

# Calibration-Free EEG-based Driver Drowsiness Detection with Online Test-Time Adaptation

Geun-Deok Jang\*, Dong-Kyun Han\*, Seo-Hyeon Park, and Seong-Whan Lee, *Fellow, IEEE*

**Abstract**—Drowsy driving is a growing cause of traffic accidents, prompting recent exploration of electroencephalography (EEG)-based drowsiness detection systems. However, the inherent variability of EEG signals due to psychological and physical factors necessitates a cumbersome calibration process. In particular, the inter-subject variability of EEG signals leads to a domain shift problem, which makes it challenging to generalize drowsiness detection models to unseen target subjects. To address these issues, we propose a novel driver drowsiness detection framework that leverages online test-time adaptation (TTA) methods to dynamically adjust to target subject distributions. Our proposed method updates the learnable parameters in batch normalization (BN) layers, while preserving pretrained normalization statistics, resulting in a modified configuration that ensures effective adaptation during test time. We incorporate a memory bank that dynamically manages streaming EEG segments, selecting samples based on their reliability determined by negative energy scores and persistence time. In addition, we introduce prototype learning to ensure robust predictions against distribution shifts over time. We validated our method on the sustained-attention driving dataset collected in a simulated environment, where drowsiness was estimated from delayed reaction times during monotonous lane-keeping tasks. Our experiments show that our method outperforms all baselines, achieving an average F1-score of 81.73%, an improvement of 11.73% over the best TTA baseline. This demonstrates that our proposed method significantly enhances the adaptability of EEG-based drowsiness detection systems in non-i.i.d. scenarios.

**Index Terms**—Driver drowsiness detection, electroencephalography, domain shift, and test-time adaptation.

## I. INTRODUCTION

**D**RIVER drowsiness is a major concern in road safety, substantially increasing the risk of traffic accidents worldwide [1]. Most road accidents can be attributed to driver fatigue, drowsiness, and impaired judgment [2]. In this context, we specifically focus on driver drowsiness, the transitional state between wakefulness and sleep, distinguishing it from fatigue [3], [4]. To effectively monitor these cognitive states, electroencephalography (EEG)-based driver drowsiness

This work was partly supported by Institute of Information and Communications Technology Planning and Evaluation (IITP) grants funded by the Korea government (No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University); No. 2021-0-02068, Artificial Intelligence Innovation Hub). (Geun-Deok Jang and Dong-Kyun Han contributed equally to this work.) (Corresponding author: Seong-Whan Lee)

Geun-Deok Jang and Seong-Whan Lee are with the Department of Artificial Intelligence, Korea University, Seongbuk-ku, Seoul 02841, Republic of Korea (e-mail: gd\_jang@korea.ac.kr; sw.lee@korea.ac.kr).

Dong-Kyun Han and Seo-Hyeon Park are with the Department of Brain and Cognitive Engineering, Korea University, Seongbuk-ku, Seoul 02841, Republic of Korea (e-mail: dk\_han@korea.ac.kr; tjugus9190@korea.ac.kr).

\* equal contributions

detection systems have been actively studied. EEG is a non-invasive method for recording electrical activity in the brain and directly assessing human cognitive states, making it increasingly favored for enhancing the performance of driver monitoring and assistance systems [5].

However, a major obstacle in developing EEG-based monitoring systems is the inherent variability of EEG signals between individuals and even within the same individual across sessions. Since this variability in EEG signals complicates the construction of robust monitoring systems, a cumbersome calibration session is required [6]. To exclude time-consuming and labor-intensive calibration sessions, recent studies [7], [8] have investigated subject-independent brain-computer interface (BCI) systems, which utilize transfer learning methods such as domain adaptation (DA) and domain generalization (DG) approaches.

DA aims to minimize the domain shifts, typically using the labeled data from the source domain and unlabeled data from the target domain. However, it is infeasible to retrieve all the target subject data in monitoring systems. On the other hand, DG aims to build robust networks without any data from target distributions, which is more practical for real-world applications. DG methods mitigate domain shifts using techniques such as data augmentation [8] and domain-invariant representations [6]. However, DG methods make it difficult to generalize the model without prior knowledge of the unseen target domain, which often leads to limited performance.

To overcome the limitations of domain transfer learning, several studies [9]–[13] in the field of computer vision have introduced the test-time adaptation (TTA) algorithm, which dynamically adapts the model parameters to the target domain during inference. The differences among DA, DG, and TTA are illustrated in Fig. 1. Inspired by these advancements, recent research has explored TTA methods in EEG-based systems such as motor imagery [14] and seizure prediction [15]. These systems demonstrated the potential of TTA techniques to improve adaptability and performance in their respective applications. Despite these efforts, research on EEG-based driver drowsiness detection systems remains limited. Moreover, no study has yet explored online TTA methods to effectively address the challenges of an EEG-based driver drowsiness detection system.

In this study, we propose a novel driver drowsiness detection framework incorporating online TTA methods that handle the variability of EEG signals. The proposed method introduces a memory bank to store streaming target data over time, which is managed by a novel removal criterion that discards unreliable samples determined by negative energy scores

and persistence time. When adapting the model, we fix the pretrained normalization statistics of the batch normalization (BN) layers and update only their affine parameters. In addition, our method incorporates prototype learning (PL) based on the memory bank data to ensure robust predictions against distribution shifts. Through our calibration-free framework, we address the critical challenge of developing an EEG-based driver drowsiness detection framework for detecting driver drowsiness. A preliminary version of this method was reported in [16].

The primary contributions of this study are as follows:

- We propose a novel driver drowsiness detection framework utilizing online TTA methods, which dynamically adjusts to varying distributions and eliminates the need for labor-intensive calibration sessions.
- Our method utilizes only the pretrained parameters from source subjects without directly accessing their raw data, ensuring the privacy of source subjects.
- We update only the learnable parameters of the BN layers through a backward pass during model updates, thereby reducing adaptation costs.
- Extensive experiments demonstrate the adaptability of our method to non-stationary EEG signals, highlighting its effectiveness for real-world driver drowsiness detection.

## II. RELATED WORK

### A. Subject-Independent Driver Drowsiness Detection

EEG-based driver drowsiness detection can significantly enhance driving safety. However, the individual variability among drivers presents a considerable challenge when developing a robust driver drowsiness detection framework. Driver drowsiness detection in the cross-subject scenario often fails to account for diverse EEG patterns exhibited by different individuals. Therefore, adaptive approaches are necessary for addressing these variations and improving the reliability of driver drowsiness detection frameworks.

Cui *et al.* [7] presented a convolutional neural network (CNN) with separable convolutions that capture patterns of EEG signals in a spatio-temporal sequence. Paulo *et al.* [6] demonstrated cross-subject drowsiness detection using spatio-temporal image encoding of EEG signals via recurrence plots and Gramian angular fields without requiring subject-specific calibration. Zhuang *et al.* [17] proposed a GNN-based model that captures EEG connectivity using self-attention, enhanced by a squeeze-and-excitation block to emphasize informative frequency bands. Gong *et al.* [18] proposed a time-frequency attention model for single-channel EEG that highlights mental-state-relevant regions and integrates them via adaptive feature fusion.

To eliminate the need for any calibration data from the target subject, some researchers have investigated domain transfer learning methods that mitigate domain shifts. Kim *et al.* [8] augmented the data at the instance level and aligned class-related features by minimizing the distance between intra-class soft labels.

