MI-EEG scenarios, as the target model tends to produce even more overconfident predictions for the non-stationary and low signal-to-noise ratio data compared to its CV counterpart [30], [31]. Specifically, the target MI-EEG network generates low-entropy, high-confidence evaluations for numerous pseudo-labels easily, which reduces its ability to evaluate pseudo-labels. An analysis of the cause of this problem is given in Section II-D and this phenomenon is visualized in Section III-F. As the target model is less reliable in pseudo-label evaluation, incorporating complementary guidance is essential for accurately mitigating the overconfident pseudo-labels.

To address the dual challenges of direct online MI-EEG decoding, we propose MI-IASW, a comprehensive framework that integrates Information-Aggregation (IA) and Source-Guided Pseudo-Label Weighting (abbreviated as Source-Guided Weighting or SW). The IA component includes Mixed and Adaptive Batch Normalization (MABN), which aggregates source-subject statistics while enabling the network to incorporate target information through effective gradient updates. Additionally, a Weight Aggregation (WA) strategy is employed to improve generalization performance, particularly in MI-EEG scenarios with limited data availability.

The SW component tackles the second challenge by evaluating potentially overconfident target predictions under the guidance of credible source centers (Source-Guided Pseudo-label Evaluation). Specifically, it evaluates the extent to which the source-center decision supports the target prediction, i.e., the source alignment level (SAL). Based on this evaluation, the contribution of each sample to the loss function is recalculated using Class-Aware Weighting (CAW). The proposed MI-IASW framework is illustrated in Figure 1. MI-IASW carefully addresses the challenges in MI-EEG TTA and demonstrates improvements over the competing methods on this challenging task. ***Therefore, MI-IASW has the potential to enable calibration-free network adaptation to new subjects, reducing the burden for participants, especially those with physical disabilities, to use BCI-based rehabilitation.***

In summary, our main contributions include:

1) We propose an online MI-EEG TTA framework called MI-IASW, which tackles the problems brought by scarcity and non-stationary nature in cross-subject online MI-EEG scenarios. Our framework aligns closely with real-world MI scenarios and holds the potential to relieve subjects from extra offline calibration experiments.
2) MI-IASW utilizes the Information-Aggregation strategies, i.e., MABN and Weight Aggregation, to enhance the performance of online MI-EEG TTA, a challenging task characterized by data scarcity.
3) MI-IASW utilizes Source-Guided Pseudo-label Evaluation and Class-Aware Weighting to tackle the overconfident pseudo-labels.

II. METHOD

A. Preliminaries

Consider a C-class MI-EEG classification transfer-learning task, where we have K labeled source subjects $\{D_s^k\}_{k=1}^K$ and one unlabeled target subject D_t. For the k-th source subject, there are m_k MI-EEG trials $D_s^k = \{(x_s^i, y_s^i)\}_{i=1}^{m_k}$, where x_s^i represents the signal and $y^i \in 1, 2, \cdots, C$ the ground-truth label. The target subject has n unlabeled trials, each arriving one by one in an online manner.

Apart from the data, we formulate a deep neural network as $\mathcal{M} = h \circ f$, where $\mathcal{M}$ is parameterized by θ. Here, f denotes the feature encoder, which maps the input signal x^i to a high-level feature embedding. Meanwhile, h represents the linear classifier that associates the extracted feature representations with class-specific templates stored in its parameters, ultimately producing class-specific logits $z^i = \{z_c^i\}_{c=1}^C$. Each logit z_c^i reflects the alignment degree between x^i and class c. Usually, the softmax function is employed to explore the relative relationships among $\{z_c^i\}_{c=1}^C$ and convert the logits into class-distribution probabilities $\{p_c^i\}_{c=1}^C$. The final classification result $\hat{y}^i$ is obtained by $\arg\max_c \{p_c^i\}_{c=1}^C$.

The source MI-EEG model needs to be trained before adaptation. In line with the 'Leaving One Subject Out' (LOSO) strategy, which is commonly used for inter-subject settings, all the source-subject data is combined to create a unified source domain $D_s = \{D_s^1, D_s^2, \cdots, D_s^K\}$. Based on the labeled dataset D_s, the source model weights W_s are trained using the cross-entropy loss in a supervised manner.

B. Test-Time Setting and the Proposed Framework

We then describe the test-time setting and our proposed framework here. During the test stage, the target model is initialized with θ_s while D_s is discarded for online computational efficiency. Our goal is to harness the useful information contained in the source model, adapt the model to the target domain using the unlabeled target data streams, and classify each target sample simultaneously. Specifically, the classification of sample x_t^i should not be influenced by any of the subsequent samples $\{x_t^j\}_{j=i+1}^n$ [32]. In this paper, whenever a mini-batch of size b is required, the previous $b-1$ samples and the current sample x_t^i are used to form the batch. Note that as long as the network has observed samples from the target subject, it is referred to as the target model.

Our proposed MI-IASW framework integrates IA and SW to tackle the dual challenges of data scarcity and MI-EEG pseudo-label evaluation. We detail IA, which consists of Mixed and Adaptive BN and Weight Aggregation, and SW, which consists of Source-Guided Pseudo-label Evaluation and Class-Aware Weighting, in the following two sections.

C. Information Aggregation

In this section, we provide a comprehensive overview of the Information Aggregation (IA) approach in MI-IASW, which aggregates available information from different perspectives to tackle the data scarcity problem. We believe that BN layers deserve special attention because they encapsulate domain-specific information and play a crucial role in ensuring the training stability of MI-EEG networks. Thus, we first propose the Mixed and Adaptive Batch-Normalization (MABN), which constitutes a component of IA.

During the test stage, TTA methods designed for CV typically discard the source statistics (μ_s, σ_s) of the BN layers and normalize network activations $\boldsymbol{a}$ using test batch statistics (μ_t, σ_t) to address domain shift [13] [15] [18]. However, we argue that retaining both source and target statistical information benefits MI-EEG TTA training with limited data, as source statistics carry rich transferable MI-EEG activation information calculated across different source subjects, while target statistics are better aligned with the target subject. Therefore, MABN implements a simple yet effective approach to recalculate the BN statistics by averaging the source statistics and target batch statistics during adaptation:

$$\tilde{\mu} = sg(avg(\mu_s, \mu_t)) + r,\ r = \mu_t - sg(\mu_t), \tag{1}$$

$$\tilde{\sigma} = sg(avg(\sigma_s, \sigma_t)) \cdot d,\ d = \frac{\sigma_t}{sg(\sigma_t)}, \tag{2}$$

where sg terms (stop gradient operation like $torch.detach$) ensure that the combined source and target statistics only affect the numerical values of $\tilde{\mu}$ and $\tilde{\sigma}$. Meanwhile, the terms r and d preserve gradients from μ_t and σ_t [33], enabling adaptive updates to the network through the BN layers. This operation ensures the integration of target-subject information through gradient updates while preserving the benefits of source statistics.

One of the key contributions of MABN is identifying that the general idea of mixing source and target BN statistics is beneficial for cross-subject MI-EEG scenarios. While we do not further explore more complicated methods such as varying mixing extents across channels and layers like some CV-TTA papers [34], [35], so that the proposed MABN remains straightforward to implement and is particularly suited for online MI-EEG decoding without a validation set. To sum up, MABN simultaneously enhances source information usage through statistical aggregation and target-information aggregation by ensuring target gradient updates.

We then detail the second part of IA: Weight Aggregation (WA). In online MI-EEG scenarios, models typically draw inferences after observing only a portion of the target data ($\{x_t^j\}_{j=1}^{i-1}$, requiring efficient use of the scarce and potentially noisy previous data without overfitting. Thus, we introduce the Weight-Aggregation Model, referred to as the Exponential Moving Average teacher (EMA-teacher), which consolidates knowledge of the target data from previous training steps and produces a noise-robust ensemble model [18], [36], [37]. During test time, the student $\mathcal{M}'$ and teacher $\mathcal{M}$ are first initialized with the source model ($\theta' \leftarrow \theta_s, \theta \leftarrow \theta_s$). The symmetric entropy loss [19], [20] is then combined with soft weighting α_t^i, which is elaborated upon in the following section, to

facilitate the student model's learning of the aggregated weight information:

$$\mathcal{L}_{\text{sym}} = \sum_i \frac{\alpha_t^i}{2|\mathbb{B}|}(-\sum_c \hat{p}_{t,c}^i \log \hat{p'}_{t,c}^i - \sum_c \hat{p'}_{t,c}^i \log \hat{p}_{t,c}^i), \quad (3)$$

where $\hat{p}_{t,c}^i$ and $\hat{p'}_{t,c}^i$ denote the probabilities assigned by the teacher and student models, respectively, to classify the i-th sample in the test mini-batch $\mathbb{B}$ as a certain class c. Following each step of student optimization, the teacher model is updated by applying an exponential moving average (EMA) to the student weights.

$$\theta \leftarrow \eta \cdot \theta + (1-\eta) \cdot \theta', \quad (4)$$

where η is the EMA momentum, set as 0.99 following [37]. ***In summary, IA effectively utilizes the available information, making good use of the sparse MI-EEG data.***

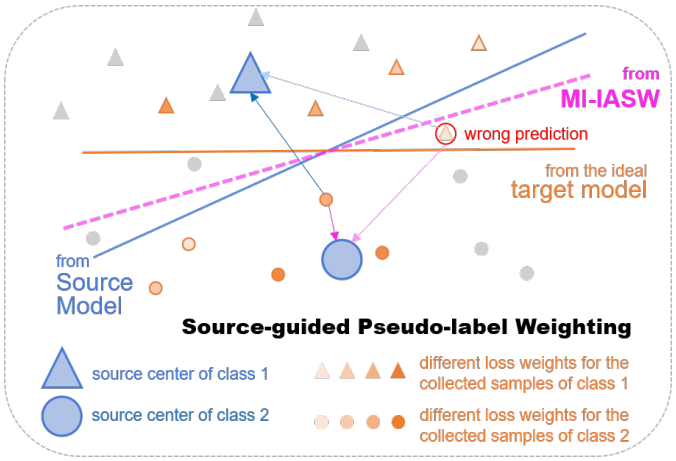

Fig. 2. This figure illustrates our SW. We leverage the well-generalized source centroids to evaluate overconfident pseudo-labels. We weight the pseudo-labels based on the result of this source-guided evaluation.

D. Source-Guided Pseudo-Label Weighting

In this section, we introduce our Source-Guided Pseudo-Label Weighting (SW) strategy, which consists of Source-Guided Pseudo-label Evaluation and Class-Aware Weighting.