Ding *et al.* [19] proposed LGGNet, a neurologically inspired GNN that models both local and global brain connectivity

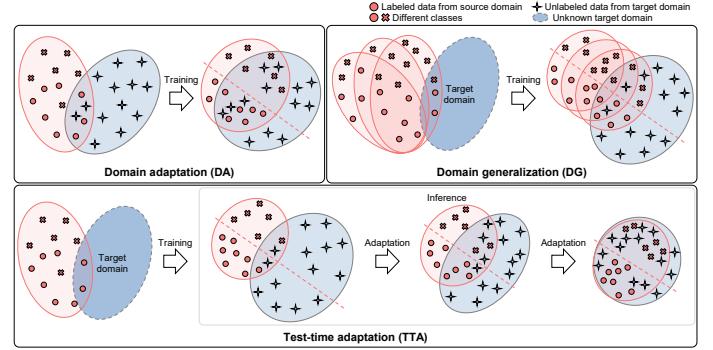


Fig. 1. Comparison of the transfer learning algorithms. DA utilizes both labeled source and unlabeled target domain data during training. Subsequently, DA makes a prediction with the pretrained model. DG uses only labeled source domain data during training and predicts with a pretrained model on unseen domains. TTA learns labeled source domain data during training, then dynamically adapts the pretrained model with input target domain data. This encourages effective predictions during inference with the adjusted model.

using functional-area-based EEG graphs, combining multiscale temporal convolutions with graph filtering.

Although recent studies often use the terms “fatigue” and “drowsiness” interchangeably, following prior works that adopt the term “drowsiness” [6]–[8], [17], [18], [20], we consistently use the term “drowsiness” throughout this paper.

### B. Test-Time Adaptation (TTA)

To mitigate the existing limitations of DA and DG in the field of computer vision, several studies have optimized the model parameters for the target domain data by dynamically adapting them to unlabeled target samples during inference. This approach enhanced the adaptability and reliability in real-world applications, particularly in environments where labeled target data are scarce or unavailable.

Several studies have explored practical TTA setups applicable to real-world scenarios [10]–[13]. For instance, Wang *et al.* [10] introduced weight-averaged and augmentation-averaged predictions to mitigate error accumulation and catastrophic forgetting in continually changing domains. Gong *et al.* [12] proposed an instance-aware BN layer and prediction-balanced reservoir sampling to overcome the challenges of non-i.i.d. target data streams. They address the challenges of non-i.i.d. target data streams in online settings. Yuan *et al.* [11] proposed a robust BN layer and category-balanced memory buffer. Their removal criterion considered timeliness and uncertainty to handle practical TTA scenarios in which the distribution continually changes and correlated sampling occurs over time.

Niu *et al.* [13] explored why TTA often fails in real-world scenarios, identifying that the normalization statistics in BN layers become unreliable under mixed domain shifts, small batch sizes, and imbalanced label distributions. To address this instability, they proposed a sharpness-aware and reliable entropy minimization method. This approach mitigates the impact of noisy test samples with large gradient norms, which can destabilize model adaptation.

In addition, Iwasawa *et al.* [21] have explored the use of PL to capture the underlying data structure. They demonstrated

that PL is appropriate for online TTA setups, where the model needs to be adjusted for input target instances.

### C. TTA for EEG-based Brain-Computer Interface Systems

In the field of EEG-based BCI systems, only a few studies have focused on online TTA to adapt BCI systems to the EEG signals from target subjects. Mao *et al.* [15] introduced a teacher-student online TTA approach based on knowledge distillation and exponential moving average strategies. This protected the patient privacy of source subjects and allowed real-time seizure prediction for timely medical intervention. Wimpff *et al.* [14] introduce an online TTA method for EEG-based motor imagery decoding to address distribution shifts between sessions and subjects without calibration data. Their approach integrates covariance alignment to align EEG covariance matrices between sessions, BN layers to replace source statistics with target statistics for better alignment with the target domain, and entropy minimization to increase prediction confidence. This combination achieves a privacy-preserving, calibration-free adaptation that effectively handles diverse EEG conditions, making BCI systems more applicable to real-world use.

Despite these efforts, research on EEG-based BCI systems that handle the distribution shifts remains limited. Moreover, no study has yet explored online TTA methods to effectively address the challenges in developing an EEG-based driver drowsiness detection system.

Among the reviewed methods, we specifically select DeepConvNet [22], EEGNet8,2 [23], ResNet1D-8 [8], InterpretableCNN [7], EEG-Conformer [24], CAGNN [17], and LGGNet-F [19] as subject-independent EEG-based benchmarks, and Tent [9], CoTTA [10], NOTE [12], RoTTA [11], T3A [21], SAR [13], and Wimpff *et al.* [14] as representative TTA baselines. These methods are compared against our proposed approach in Section IV. EXPERIMENTS.

## III. METHOD

EEG signals exhibit temporal dynamics characterized by strong temporal correlations and gradual, yet continuous, shifts in cognitive states. Unlike static or independently sampled data, EEG signals during driving reflect correlated mental states and patterns across consecutive samples. This temporal dependency violates the common assumption of sample independence inherent in many existing TTA methods, making them ineffective for EEG adaptation. To address the non-stationarity and domain shift of driver drowsiness detection systems, we introduce a novel driver drowsiness detection framework that utilizes online TTA methods. We updated only the learnable parameters of the BN layers with entropy minimization regularized by energy-bounded loss, while preserving the pretrained normalization statistics. Our proposed framework manages a memory bank based on the confidence score, which considers the energy score and persistence time. The prototypes are then updated and constructed with the features and pseudo-labels using the memory data and the adjusted model. Predictions are made by computing the similarity between the features of the current input data and the prototypes. In particular, the proposed

framework is designed to accommodate the temporal dynamics of EEG signals, which often exhibit gradual shifts in cognitive states. By incorporating persistence-aware memory filtering and sequential prototype updates, our method continuously adapts to the evolving distribution of streaming EEG data.

Fig. 2 illustrates the overall process of our proposed framework. We describe the procedure as a pseudo-code in Algorithm 1.

### A. Preliminaries

Let  $P_S$  and  $P_T$  denote the probability distributions of data from the source and target domains, respectively. Most domain transfer learning methods aim to solve the covariate shift between the source and target domains  $P_S(\mathbf{x}) \neq P_T(\mathbf{x})$  and  $P_S(y|\mathbf{x}) = P_T(y|\mathbf{x})$  [9], [10]. However, the driver's states in the driver drowsiness detection system cause the correlation between test samples, which leads to a difference between the observed distribution over time  $P_T(\mathbf{x}, y|t_1) \neq P_T(\mathbf{x}, y|t_2)$  [11], [12], making it challenging to work well on the previous TTA methods.

### B. Configuration of the BN layers for non-i.i.d. EEG samples

The BN layer [25] is widely used in neural network training due to its ability to promote convergence and stability. Given the input feature  $\mathbf{z} \in \mathbb{R}^{N \times C \times H \times W}$  and channel-wise statistics<sup>1</sup> ( $\mu_c$  and  $\sigma_c^2$ ), BN layers normalize the feature for each channel:

$$\text{BN}(\mathbf{z}_{:,c,:,:}; \mu_c, \sigma_c^2) = \gamma_c \frac{\mathbf{z}_{:,c,:,:} - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}} + \beta_c, \quad (1)$$

where  $\gamma_c$  and  $\beta_c$  denote channel-wise learnable parameters (scale and shift) and  $\epsilon$  denotes a small positive constant that prevents numerical instability.

Given the channel-wise mini-batch statistics ( $\hat{\mu}_c$  and  $\hat{\sigma}_c^2$ ) calculated from the input data, the normalization statistics of the BN layers are continuously updated using a momentum term:

$$\hat{\mu}_c = \frac{1}{NHW} \sum_{n,h,w} \mathbf{z}_{n,c,h,w}, \quad (2a)$$

$$\hat{\sigma}_c^2 = \frac{1}{NHW} \sum_{n,h,w} (\mathbf{z}_{n,c,h,w} - \hat{\mu}_c)^2, \quad (2b)$$

$$\mu_c = \delta \cdot \mu_c + (1 - \delta) \cdot \hat{\mu}_c, \quad (2c)$$

$$\sigma_c^2 = \delta \cdot \sigma_c^2 + (1 - \delta) \cdot \hat{\sigma}_c^2, \quad (2d)$$

where  $\delta$  is the momentum term that balances the influence of new mini-batch statistics against the running estimates.

Following recent studies [9]–[12], [14], we focus on optimizing the BN layers on a given feature extractor during inference. Existing TTA methods can be categorized based on how they apply normalization statistics in BN layers. Some studies [9], [10], [14] normalize target data using only mini-batch statistics (2a and 2b) for both updating the parameters and prediction. Conversely, other studies [11], [12] use a continuous tracking

<sup>1</sup>We use the terms *statistics* and *normalization statistics* interchangeably in the context of the BN layers.

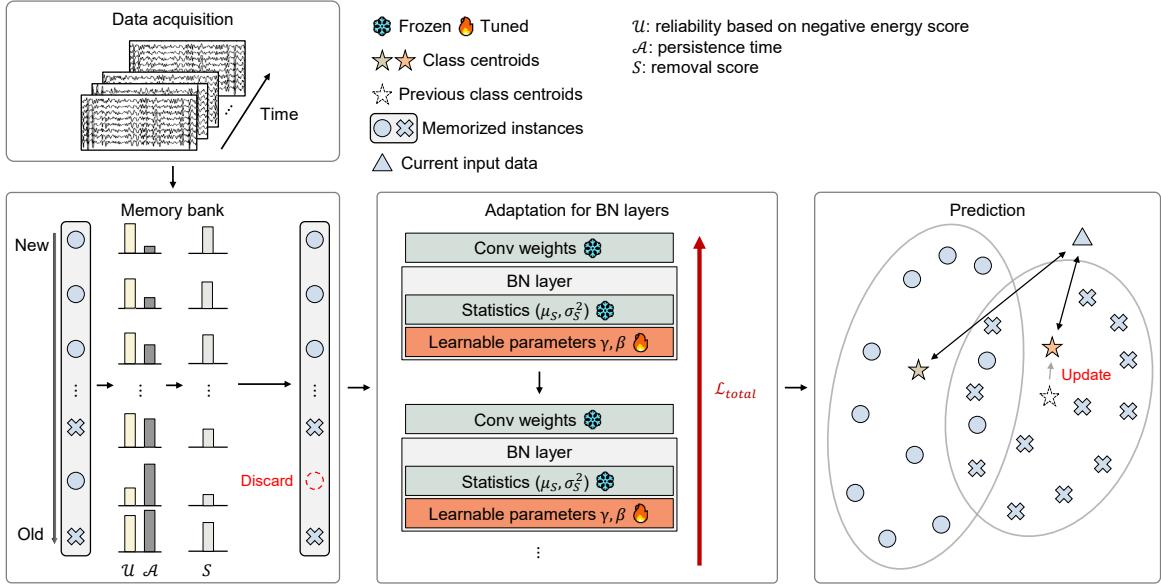


Fig. 2. Overview of the proposed test-time adaptation framework. Incoming EEG segments from the target subject are filtered by entropy and stored in a memory bank based on energy score and persistence. The memory bank is used to update only the learnable parameters of the BN layers via entropy minimization with energy-based regularization. Class prototypes are continuously updated from the memory to support robust prediction under distribution shifts.

of statistics from input target features to update parameters (2c and 2d). Subsequently, they apply these updated statistics for prediction.

While both approaches use channel-wise statistics to update learnable parameters based on target domain data, they differ in how they handle statistics for prediction: the first approach relies on the current batch statistics, whereas the second continuously updates statistics at each time step. However, estimating statistics from individual streaming instances (batch size of 1) can ruin BN statistics [13]. In addition, estimating statistics from correlated samples may introduce a bias toward classes that dominate the distribution [11], [12].

To address this issue, we update only the learnable parameters within the BN layers using input test samples, while preserving the pretrained statistics from the multiple source subjects. This approach marks a key distinction from prior TTA studies, which typically adjust both parameters and statistics during adaptation. Given the feature, our proposed BN layers compute:

$$\text{BN}_{\text{ours}}(\mathbf{z}_{:, \mathbf{c}, :, :}; \mu_c^S, (\sigma_c^S)^2) = \gamma_c \frac{\mathbf{z}_{:, \mathbf{c}, :, :} - \mu_c^S}{\sqrt{(\sigma_c^S)^2 + \epsilon}} + \beta_c, \quad (3)$$

where  $\mu_c^S$  and  $(\sigma_c^S)^2$  indicate source pretrained statistics, which are fixed during the adaptation process.

### C. Entropy Minimization with Energy-bounded Loss

We optimize the BN parameters using the entropy minimization, which is widely used in TTA to adapt to the target distribution [9], [11], [12]:

$$\mathcal{L}_{\text{ent}} = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^C p_{i,k} \log p_{i,k}, \quad (4)$$

where  $p_{i,k}$  denotes the probability that an instance belongs to class  $k$ .

In addition to the entropy minimization, we introduce the energy-bounded loss as a regularization term:

$$E(\mathbf{x}; f; \tau) = -\log \sum_{k=1}^C e^{f_k(\mathbf{x})/(\tau^2)}, \quad (5)$$

$$\begin{aligned} \mathcal{L}_{\text{energy}} &= \mathbb{E} [\max(0, E(\mathbf{x}; f) - m_{in})]^2 \\ &\quad + \mathbb{E} [\max(0, m_{out} - E(\mathbf{x}'; f))]^2, \end{aligned} \quad (6)$$

where  $\mathbf{x}$  represents the input data,  $\mathbf{x}'$  denotes the augmented data used to simulate domain shifts,  $E(\cdot)$  is the energy function,  $\tau$  is the temperature scaling factor (set to 1 in energy-bounded loss),  $m_{in}$  is the margin for in-distribution samples,  $m_{out}$  is the margin for out-of-distribution samples,  $f$  represents the neural network, and  $C$  is the total number of classes.

The energy-bounded loss encourages the model to treat the target samples as in-distribution by assigning lower energy scores to them [26]. To explicitly model potential distribution shifts, we apply augmentations to the test samples. Specifically, the loss penalizes original target samples whose energy scores exceed the threshold  $m_{in}$  and augmented (shifted) samples whose scores fall below the threshold  $m_{out}$ .

The parameters are optimized using the following comprehensive objective function:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{ent}} \mathcal{L}_{\text{ent}} + \lambda_{\text{eng}} \mathcal{L}_{\text{energy}}, \quad (7)$$

where  $\lambda_{\text{ent}}$  and  $\lambda_{\text{eng}}$  denote the weights for balancing loss.