In the absence of ground-truth labels, evaluating pseudo-label reliability is crucial for stable MI-EEG network training. However, we find that the overconfidence problem of the target MI-EEG model is severe. This issue arises from the sparse nature of the MI-EEG feature space (caused by the non-stationary characteristics and the low signal-to-noise ratio of the data [38]). In the sparse feature space, the unsupervised entropy minimization loss [15], [19] may drive the model to rely heavily on seemingly significant features that lack true discriminative power. Without actual confidence error guidance, this reliance leads to more extreme but potentially incorrect outputs. Consequently, evaluating noisy pseudo-labels becomes challenging for the target MI-EEG model, as it tends to treat most of the pseudo-labels it predicts as equally confident. ***In summary, we argue that the target model cannot accurately evaluate overconfident MI-EEG pseudo-labels***

To improve the evaluation of noisy pseudo-labels, we leverage the well-posed source centers (centroids) to evaluate how closely the predicted MI class of a pseudo-label aligns with the decision of these anchors. This approach is based on the observation that a fully supervised source model, trained with actual confidence errors (cross-entropy loss), is less reliant on the potentially spurious MI-EEG features and better retains the ability to withhold confidence when predictions are incorrect. Therefore, pseudo-labels deemed accurate based on source information are more likely to indicate true label reliability. Prioritizing such pseudo-labels (samples) also preserves source information and prevents model forgetting. ***Thus, we argue that guidance from the source model enables more accurate evaluation of overconfident MI-EEG pseudo-labels.***

Given the source centroid of each MI class $\{o_{s,c}\}_{c=1}^C$, we define the source alignment level (SAL) of a target sample, soft pseudo-label, and classification result tuple $(x_t^i, p_t^i, \hat{y}_t^i)$ for pseudo-label evaluation as follows:

$$\begin{aligned} SAL^i &= SAL_{rel}^i + SAL_{cons}^i \\ &= \frac{1}{C}\sum_c (d(f_s(x_t^i), o_{s,\hat{y}_t^i}) - d(f_s(x_t^i), o_{s,c})) \\ &\quad + d(f_s(x_t^i), o_{s,\hat{y}_t^i}), \end{aligned} \quad (5)$$

where $f_s(x_t^i)$ is the representation of sample x_t^i derived from the source encoder f_s, $o_{s,\hat{y}_t^i}$ is the source centroid of the predicted MI class $\hat{y}_t^i = \arg\max_c \{p_{t,c}^i\}_{c=1}^C$, and d is a similarity metric. $SAL_{cons}^i = d(f_s(x_t^i), o_{s,\hat{y}_t^i})$ measures the consistency between the source centroid $o_{s,\hat{y}_t^i}$ and the sample representation $f_s(x_t^i)$ obtained from the source model. Meanwhile, $SAL_{rel}^i = \frac{1}{C}\sum_c (d(f_s(x_t^i), o_{s,\hat{y}_t^i}) - d(f_s(x_t^i), o_{s,c}))$ is a relative term measuring the superiority of the consistency comparing to other MI classes. If the source alignment level of a target sample x_t^i is high, then the source model not only supports its pseudo-label's classification result $\hat{y}_t^i$, but does so with stronger confidence. Thus, it is safer to adapt the MI-EEG network using this sample.

With only the source model weights accessible in the fully TTA setting, source class centroids can be approximated by class prototypes stored in the source classifier [39]. Then, the similarities between the source centroids and the target MI-EEG sample can be represented by the source logits:

$$SAL^i \approx \frac{1}{C}\sum_c (z_{s,\hat{y}_t^i}^i - z_{s,c}^i) + z_{s,\hat{y}_t^i}^i, z_s^i = \mathcal{M}_s(x_t^i), \quad (6)$$

where z_s^i is the logit vector predicted by the source model $\mathcal{M}_s$ for x_t^i. Given samples' SAL values (the result of Source-Guided Pseudo-label Evaluation), we then differentiate each sample's contribution to $\mathcal{L}_{\text{sym}}$ through Class-Aware Weighting (CAW). For a test batch $\mathbb{B}$ and a sample x_t^i with its target-predicted MI class $c = \hat{y}_t^i$, we use $\mathbb{B}_c$ to denote the set of samples in $\mathbb{B}$ classified as class c. The CAW value α_t^i assigned to the i-th target sample is then defined as:

$$\alpha_t^i = \frac{1}{\epsilon + \exp\left(\beta\left(-SAL^i + \min_{j, x_t^j \in \mathbb{B}_c}(SAL^j)\right)\right)}, \quad (7)$$

TABLE I
TTA Classification Accuracies (%) on BCI-IV2a. The **bold** number shows the best result, and the underlined one the second best.

Methods	A01	A02	A03	A04	A05	A06	A07	A08	A09	avg
Source	67.71	50.00	77.26	54.86	51.39	49.83	67.88	74.31	64.93	62.02
TBN	71.53	46.35	82.81	**58.68**	**67.36**	55.90	72.92	76.91	64.58	66.34
PBN	71.53	47.05	82.81	58.16	65.10	55.73	73.09	77.08	64.58	66.13
PL	71.53	46.35	82.99	**58.68**	67.19	55.73	72.57	76.91	64.58	66.28
Eta	71.35	45.66	82.47	**58.68**	**67.36**	55.73	72.22	77.08	64.93	66.17
CoTTA	71.18	46.01	83.51	57.47	64.93	55.90	72.74	77.08	64.76	65.95
RMT	73.61	43.06	84.03	55.21	64.58	58.51	75.52	78.30	70.31	67.01
ROID	71.35	47.92	82.81	57.29	67.36	55.38	72.22	77.60	65.45	66.38
Versa	70.83	46.88	82.29	58.51	66.49	57.81	73.96	75.00	66.49	66.47
Ours	**76.39**	**50.17**	**84.55**	56.94	66.84	**58.68**	**76.22**	**79.51**	**71.35**	**68.94**

TABLE II
TTA Classification Accuracies(%) on BCI-IV2b. The **bold** number shows the best result, and the underlined one the second best.