### D. Memory Bank

The reliability of normalization statistics is highly dependent on the batch size of the input data [13]. Using a smaller

batch size may hinder the network's ability to capture the target distribution [27], posing significant challenges for practical driver drowsiness detection systems. Specifically, in an EEG-based driver drowsiness detection system, estimating and updating statistics can be affected by the individual features  $\mathbf{z} \in \mathbb{R}^{1 \times C \times H \times W}$ . Furthermore, the correlation among target instances presents discrepancies between the observed distribution and the genuine target distribution, leading to inaccurate statistics of the BN layers [11], [12].

To address this challenge, we propose a memory bank with a novel removal criterion that utilizes the energy score and persistence time. When the first input arrives, instead of waiting to fill the memory bank over time, we initialize it by applying augmentation algorithms to the input. Specifically, we apply white Gaussian noise and a permutation algorithm [28] that mixes segments of the input EEG signals to generate diverse augmented samples simulating the target distribution. When the memory bank is full, we compute the negative energy score  $-E(\mathbf{x})$  [26] for each sample to identify and remove unreliable samples from the memory. The energy score acts as a confidence measure, with lower scores indicating samples that the model considers less reliable or potentially out-of-distribution. In addition to the negative energy score, we further regularize it with persistence time, which indicates how long the data have been stored in the memory bank. This allows the removal mechanism to penalize not only low-confidence samples but also outdated entries that no longer reflect the current distribution. We use the persistence time as a parameter for the temperature of the negative energy score, gradually decreasing the confidence of outdated samples over time. The overall removal score is as follows:

$$S(\mathbf{x} | \mathbf{x} \in \mathcal{M}) = -E(\mathbf{x}; f, \mathcal{A}) = \log \sum_{k=1}^C e^{f_k(\mathbf{x})/(\mathcal{A}^2)}, \quad (8)$$

where  $\mathbf{x}$  indicates samples in the memory bank,  $\mathcal{M}$  represents the memory bank, and  $\mathcal{A}$  represents the persistence time.

Our proposed removal criterion selectively removes over-confident and outdated samples, allowing the memory bank to align with the current distribution of the streaming data.

### E. Prototype Learning

Prototype learning (PL) aligns with the principles of few-shot learning, where a model learns from limited labeled data. However, unlike few-shot learning, online TTA operates with only a single unlabeled target sample available at each time step. Since PL generally involves updating class center representations, it is computationally efficient and has low latency, making it suitable for real-world applications.

In our approach, we combine the PL with a memory bank that captures the current target distributions. To initialize the prototypes, we leverage the weights from a source pretrained classifier, incorporating source knowledge into the prototypes. The prototypes are created by averaging the features and pseudo-labels of the memory data using the adapted model.

To ensure reliable prototype estimation, we consider only the memory samples that are pseudo-labeled as class  $k$  and

---

### Algorithm 1: Proposed Online Test-Time Adaptation

---

```

1 Require: Streaming input EEG sample  $\mathbf{x}_t$  at time  $t$   

 $(t \leq T)$ ; Memory bank  $\mathcal{M}$  with capacity  $\mathcal{N}$ ;  

Occupancy  $\mathcal{O}$  of  $\mathcal{M}$ ; Pretrained BN statistics  

 $(\mu_c^{\mathcal{S}}, (\sigma_c^{\mathcal{S}})^2)$ ; Learnable affine parameters  $(\gamma_c, \beta_c)$  in  

BN layers (scale and shift)  

2 for  $t = 1$  to  $T$  do  

    // 1. Store current input  

3   Insert  $\mathbf{x}_t$  into  $\mathcal{M}$   

    // 2. Initialize memory bank  

4   if  $t = 1$  then  

5     while  $\mathcal{O} < \mathcal{N}$  do  

6       Generate augmentations of  $\mathbf{x}_t$  (Gaussian  

            noise & permutation)  

7       Add augmented samples to  $\mathcal{M}$   

8     end  

9     Initialize prototypes using pretrained classifier  

            weights  

10   end  

    // 3. Memory bank update  

11   Compute removal score all samples in  $\mathcal{M}$  via (8)  

12   Discard the sample with the highest removal score  

    // 4. BN parameter adaptation  

13   Preserve source pretrained statistics  $(\mu_c^{\mathcal{S}}, (\sigma_c^{\mathcal{S}})^2)$   

14   Update BN affine parameters  $(\gamma_c, \beta_c)$  via (7)  

    // 5. Prototype learning  

15   Compute pseudo prototypes  $\hat{\mathcal{P}}_k$  for each class  $k$   

16   Update each prototype  $\mathcal{P}_k$  via (11)  

    // 6. Prediction  

17   Prediction based on the similarity as in (12)  

18 end

```

---

have energy scores above the threshold  $m_{out}$ . Let  $\mathcal{M}_k$  denote the set of such samples:

$$\mathcal{M}_k = \{\mathbf{x}_i \in \mathcal{M} \mid \hat{y}_i = k, E(\mathbf{x}_i) > m_{out}\}, \quad (9)$$

where  $\hat{y}_i$  is the pseudo-label assigned to sample  $\mathbf{x}_i$ .

The pseudo prototype  $\hat{\mathcal{P}}_k$  for class  $k$  is computed by averaging the feature representations  $\mathbf{z}_i$  of all samples in  $\mathcal{M}_k$ :

$$\hat{\mathcal{P}}_k = \frac{1}{|\mathcal{M}_k|} \sum_{\mathbf{x}_i \in \mathcal{M}_k} \mathbf{z}_i. \quad (10)$$

We then update the final prototype  $\mathcal{P}_k$  using exponential moving average (EMA):

$$\mathcal{P}_k = \alpha \mathcal{P}_k + (1 - \alpha) \hat{\mathcal{P}}_k, \quad (11)$$

where  $\alpha$  is the smoothing factor that controls the update rate.

Finally, the prediction is made by measuring the similarity between the current input feature  $\mathbf{z}_t$  and each class prototype  $\mathcal{P}_k$  via dot-product attention:

$$\arg \max_k P(y = k \mid \mathbf{z}_t) = \frac{\exp(\mathbf{z}_t \cdot \mathcal{P}_k)}{\sum_j \exp(\mathbf{z}_t \cdot \mathcal{P}_j)}. \quad (12)$$

## IV. EXPERIMENTS

### A. Materials

We used the preprocessed version of the sustained-attention driving dataset from a previous study [7], [29]. The original dataset comprises 62 sessions collected from 27 healthy participants, each of whom engaged in a 90-minute driving task using a virtual reality (VR) driving simulator. During the experiment, participants were instructed to maintain their vehicle at the center of the lane, while the system introduced unexpected lane-departure events to simulate a non-ideal road surface that caused the car to drift left or right. Participants were required to respond immediately by steering the vehicle back to the center lane, simulating the continuous vigilance required in real driving. The durations of lane departure, steering, and return to the center of the lane are referred to as deviation onset, response onset, and response offset, respectively. In the absence of camera-based measures (e.g., eye closure), these lane-departure events served as the primary behavioral probes to quantify drowsiness via reaction times. Furthermore, occurring at randomized intervals, these events established a monotonous task environment designed to induce passive drowsiness throughout the 90-minute session. Each lane-departure event was treated as a distinct trial and segmented accordingly. Experimental paradigms of the sustained-attention driving dataset are illustrated in Fig. 3.

EEG data were continuously recorded at 500 Hz using a wired cap with 30 EEG electrodes and 2 reference electrodes, which were placed according to a modified international 10–20 system. EEG signals were preprocessed with a 1–50 Hz band-pass filter. To address the contamination of EEG signals by eye blinks, the recordings were manually inspected, and segments containing apparent blink artifacts were visually rejected. Ocular and muscular artifacts were further removed using the Automatic Artifact Removal (AAR) plugin of EEGLAB. The preprocessed EEG data were then downsampled to 128 Hz, and 3-second segments were extracted prior to each deviation onset to capture neural activity preceding behavioral responses.

To ensure data quality and class balance, additional filtering steps were performed as follows. First, sessions with less than 50 samples in either class were excluded. Second, if there were multiple sessions from the same subject, the session with the most balanced class distribution was selected. Both raw and preprocessed datasets are publicly available.

According to prior studies [6], [7], [30]–[32], we followed two metrics for each sample. Local reaction time (RT) is the duration taken to respond to a lane deviation event. Global RT is the average of RTs recorded within the 90-second window before the lane departure event. Samples with both local RT and global RT shorter than 1.5 times the alert RT were classified as alertness, whereas those with both longer than 2.5 times the alert RT were classified as drowsiness. To ensure the balance and integrity of the data, sessions with fewer than 50 samples in either class were excluded. The session with the most balanced class distribution was selected if multiple sessions were available for the same participant.

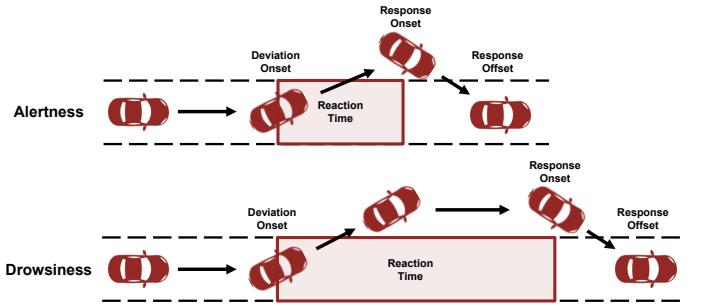


Fig. 3. Experimental paradigms of the sustained-attention driving dataset were conducted in the following sequence for each trial: response offset, deviation onset, and response onset. Drowsy drivers exhibited longer RTs compared to alert drivers.

### B. Baselines

We established two categories of baselines for comparison. We referred to the pretrained models evaluated directly on the target subject without any adaptation methods as “Source”. “Source” includes DeepConvNet [22], EEGNet8,2 [23], ResNet1D-8 [8], InterpretableCNN [7], EEG-Conformer [24], CAGNN [17], and LGGNet-F [19].

Specifically, we adopted a diverse set of “Source” baselines that use conventional CNN-based architectures (DeepConvNet [22], EEGNet8,2 [23], ResNet1D-8 [8], InterpretableCNN [7]), graph-based architecture (CAGNN [17]), and advanced Transformer variants (LGGNet [19], EEG-Conformer [24]). Some of them have been previously validated on the same sustained-attention driving dataset in their respective publication. These models serve as strong baselines that reflect both foundational and state-of-the-art approaches in EEG-based BCI systems.

For “TTA” baselines, we selected methods that represent a broad range of TTA strategies. Specifically, BN adaptation approaches (Tent [9], SAR [13]), memory-based and augmentation-aware methods (CoTTA [10], NOTE [12], RoTTA [11]), prototype-based classifiers (T3A [21]), and EEG-specific online TTA (Wimpff *et al.* [14]) were evaluated.

Each TTA method was implemented to adapt the same source-pretrained EEGNet8,2 [23] architecture to unseen target subjects under identical hyperparameter settings and evaluation protocols, ensuring a fair and meaningful comparison.

### C. Implementation Details

To evaluate the robustness of our TTA methods, we established a sophisticated experimental setup that contained both pretraining and inference stages. Our experiments employed leave-one-subject-out (LOSO) cross-validation [33], where one subject was designated as the target subject for inference, while the others were used for pretraining. This process was repeated for each subject.

To guide the experimental design of this study, we selected benchmark models from the reviewed literature that satisfy two key criteria. First, public availability of the source code. Second, compatibility with the subject-independent LOSO cross-validation protocol used in our EEG-based drowsiness detection task. These choices ensure reproducibility and fairness in performance comparison.

TABLE I  
OVERALL F1-SCORE (%) COMPARISON OF DRIVER DROWSINESS DETECTION ON A SUSTAINED-ATTENTION DRIVING DATASET

Category	Method	S1*	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	Avg. (std.)
Source	DeepConvNet [22]	81.44	8.66	56.13	62.50	70.71	81.00	80.79	62.61	77.19	49.32	60.91	62.84 (21.06)
	EEGNet8,2 [23]	73.86	29.52	62.04	68.97	76.04	87.04	83.02	64.23	77.23	52.50	65.12	67.23 (15.96)
	ResNet1D-8 [8]	81.97	33.46	50.98	66.02	52.63	73.91	80.57	67.12	77.66	60.98	75.76	65.55 (15.00)
	InterpretableCNN [7]	67.47	19.33	64.23	66.67	80.00	73.85	83.81	65.59	80.79	54.55	75.76	66.55 (17.89)
	EEG-Conformer [24]	72.39	31.03	42.42	58.49	73.24	84.47	77.55	63.32	75.09	62.65	74.45	65.01 (16.01)
	CAGNN [17]	77.97	31.18	26.67	49.75	26.47	68.37	52.90	48.30	65.43	65.12	65.99	52.56 (17.98)
	LGGNet-F [19]	67.00	<b>36.69</b>	49.61	57.97	68.42	68.75	74.87	63.59	77.56	53.33	71.91	62.70 (12.25)
TTA	Tent [9]	3.51	28.21	1.10	12.00	0.00	7.75	1.83	0.00	0.00	11.02	5.59	6.46 (8.43)
	CoTTA [10]	16.82	27.69	1.10	19.78	4.96	4.69	0.00	9.94	3.75	16.36	10.07	10.47 (8.75)
	T3A [21]	75.98	34.11	67.13	<b>69.70</b>	79.40	87.11	82.63	70.50	81.90	51.22	70.27	70.00 (15.39)
	NOTE [12]	40.25	9.26	32.62	57.14	24.05	67.89	56.97	29.91	1.10	45.21	33.14	36.14 (20.34)
	RoTTA [11]	68.97	5.80	62.59	49.03	83.61	89.27	83.02	46.67	20.41	0.00	65.18	52.23 (31.30)
	SAR [13]	39.05	6.11	56.15	65.85	77.25	88.29	31.01	66.14	75.72	51.95	48.42	55.09 (23.56)
	Wimpff <i>et al.</i> [14]	44.44	25.36	42.54	40.00	45.19	54.82	46.84	40.43	46.26	31.06	45.28	42.02 (7.98)
Ours		<b>94.68</b>	28.33	<b>95.63</b>	64.49	<b>95.73</b>	<b>94.42</b>	<b>84.51</b>	<b>87.29</b>	<b>88.66</b>	<b>75.27</b>	<b>90.04</b>	<b>81.73 (20.16)</b>

\* S: Subject

During the pretraining phase, we utilized the Adam optimizer with a learning rate of 0.001. We set the batch size to 32 and trained the model for 100 epochs. This process was performed for each target subject following a LOSO cross-validation.

An essential component of our proposed framework was the configuration of the BN layers after pretraining. As described in Section III-C, we ensured that the BN layers do not alter the normalization statistics pretrained with multiple source subjects, while focusing on updating only the learnable parameters.