Methods	B01	B02	B03	B04	B05	B06	B07	B08	B09	avg
Source	74.06	60.00	62.50	86.25	94.38	75.94	80.00	79.38	85.62	77.57
TBN	71.88	60.36	65.62	93.75	93.75	73.12	83.75	91.88	84.38	79.83
PBN	72.50	61.07	65.94	93.44	93.75	73.44	83.75	91.88	85.00	80.08
PL	72.81	58.93	65.00	93.75	93.75	72.19	83.44	92.19	85.31	79.71
Eta	73.44	58.93	65.94	93.75	93.44	71.88	83.75	92.19	85.62	79.88
CoTTA	72.81	59.29	65.62	93.44	93.75	73.12	83.75	91.88	84.38	79.78
RMT	72.50	59.29	72.81	**96.25**	**94.69**	84.38	86.25	**92.50**	88.12	82.98
ROID	74.38	58.93	65.94	93.75	93.44	75.94	83.44	91.88	85.00	80.30
Versa	69.69	58.57	74.69	95.00	92.81	81.56	85.62	90.31	85.31	81.51
Ours	**75.00**	**63.57**	**75.31**	95.62	93.44	**88.12**	**87.50**	91.25	**88.44**	**84.25**

where ϵ is a small value for numerical stability; β is a hyper-parameter that regulates the variation of CAW values (α_t) based on the evaluation differences among pseudo-labels. Specifically, it determines how quickly the weight decreases as samples approach the worst-performing sample ($\min_{j,x_t^j \in \mathbb{B}_c} \left(SAL^j\right)$) within the same class in the batch. An extremely large β may lead to overfitting on certain samples. Conversely, a small β allows flexible adaptation but inevitably results in error accumulation. We find that simply setting β to 1 achieves a favorable balance. α_t is normalized at the batch level to ensure numerical stability. The process of SW is illustrated in Figure 2. ***The introduction of α_t prioritizes safer pseudo-labels while retaining others, balancing stable learning and effective adaptation for online MI-EEG data.***

Note that our source-guided CAW only rescales the contribution of target pseudo-labels to the IA component. We do not use source model predictions to supervise the target network directly, as this hinders adaptation to the target subject. Unlike [17], our method dynamically adjusts the margin at the batch level to capture variations in non-stationary MI-EEG data. We also account for class-specific differences, as learning progress varies across MI categories, and ignoring these differences may hinder adaptation. Our belief also differs from previous source-prediction-based CV-TTA work [40], in which the confidence of the target model still dominates. The proposed Source-Guided Pseudo-Label Weighting, consisting of SAL and CAW, is a new concept to MI-EEG TTA.

III. Experiments and Results

A. Dataset

We evaluate online adaptation performance on two public MI-EEG datasets. BCI-IV2a [41] includes data from nine subjects (A01-A09), each completing 576 trials across four MI tasks (left hand, right hand, both feet, and tongue). As we adopt the LOSO setting, when a specific subject is designated as the target, data from all other subjects are aggregated to form the source domain (4608 trials). BCI-IV2b [42] contains left-hand and right-hand MI data from 9 subjects (B01-B09), with three offline sessions (about 400 trials) and two online sessions with feedback (approximately 320 trials) per subject. To assess the effectiveness of our algorithm in an online scenario with feedback, we merge the final two sessions of the target subject to create the test set. The offline sessions from the other subjects are used as the source domain.

B. Experimental Settings

The experiments are conducted on a single NVIDIA 3090. We first pretrain the backbone EEGNet [43]: 80% of the data from the source subjects is used as the training set, and the remaining 20% is used as the validation set. Following [3], we utilize the Adam optimizer with a learning rate of 0.0009, training epochs of 1000, an early-stopping patience of 300 using the validation-set, and a batch size of 64. When it comes to the target domain, all adaptation methods are initialized with the same source weights to ensure fairness. During testing, target data arrives one by one sequentially in an online manner. The Sliding batch strategy [32] is adopted for inference and training. For all experiments, the batch size is reduced to 32 for BCI-IV2a and 16 for BCI-IV2b, as the number of classes in BCI-IV2b is half that of BCI-IV2a. The proposed method updates the entire model during test time. The baseline methods can be categorized into two groups based on their adaptation settings. 1) Gradient-free methods: *Source* (source model), *TBN* [13], *PBN* [14]; 2) Gradient-based adaptation methods *PL* [16], *Eta* [17], *CoTTA* [18], *RMT* [19], *ROID* [21], *Versa* [22]. Method-specific hyper-parameters for comparison methods listed in Sections III-C and III-E are set according to the guidelines provided in their original papers.

All sample predictions strictly adhere to the causal principle: during inference for the i-th sample, the model has only observed test data from 1 to $i-1$, preventing any future influence. We then accumulate these predictions to compute the overall online TTA classification accuracy for a subject.

C. Online Classification Accuracies

Table I displays a comprehensive summary of online TTA classification accuracies on BCI-IV2a. Each column corresponds to a specific task where one subject is the target domain, while the others form the source domain. On BCI-IV2a, most baselines find it challenging to improve TBN significantly, likely because some methods only update the BN layers or suffer from degradation when applied to online MI-EEG data. Compared to the source model, our method enhances accuracy by 6.92%. Versa relies on target-model-based pseudo-label thresholds and MI-IASW improves it by 2.47%. Compared to the leading baseline RMT, our method enhances accuracy by 1.93%. Table II presents the classification outcomes for BCI-IV2b. Our framework achieves an average accuracy of 84.25% across different subjects and enhances the source model by 6.68%. Compared to ROID, which only adapts the BN affine parameters, our method achieves an improvement of 3.95%; Compared to RMT, MI-IASW improves the accuracy by 1.27%. In summary, MI-IASW consistently surpasses the baseline approaches in challenging online MI-EEG scenarios.