During the adaptation process with the streaming target data, we used the source pretrained EEGNet8,2 [23], which showed the highest performance among all source baselines without using any adaptation methods. We utilized the AdamW optimizer [34] with a learning rate of 0.001 and a weight decay of 0.1. We used a test batch size of one and set the adaptation step to one, minimizing the latency of the adaptation process. We set the margin values  $m_{in}$  and  $m_{out}$  in (6) to -15 and -7, respectively. Additionally, we set the smoothing factor  $\alpha$  to 0.9 to facilitate the update of the adaptive prototypes. Finally, we set the balance weights  $\lambda_{ent}$  and  $\lambda_{eng}$  in (7) to 2 and 0.01, respectively.

For the TTA baselines, we used the hyperparameter values reported in their respective publications or those provided in their official implementation code. We set the memory bank capacity to 16 for all TTA baselines that used a memory bank, based on preliminary experiments where we tested capacities from 8 to 64 and found 16 to yield the best performance.

We conducted experiments on an NVIDIA RTX 3090 GPU.

#### D. Results

In this study, we assessed the effectiveness of our proposed framework against the source baselines and various TTA baselines on a sustained-attention driving dataset [7]. We evaluated all methods using the F1-score metric, which represents the harmonic mean of precision and recall. F1-score is particularly useful in scenarios with imbalanced label

distributions. In Table I, we evaluated the performances of our proposed method compared to source baselines and TTA baselines on a sustained-attention driving dataset.

The “Source” methods achieve an average F1-score ranging from 62.84% to 67.23%. Among the TTA methods, Tent [9] shows a marked performance degradation, achieving an average F1-score of 6.46%. Due to its dependence on mini-batch statistics for individual input instances, this causes model collapse across all target subjects [13]. CoTTA [10] obtains an average F1-score of 10.47%. Its performance is very low across all target subjects, indicating issues with normalization statistics when using only input instances. T3A [21], designed for online TTA using PL, achieves an average F1-score of 70%, demonstrating the potential of PL for driver drowsiness detection. However, storing all past data in T3A can lead to memory and computational constraints and may retain outdated information as the target distribution evolves. NOTE [12] shows a decrease in performance with an average F1-score of 36.14%. This indicates that its proposed prediction-balanced memory bank failed to mitigate the label distribution shifts in the driver drowsiness detection task, due to the dominance of certain classes over time. RoTTA [11] performs better with an average F1-score of 52.23%. However, it still does not surpass the performance of the source models that did not attempt to adjust to the target distribution. This performance decrease is attributed to the updating of global statistics based on the target data, which can introduce noise and reduce robustness. Although SAR [13] was designed for online TTA by introducing batch-agnostic normalization layers that prevent a collapse of normalization statistics, SAR decreases performance with an average F1-score of 55.09% even after the adaptation process. The online TTA method proposed by Wimpff *et al.* [14] shows a decreased performance with an average F1-score of 42.02%, indicating that using the target statistics ruins the normalization statistics of the BN layers.

Our proposed method notably outperforms all other base-

TABLE II  
PERFORMANCE COMPARISON OF VARYING ABLATION SETTINGS

Method	F1-score	AUROC	Precision	Recall
EEGNet8,2 (source)	67.23	73.18	74.90	67.46
Ours w/o BN updates	69.23	75.66	<b>80.22</b>	68.77
Ours w/o memory and PL	75.06	77.84	70.92	86.95
Ours w/o PL	75.84	78.40	71.32	87.84
Ours	<b>81.73</b>	<b>83.83</b>	79.75	<b>91.06</b>

TABLE III  
COMPARISON OF EXISTING BN CONFIGURATIONS

Method	Track Statistics	Initialize Statistics	F1-score	AUROC	Precision	Recall
Tent [9]			6.46	46.31	22.73	9.11
CoTTA [10]	x	✓	10.47	50.15	48.72	9.00
Wimpff <i>et al.</i> [14]			42.02	52.74	48.20	41.26
NOTE [12]	✓	x	36.14	53.65	46.53	33.50
RoTTA [11]			52.23	65.14	59.79	49.09
Ours	x	✓	0.00	48.69	0.00	0.00
Ours	✓	x	40.72	52.29	43.78	42.98
Ours	x	x	<b>81.73</b>	<b>83.83</b>	79.75	<b>91.06</b>

lines, achieving the highest average F1-score of 81.73%. This shows an average improvement of approximately 11% over the best performance among the TTA baselines. This significant improvement validates the effectiveness of our approach in a strictly calibration-free setting. By accounting for the non-stationary and temporally correlated nature of EEG signals, our framework enables robust adaptation to target subjects, effectively overcoming the domain shifts that typically challenge subject-independent models. The average per-sample inference time was measured at 15.94 ms.

### E. Ablation Study

We performed an ablation study to evaluate the individual contribution of each component by removing it. The “EEGNet8,2 (source)” configuration serves as the baseline. The performance of the ablation study for our proposed method is illustrated in Table II.

In the scenario where we did not update the BN layers, referred to as “Ours w/o BN updates”, we evaluated the proposed PL method based on the memory bank for TTA. This configuration demonstrates a performance improvement over the “EEGNet8,2 (source)” model, highlighting the effectiveness of PL for driver drowsiness detection in TTA settings. When we update only the learnable parameters of the BN layers, termed as “Ours w/o memory and PL”, we observe significant increases in F1-score and recall over the source baseline. This indicates that the BN layers can effectively adapt to even a single target instance of the target domain by preserving the source normalization statistics, showing their robustness against smaller batch sizes and distribution shifts. “Ours w/o PL” shows slight improvements across all metrics compared to “Ours w/o memory and PL”. This indicates that introducing the memory bank to update the BN layers enhances stability, compared to updating the model based on individual instances.

Our proposed framework exhibits significant performance improvements across all metrics. Notably, the F1-score increases to 81.73%, which highlights the strong capability of the model to discriminate between the driver’s mental states under domain shift environments.

### F. Normalization Configurations of the BN Layers

In Table III, we compared the performance of the proposed method when adapting the various normalization configurations of the BN layers that were employed in the existing TTA baselines [9]–[12], [14]. First, we configured the BN layers to normalize the input samples using only input batch statistics. Then, we transformed the input with updated learnable parameters for prediction. However, as illustrated in Table III, this setup showed the most severe performance degradation for the F1-score to zero, indicating that the model collapses over all time steps. This was in the same vein as Tent [9] and CoTTA [10] with this setup that showed the worst performance compared to the other baselines, as presented in Table I.

Second, we modified the BN layers to track normalization statistics without initializing them. Subsequently, we normalized the input samples with updated statistics and transformed them with updated learnable parameters for prediction. Compared to the first configuration, this setup showed a performance improvement across all metrics. However, since the BN layers updated the normalization statistics based on the statistics of individual samples, the performance was still poor, which was consistent with the claim of [13].

Lastly, we configured the BN layers with our proposed BN configuration, which used the fixed normalization statistics pretrained with the multiple source subjects. We normalized the input samples with fixed normalization statistics and transformed them with updated learnable parameters for prediction. This setup showed the best performance compared to the other configurations, which indicates that using fixed normalization statistics can enhance stability and adaptability when adapting to the target distribution for driver drowsiness detection systems.

### G. Configuration of the BN Layers for the TTA Baselines

To demonstrate the effectiveness of the proposed BN configuration, we applied it to the existing TTA baselines that involve updating the BN layers [9]–[12], [14]. Tent [9], CoTTA [10], Wimpff *et al.* [14] normalized the input target data with estimated channel-wise statistics of them. On the other hand, NOTE [12] and RoTTA [11] updated the normalization statistics.