Figure 3 compares the MI-EEG TTA performance of MI-IASW against three representative methods, illustrating performance trends as the processed sample size increases. The horizontal axis represents the sample number that has been observed and processed by the TTA algorithm during training and inference, while the vertical axis shows the cumulative

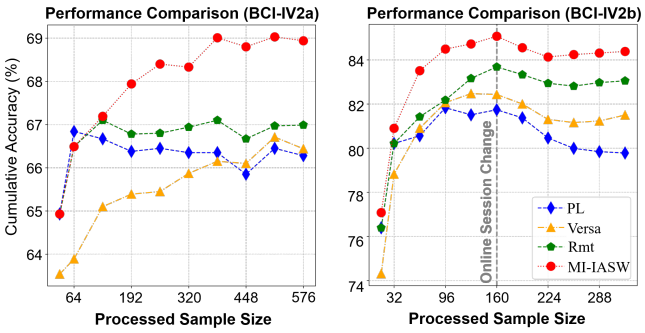

Fig. 3. Performance comparison throughout the adaptation process.

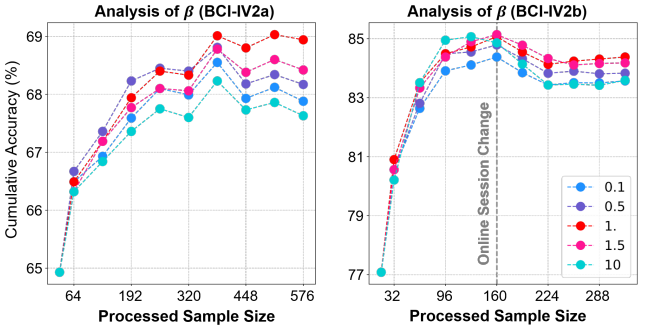

Fig. 4. Hyper-parameter Analysis.

accuracy of the TTA algorithm up to that sample number. Besides, the cumulative accuracy is calculated across all the subjects within a specific dataset. On BCI-IV2a, MI-IASW demonstrates superior performance compared to other methods, particularly in the later stages. Versa, adopting pseudo-label calibration, struggles in the early stages due to sample sparsity. While PL achieves higher initial cumulative accuracy due to its faster learning of hard pseudo-labels, MI-IASW maintains consistent improvement through sustained adaptation rather than experiencing performance degradation. This is achieved through IA's full utilization of available information and SW's enhancement of stable adaptation. On BCI-IV2b, the occurrence of online session changes results in a significant drop in performance for all algorithms, attributed to the severe domain shift. Nevertheless, MI-IASW recovers more quickly than other competitive methods and gradually enhances its performance over time.

MI-IASW is currently being deployed in real-world applications. It is capable of providing real-time feedback, averaging 59 ms per batch for both training and inference, meeting the timing requirements of MI-EEG rehabilitation paradigms. ***With its outstanding performance, MI-IASW has the potential to enhance user-friendly BCIs by eliminating the need for offline calibration trials.***

D. Ablation Studies

We conduct ablation experiments on the two datasets and the results are shown in Table IV. 'base' represents tuning the whole network in an unsupervised manner, with TBN equipped. The addition of IA (WA and MABN) facilitates better information mining, leading to a substantial improvement of 1.55% and 3.43% on BCI-IV2a and BCI-IV2b re-

TABLE III
COMPARISON WITH RELATED METHODS: SOURCE-MODEL REGULARIZATION AND SAMPLE WEIGHTING

Method / Model used	Weight Regularization (1)	Weight Interpolation (2)	Sample Weighting				
			Hard Weighting		Soft Weighting		
			Dy-Thre (3)	**Agree** (4)	**Eta** (5)	**Exp** (6)	**CAW**
Target only	✗	✗	✗	✗	$-\sum_c p_{t,c} \log p_{t,c}$	DIV	TAL (7)
Target + Source	$\beta^l \left\|\theta_t^l - \theta_s^l\right\|_2^2$	$\eta \cdot \theta_t + (1-\eta) \cdot \theta_s$	$\mathbb{I}_{\{p_{t,\hat{y}_t^i} > p_{s,\hat{y}_t^i}\}}$	$\mathbb{I}_{\{\hat{y}_t^i = \hat{y}_s^i\}}$	✗	✗	SAL (8)

TABLE IV
ABLATION STUDY ON BCI-IV2a AND BCI-IV2b.

Methods	BCI-IV2a	BCI-IV2b	avg
base	65.93	79.13	72.53
+ WA	67.03	82.56	74.80
+ MABN	67.48	83.35	75.42
+ CAW & TAL	67.69	83.72	75.71
MI-IASW	**68.94**	**84.25**	**76.60**

TABLE V
DETAILED COMPARISON RESULTS ON BCI-IV2a AND BCI-IV2b.

Methods	BCI-IV2a	BCI-IV2b	avg
base (IA)	67.48	83.35	75.42
(1) Weight Regularization [22]	67.63	83.12	75.38
(2) Weight Interpolation [21]	67.59	83.18	75.39
(3) Dy-thre [40]	66.32	83.62	74.97
(4) Agree	66.59	83.66	75.13
(5) Eta & ENT [17]	67.40	83.34	75.37
(6) Exp & DIV [21] [22]	67.23	83.64	75.44
(7) CAW & TAL	67.69	83.72	75.71
(8) CAW & SAL_{cons}	68.25	83.88	76.07
(8) CAW & SAL_{rel}	68.33	83.88	76.11
(8) CAW & SAL (***MI-IASW***)	**68.94**	**84.25**	**76.60**

spectively. In particular, WA, comprising the EMA-teacher and symmetric entropy loss, proves especially essential for BCI-IV2b.In brief, IA lays a solid foundation for MI-EEG TTA. 'CAW & TAL' represents the combination of target alignment level (TAL), which uses the target logits for evaluation ($TAL^i \approx \frac{1}{C}\sum_c (z^i_{t,\hat{y}^i_t} - z^i_{t,c}) + z_{t,\hat{y}^i_t}, z^i_t = \mathcal{M}_t(x^i_t)$), and CAW values α_t calculated using TAL. Finally, replacing TAL with SAL leads to our MI-IASW. Harnessing source guidance, it further improves upon the well-performing IA component and achieves the best average accuracy of 76.60% across different subjects and datasets.

Next, we analyze the impact of different settings for the hyper-parameter β used in CAW. The cumulative accuracies of different β are presented in Figure 4. Setting β=1 strikes a favorable balance between safe learning and adaptation to intra-subject shifts by placing appropriate emphasis on reliable samples. When β deviates from 1, the cumulative accuracy shows a similar trend during TTA but results in varying final accuracies. Specifically, when β is closer to 1 (0.5 and 1.5, represented by purple and pink, respectively), the cumulative accuracy remains competitive across both datasets, achieving a final accuracy greater than 68.1% on BCI-IV2a and 83.8% on BCI-IV2b. Setting β to 0.1 tends to weaken the weight distinction between pseudo-labels, while setting it to 10 may result in overfitting. Both settings show slightly inferior performance, aligning closer to the baseline (IA, 67.48% on BCI-IV2a and 83.35% on BCI-IV2b).

E. Detailed Comparison with Related Methods

Most strategies proposed in the baseline methods are used in combination with others, complicating the assessment of their individual impact on adaptation performance. To enable a direct comparison with methods related to SW, namely regularization and sample weighting (pseudo-label handling) strategies, we integrate all these methods with IA. This ensures a more consistent comparison under online MI-EEG scenarios.

The comparison methods are listed in Table III. The source-model regularization methods include Weight Regularization (1), used in [22], and Weight Interpolation (2), proposed in [21]. (1) applies a layer-wise penalty on the difference between target and source model weights, formulated as $\beta^l \left\|\theta_t^l - \theta_s^l\right\|_2^2$, where β^l controls the regularization strength at each layer l. (2) instead updates the target weight with $\eta \cdot \theta_t + (1-\eta) \cdot \theta_s$, where η determines the degree of interpolation between the source and target models. Sample weighting is divided into two categories: Hard Weighting (discarding) and Soft Weighting. Hard Weighting includes Dy-Threshold (3), which discards samples whose target maximum softmax probability (MSP) is smaller than its source counterpart [40], i.e., $\mathbb{I}_{\{p_{t,\hat{y}^i_t} > p_{s,\hat{y}^i_t}\}}$, and Agree (4), which retains only samples whose source-model classification results agree with the target ones, i.e., $\mathbb{I}_{\{\hat{y}^i_t = \hat{y}^i_s\}}$. Eta (5) denotes the entropy-based ($ENT = -\sum_c p_{t,c} \log p_{t,c}$) weighting loss from [17], with its sample discarding part excluded to address data scarcity. (6) employs a variant of ENT that incorporates diversity into pseudo-label evaluation [21] [22]. This sample evaluation metric, DIV, is combined with the exponential (Exp) function using a temperature for sample weighting. (7) is a variant of our method, which uses target logits for alignment calculation (TAL). Methods (5)–(7) rely solely on the predictions of the target model for pseudo-

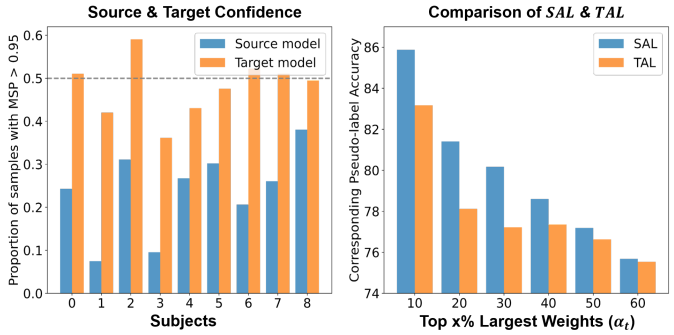

Fig. 5. Source-Target Evaluation Comparison

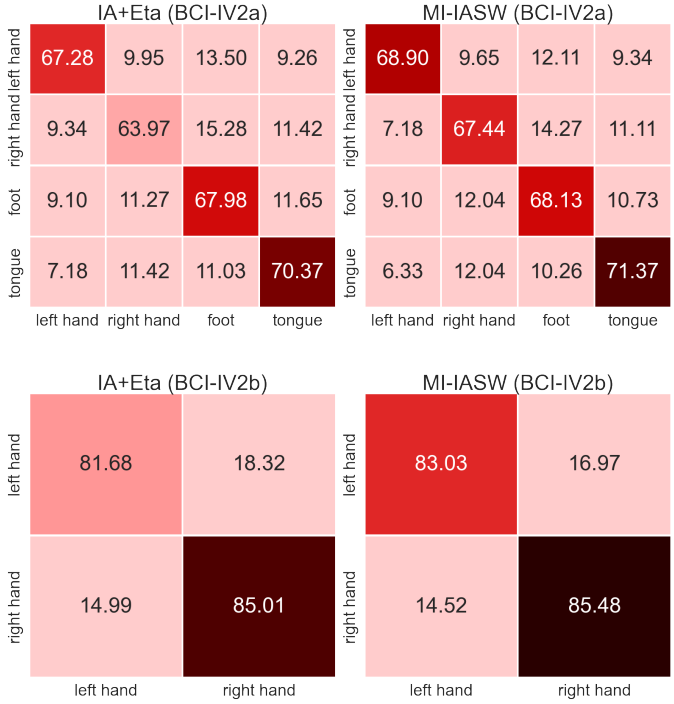

Fig. 6. Confusion Matrices Analysis

label evaluation. (8) is our proposed SW, where CAW and SAL are both applied. We provide a detailed comparison of (8), including the exclusive use of SAL_{cons} and SAL_{rel}.