As shown in Fig. 5, the overall performance increases significantly after applying our proposed configuration for the BN layers. Specifically, Tent [9] and CoTTA [10] not only significantly boost their average F1-scores but also surpass the average F1-score of the source baseline model. In addition, NOTE [12] and Wimpff *et al.* [14] also show substantial improvements when applying our configuration. In contrast, RoTTA [11] does not exhibit performance improvements, indicating that its removal criterion fails to filter out noisy and outdated samples effectively.

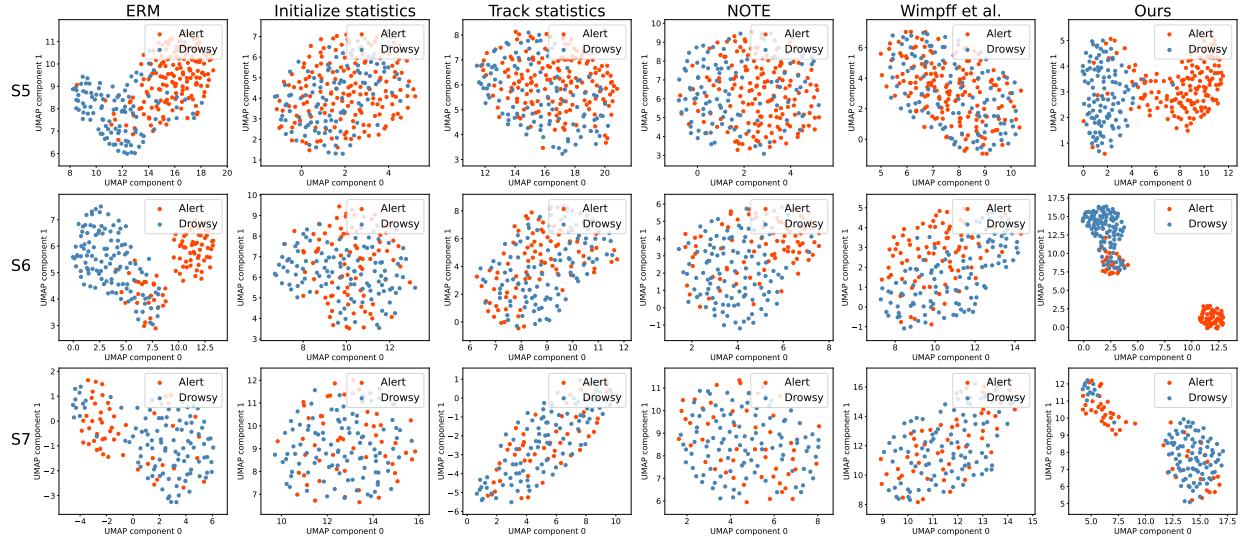


Fig. 4. UMAP feature visualization from target subjects. The row and column indicate the target subjects and adaptation methods, respectively. The first column shows the features from the ERM (no adaptation) model. The second and third columns show the features when adapting the model using existing BN configurations. The fourth and fifth columns show the features when adapting the model using TTA methods. The last column shows the features of our proposed method. We show that the proposed method significantly enhances performance and disentanglement.

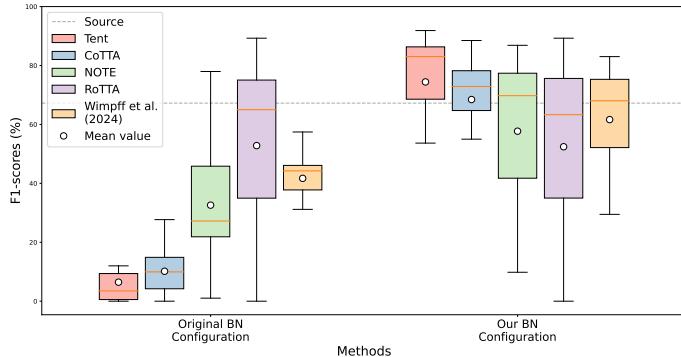


Fig. 5. Performance comparison when applying our BN configuration to the existing TTA baselines. The gray horizontal line denotes the average F1-score of the source pretrained EEGNet8.2 without adaptation methods. Existing baselines with their original BN configuration are presented on the left, while the baselines with our BN configuration are shown on the right. The overall performance of the existing TTA baselines significantly improves after manipulating the BN layers.

The significant performance improvements across different TTA methods highlight the adaptability and effectiveness of the proposed method. Overall, our findings indicate that the configuration of the BN layers is crucial for improving the performance and robustness to distribution shifts in EEG-based driver drowsiness detection framework.

## V. DISCUSSION

### A. Feature Representation Analysis of Adaptation Methods

In Fig. 4, we present a UMAP visualization of features from target subjects (S5, S6, and S7) to compare different adaptation methods. Each row corresponds to a target subject, and each column represents a specific adaptation method. The first column, showing the results of the empirical risk minimization

(ERM) model without adaptation, reveals a significant overlap between the alert and drowsy states.

In the second and third columns, we present features adapted using existing BN configurations, labeled “Initialize statistics” and “Track statistics”. Even after adapting the model to the target subjects, the boundaries between alert and drowsy states remain unclear, with substantial ambiguity between the classes. The fourth and fifth columns present results from the TTA methods NOTE [12] and Wimpff *et al.* [14]. These methods show moderate improvement in separating states; however, noticeable regions of overlap persist. Finally, in the last column, we show the features from our proposed method, where the separation between the alert and drowsy states is significantly more distinct than in the other approaches. This clear distinction suggests that our method improves performance and feature disentanglement more effectively than existing methods, indicating its greater suitability for addressing the challenges of driver drowsiness detection.

### B. Limitation and Future Work

The proposed framework demonstrates significant improvements in driver drowsiness detection within TTA environments. However, some limitations remain. First, this study was conducted in a controlled simulator, which lacks real-world factors such as vehicle vibrations, changing weather, and external noise. Future work should validate the framework in real-world driving environments to ensure robustness against environmental noise and diverse sensory stimuli. Second, we relied on binary labels derived from reaction time thresholds. Since drowsiness is inherently a continuous physiological phenomenon, this simplification may not fully capture transitional phases. Future studies could utilize fine-grained labels (e.g., Karolinska Sleepiness Scale [35]) and explore regression or ordinal classification to better model the gradual progression

of drowsiness. Third, we employed a fixed 3-second time window for EEG segments. Future investigations could analyze the impact of varying window sizes to optimize the trade-off between detection latency and information content. Finally, while this study focuses on drowsiness, our calibration-free framework is extensible to other EEG tasks [36]–[38]. Future research could explore its adaptability to broader BCI applications and wearable devices [4], [39] in low-resource settings.

## VI. CONCLUSION

We proposed a novel driver drowsiness detection framework utilizing online TTA methods, addressing the variability of EEG signals. We update only the learnable parameters of the BN layers with entropy minimization regularized by energy-bounded loss, while preserving the pretrained normalization statistics. The memory bank helps maintain a dynamic view of the target domain by using a novel removal criterion based on the energy function and persistence time. In addition, we integrate prototype learning for robust predictions against distribution shifts. Extensive experiments validated the robustness of our proposed framework against the distribution shifts and demonstrated the effectiveness and reliability of online TTA methods for EEG-based driver drowsiness detection systems.

## REFERENCES

- [1] G. Li and W.-Y. Chung, “Combined EEG-gyroscope-tDCS brain machine interface system for early management of driver drowsiness,” *IEEE Trans. Hum.-Mach. Syst.*, vol. 48, no. 1, pp. 50–62, 2017.
- [2] T. G. Monteiro, C. Skourup, and H. Zhang, “Using EEG for mental fatigue assessment: A comprehensive look into the current state of the art,” *IEEE Trans. Hum.-Mach. Syst.*, vol. 49, no. 6, pp. 599–610, 2019.
- [3] A. Giorgi *et al.*, “Driving fatigue onset and visual attention: An electroencephalography-driven analysis of ocular behavior in a driving simulation task,” *Behav. Sci.*, vol. 14, no. 11, p. 1090, 2024.
- [4] V. Ronca *et al.*, “Validation of an EEG-based neurometric for online monitoring and detection of mental drowsiness while driving,” in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2022, pp. 3714–3717.
- [5] H. Fujisawa, H. Sako, Y. Okada, and S.-W. Lee, “Information capturing camera and developmental issues,” in *Proc. Int. Conf. Doc. Anal. (ICDAR)*, 1999, pp. 205–208.
- [6] J. R. Paulo, G. Pires, and U. J. Nunes, “Cross-subject zero calibration driver’s drowsiness detection: Exploring spatiotemporal image encoding of EEG signals for convolutional neural network classification,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 905–915, 2021.
- [7] J. Cui, Z. Lan, O. Sourina, and W. Müller-Wittig, “EEG-based cross-subject driver drowsiness recognition with an interpretable convolutional neural network,” *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 7921–7933, 2022.
- [8] D.-Y. Kim, D.-K. Han, J.-H. Jeong, and S.-W. Lee, “EEG-based driver drowsiness classification via calibration-free framework with domain generalization,” in *Proc. IEEE Int. Conf. Syst. Man Cybern. (SMC)*, 2022, pp. 2293–2298.
- [9] D. Wang *et al.*, “Tent: Fully test-time adaptation by entropy minimization,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–12.
- [10] Q. Wang, O. Fink, L. Van Gool, and D. Dai, “Continual test-time domain adaptation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 7201–7211.
- [11] L. Yuan, B. Xie, and S. Li, “Robust test-time adaptation in dynamic scenarios,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 15 922–15 932.
- [12] T.-S. Gong *et al.*, “Note: Robust continual test-time adaptation against temporal correlation,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 27 253–27 266.
- [13] S. Niu *et al.*, “Towards stable test-time adaptation in dynamic wild world,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023, pp. 1–14.
- [14] M. Wimpff, M. Döbler, and B. Yang, “Calibration-free online test-time adaptation for electroencephalography motor imagery decoding,” in *Proc. IEEE Int. Winter Conf. Brain-Comput. Interface (BCI)*, 2024, pp. 1–6.
- [15] T. Mao *et al.*, “Online test-time adaptation for patient-independent seizure prediction,” *IEEE Sens. J.*, vol. 23, no. 19, pp. 23 133–23 144, 2023.
- [16] G.-D. Jang, D.-K. Han, and S.-W. Lee, “Test-time adaptation for EEG-based driver drowsiness classification,” in *Int. Conf. Pattern Recognit. Artif. Intell. (ICPRAI)*, 2024.
- [17] Z. Zhuang *et al.*, “A connectivity-aware graph neural network for real-time drowsiness classification,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 83–93, 2023.
- [18] P. Gong *et al.*, “TFAC-Net: A temporal-frequenital attentional convolutional network for driver drowsiness recognition with single-channel EEG,” *IEEE Trans. Intell. Transp. Syst.*, pp. 1–13, 2024.
- [19] Y. Ding *et al.*, “LGGNet: Learning from local-global-graph representations for brain–computer interface,” *IEEE Trans. Neural Netw. Learn. Syst.*, 2023.
- [20] J. Cui *et al.*, “Subject-independent drowsiness recognition from single-channel EEG with an interpretable CNN-LSTM model,” in *Int. Conf. on Cyberworlds (CW)*, 2021, pp. 201–208.
- [21] Y. Iwasawa and Y. Matsuo, “Test-time classifier adjustment module for model-agnostic domain generalization,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 2427–2440.
- [22] R. T. Schirrmeister *et al.*, “Deep learning with convolutional neural networks for EEG decoding and visualization,” *Hum. Brain Mapp.*, vol. 38, no. 11, pp. 5391–5420, 2017.
- [23] V. J. Lawhern *et al.*, “EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces,” *J. Neural Eng.*, vol. 15, no. 5, p. 056013, 2018.
- [24] Y. Song, Q. Zheng, B. Liu, and X. Gao, “EEG Conformer: Convolutional transformer for EEG decoding and visualization,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 710–719, 2023.
- [25] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 448–456.
- [26] W. Liu, X. Wang, J. Owens, and Y. Li, “Energy-based out-of-distribution detection,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 21 464–21 475.
- [27] T. Duan *et al.*, “Online continual decoding of streaming EEG signal with a balanced and informative memory buffer,” *Neural Netw.*, vol. 176, p. 106338, 2024.
- [28] X. Jiang, J. Zhao, B. Du, and Z. Yuan, “Self-supervised contrastive learning for EEG-based sleep staging,” in *Proc. IEEE Int. Jt. Conf. Neural Netw. (IJCNN)*, 2021, pp. 1–8.
- [29] Z. Cao, C.-H. Chuang, J.-K. King, and C.-T. Lin, “Multi-channel EEG recordings during a sustained-attention driving task,” *Sci. Data*, vol. 6, p. 19, 2019.
- [30] C.-S. Wei, Y.-T. Wang, C.-T. Lin, and T.-P. Jung, “Toward drowsiness detection using non-hair-bearing EEG-based brain-computer interfaces,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 400–406, 2018.
- [31] R. Li, R. Gao, and P. N. Suganthan, “A decomposition-based hybrid ensemble CNN framework for driver fatigue recognition,” *Inf. Sci.*, vol. 624, pp. 833–848, 2023.
- [32] Y. Liu *et al.*, “Inter-subject transfer learning for EEG-based mental fatigue recognition,” *Adv. Eng. Inform.*, vol. 46, p. 101157, 2020.
- [33] A. M. Ray *et al.*, “A subject-independent pattern-based brain-computer interface,” *Front. Behav. Neurosci.*, vol. 9, pp. 1–15, 2015.
- [34] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.
- [35] D.-Y. Kim, D.-K. Han, J.-H. Jeong, and S.-W. Lee, “Calibration-free driver drowsiness classification with prototype-based multi-domain mixup,” *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 3, pp. 2955–2966, 2025.
- [36] M. Buzzelli, S. Bianco, and P. Napoletano, “Unified framework for identity and imagined action recognition from EEG patterns,” *IEEE Trans. Hum.-Mach. Syst.*, vol. 53, no. 3, pp. 529–537, 2023.
- [37] K. P. Thomas, N. Robinson, and V. A. Prasad, “Separability of motor imagery directions using subject-specific discriminative eeg features,” *IEEE Trans. Hum.-Mach. Syst.*, vol. 51, no. 5, pp. 544–553, 2021.
- [38] M.-C. Roh, H.-K. Shin, and S.-W. Lee, “View-independent human action recognition with volume motion template on single stereo camera,” *Pattern Recognit. Lett.*, vol. 31, no. 7, pp. 639–647, 2010.
- [39] S.-W. Lee, *Advances in Biometrics: International Conference, ICB 2007, Seoul, Korea, August 27–29, 2007, Proceedings*. Springer Science & Business Media, 2007, vol. 4642.