The comparison results are listed in Table V. It can be observed that the regularization methods fail to provide consistent improvements, as they impose strict constraints on the weight domain. (3) and (4) result in model degradation on BCI-IV2a, as they discard some samples and make the data scarcer. (5)–(7) fail to bring consistent improvement, as they rely on the overconfident target MI-EEG network. The influence of overconfidence persists even when the evaluation difference is amplified by a manually selected temperature in (6). Only the methods under (8) consistently perform well on both datasets by effectively leveraging source guidance for target adaptation. They achieve this by applying soft constraints to the target MI-EEG pseudo-labels without directly limiting the updates to model weights, ensuring safe adaptation to the target subject. The two variants of SAL perform well when used individually, but their combined application achieves the best performance. To summarize, in MI-EEG scenarios, the most effective approach is to guide the overconfident target model with source information using Soft Weighting.

F. Visualization

Figure 5 compares the confidence levels of the target and source models on BCI-IV2a. It can be seen that the target models tend to predict high-confidence pseudo-labels. For each subject, nearly 50% of the samples were assigned with a MSP (confidence level) greater than 0.95 by the target model. This means that the overconfidence issue for MI-EEG TTA is more severe than that observed in a CV-TTA work [30], where the proportion of samples with $MSP > 0.95$ is significantly below 0.5. Such excessive confidence hinders the target MI-EEG network's ability to recognize its mistakes and may reduce its effectiveness in evaluating pseudo-labels. We then investigate how incorporating source information can improve the evaluation of pseudo-labels. Specifically, we compare the pseudo-label evaluation between target-model-based evaluation (TAL) and source-model-based evaluation (SAL) in the top half of batches with the highest confidence. By analyzing the pseudo-label accuracy of samples within the top 10% to 60% highest CAW values, we find that SAL consistently achieves higher accuracy. This suggests that using source information in evaluation assigns higher CAW values to more accurate pseudo-labels, promoting safer adaptation. In contrast, TAL and the target model are prone to being misled by falsely confident samples, persistently assigning higher weights to less accurate pseudo-labels. ***These comparison results support our insight that incorporating source information enables better evaluation of the overconfident MI-EEG pseudo-labels.***

Figure 6 presents the confusion matrices for IA integrated with the entropy-based loss proposed in [17] (IA+Eta, left) and IA combined with our SAL-based CAW (MI-IASW, right). The inclusion of SAL-based CAW in our method leads to notable improvements for classes prone to confusion (e.g., left hand and right hand), with accuracy gains of 1.62% and 3.47% on BCI-IV2a. Additionally, tongue classification also demonstrates consistent relative improvement. Furthermore, SW improves classification accuracy for both classes on BCI-IV2b as well. These findings highlight the effectiveness of SW (SAL-based CAW) in mitigating class confusion by incorporating class-aware information and source guidance.

Conclusion

This study focuses on online unsupervised cross-subject adaptation during test-time for MI-EEG classification. Specifically, we propose a practical yet effective MI-EEG TTA framework called MI-IASW, enabling calibration-free model adaptation to new subjects during testing. It innovatively tackles the challenges of online MI-EEG data scarcity and unreliable MI-EEG pseudo-labels, through Information Aggregation (MABN and Weight Aggregation) and Source-Guided Pseudo-label Weighting strategies (SAL and CAW).

Extensive experimental evaluations highlight the superiority of our proposed framework. One limitation of the current approach is that it does not account for the gradually changing level of overconfidence in the target model during adaptation. Future work may incorporate this factor into the pseudo-label evaluation process to further improve MI-EEG TTA accuracy. Our work demonstrates the potential to eliminate offline calibration requirements for subjects. This possibility is especially valuable in BCI applications, where reducing user burden and improving accessibility are crucial.

ACKNOWLEDGMENT

This work is supported in part by the National Natural Science Foundation of China under Grant 62403140, in part by the National Key Research and Development Program of China under Grant 2022YFC361400 / 2022YFC3601401, in part by the National Natural Science Foundation of China under Grant U1913216, and in part by the Independent Deployment of Scientific Research Projects of Jihua Laboratory under Grant X190051TB190.

REFERENCES

[1] B. He, B. Baxter, B. J. Edelman, C. C. Cline, and W. Y. Wenjing, "Non-invasive brain-computer interfaces based on sensorimotor rhythms," *Proceedings of the IEEE*, vol. 103, no. 6, pp. 907–925, 2015.

[2] X. Gu et al., "EEG-based brain-computer interfaces (BCIs): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 18, no. 5, pp. 1645–1666, 2021.

[3] H. Altaheri, G. Muhammad, and M. Alsulaiman, "Physics-informed attention temporal convolutional network for EEG-based motor imagery classification," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 2249–2258, 2022.

[4] A. Zimmermann-Schlatter, C. Schuster, M. A. Puhan, E. Siekierka, and J. Steurer, "Efficacy of motor imagery in post-stroke rehabilitation: a systematic review," *Journal of neuroengineering and rehabilitation*, vol. 5, no. 1, pp. 1–10, 2008.

[5] W. Zhang and D. Wu, "Manifold embedded knowledge transfer for brain-computer interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 5, pp. 1117–1127, 2020.

[6] X. Hong et al., "Dynamic joint domain adaptation network for motor imagery classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 556–565, 2021.

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

[8] W. Zhang, Z. Wang, and D. Wu, "Multi-source decentralized transfer for privacy-preserving BCIs," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 2710–2720, 2022.

[9] W. Zhang and D. Wu, "Lightweight source-free transfer for privacy-preserving motor imagery classification," *IEEE Transactions on Cognitive and Developmental Systems*, 2022.

[10] K. Xia, L. Deng, W. Duch, and D. Wu, "Privacy-preserving domain adaptation for motor imagery-based brain-computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 11, pp. 3365–3376, 2022.

[11] S. Li, Z. Wang, H. Luo, L. Ding, and D. Wu, "T-TIME: Test-time information maximization ensemble for plug-and-play BCIs," *IEEE Transactions on Biomedical Engineering*, 2023.

[12] M. Wimpff, M. Döbler, and B. Yang, "Calibration-free online test-time adaptation for electroencephalography motor imagery decoding," in *2024 12th International Winter Conference on Brain-Computer Interface (BCI)*. IEEE, 2024, pp. 1–6.

[13] Z. Nado et al., "Evaluating prediction-time batch normalization for robustness under covariate shift," *arXiv preprint arXiv:2006.10963*, 2020.

[14] S. Schneider et al., "Improving robustness against common corruptions by covariate shift adaptation," *Advances in neural information processing systems*, vol. 33, pp. 11 539–11 551, 2020.

[15] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," *arXiv preprint arXiv:2006.10726*, 2020.

[16] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2. Atlanta, 2013, p. 896.

[17] S. Niu et al., "Efficient test-time model adaptation without forgetting," in *International conference on machine learning*. PMLR, 2022, pp. 16 888–16 905.

[18] Q. Wang, O. Fink, L. Van Gool, and D. Dai, "Continual test-time domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7201–7211.

[19] M. Döbler, R. A. Marsden, and B. Yang, "Robust mean teacher for continual and gradual test-time adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7704–7714.

[20] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 322–330.

[21] R. A. Marsden, M. Döbler, and B. Yang, "Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 2555–2565.

[22] X. Yang, X. Chen, M. Li, K. Wei, and C. Deng, "A versatile framework for continual test-time domain adaptation: Balancing discriminability and generalizability," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 731–23 740.

[23] O. Press, S. Schneider, M. Kümmerer, and M. Bethge, "Rdumb: A simple approach that questions our progress in continual test-time adaptation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[24] T. Gong, J. Jeong, T. Kim, Y. Kim, J. Shin, and S.-J. Lee, "Note: Robust continual test-time adaptation against temporal correlation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 253–27 266, 2022.

[25] B. Zhao, C. Chen, and S.-T. Xia, "DELTA: degradation-free fully test-time adaptation," *arXiv preprint arXiv:2301.13018*, 2023.

[26] S. Niu, J. Wu, Y. Zhang, Z. Wen, Y. Chen, P. Zhao, and M. Tan, "Towards stable test-time adaptation in dynamic wild world," *arXiv preprint arXiv:2302.12400*, 2023.

[27] L. Yuan, B. Xie, and S. Li, "Robust test-time adaptation in dynamic scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 922–15 932.

[28] Y. Su, X. Xu, and K. Jia, "Towards real-world test-time adaptation: Tri-net self-training with balanced normalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 13, 2024, pp. 15 126–15 135.

[29] Z. Wang, Z. Chi, Y. Wu, L. Gu, Z. Liu, K. Plataniotis, and Y. Wang, "Distribution alignment for fully test-time adaptation with dynamic online data streams," in *European Conference on Computer Vision*. Springer, 2025, pp. 332–349.

[30] H. Yang, M. Wang, J. Jiang, and Y. Zhou, "Towards test time adaptation via calibrated entropy minimization," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 3736–3746.

[31] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.

[32] Y. Su, X. Xu, and K. Jia, "Revisiting realistic test-time training: Sequential inference and adaptation by anchored clustering," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 543–17 555, 2022.

[33] S. Ioffe, "Batch renormalization: Towards reducing minibatch dependence in batch-normalized models," *Advances in neural information processing systems*, vol. 30, 2017.

[34] H. Lim, B. Kim, J. Choo, and S. Choi, "Ttn: A domain-shift aware batch normalization in test-time adaptation," in *The Eleventh International Conference on Learning Representations*.

[35] J. Kang, N. Kim, J. Ok, and S. Kwak, "Membn: Robust test-time adaptation via batch norm with statistics memory," in *European Conference on Computer Vision*. Springer, 2025, pp. 467–483.
[36] Y.-C. Liu et al., "Unbiased teacher for semi-supervised object detection," *arXiv preprint arXiv:2102.09480*, 2021.
[37] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.
[38] H. Zhao, Q. Zheng, K. Ma, H. Li, and Y. Zheng, "Deep representation-based domain adaptation for nonstationary eeg classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 535–545, 2020.
[39] Y. Iwasawa and Y. Matsuo, "Test-time classifier adjustment module for model-agnostic domain generalization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2427–2440, 2021.
[40] J. Lee, D. Das, J. Choo, and S. Choi, "Towards open-set test-time adaptation utilizing the wisdom of crowds in entropy minimization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 380–16 389.
[41] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "BCI Competition 2008–Graz data set A," *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology*, vol. 16, pp. 1–6, 2008.
[42] R. Leeb, C. Brunner, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "BCI Competition 2008–Graz data set B," *Graz University of Technology, Austria*, vol. 16, pp. 1–6, 2008.
[43] V. J. Lawhern et al., "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018.