Fully Test-time Adaptation for Tabular Data

Zhi Zhou[1, *]**, Kun-Yang Yu**[1, 2, *]**, Lan-Zhe Guo**[1, 3, †]**, Yu-Feng Li**[1, 2, †]

[1]National Key Laboratory for Novel Software Technology, Nanjing University, China
[2]School of Artificial Intelligence, Nanjing University, China
[3]School of Intelligence Science and Technology, Nanjing University, China.
{zhouz, yuky, guolz, liyf}@lamda.nju.edu.cn

Abstract

Tabular data plays a vital role in various real-world scenarios and finds extensive applications. Although recent deep tabular models have shown remarkable success, they still struggle to handle data distribution shifts, leading to performance degradation when testing distributions change. To remedy this, a robust tabular model must adapt to generalize to unknown distributions during testing. In this paper, we investigate the problem of fully test-time adaptation (FTTA) for tabular data, where the model is adapted using only the testing data. We identify three key challenges: the existence of label and covariate distribution shifts, the lack of effective data augmentation, and the sensitivity of adaptation, which render existing FTTA methods ineffective for tabular data. To this end, we propose the *Fully Test-time Adaptation for Tabular data*, namely FTAT, which enables FTTA methods to robustly optimize the label distribution of predictions, adapt to shifted covariate distributions, and suit a variety of tasks and models effectively. We conduct comprehensive experiments on six benchmark datasets, which are evaluated using three metrics. The experimental results demonstrate that FTAT outperforms state-of-the-art methods by a margin.

Project Homepage — https://wnjxyk.github.io/FTTA
Extended Version — https://arxiv.org/abs/2412.10871

1 Introduction

Tabular data (Altman and Krzywinski 2017) plays a vital role in numerous practical applications, including economics (Salehpour and Samadzamini 2024), healthcare (Ching et al. 2018), finance (Ozbayoglu, Gudelek, and Sezer 2020), and manufacturing (Hein et al. 2017). Deep neural networks (DNNs) have recently shown remarkable success in handling tabular data, often surpassing traditional statistical methods when training and test data share the same distribution (Arik and Pfister 2021; Gorishniy et al. 2021). However, real-world applications often experience shifts in data distributions during testing, leading to significant performance degradation in existing methods (Kolesnikov 2023).

---

*These authors contributed equally.

†Corresponding author

To address distribution shifts during testing, fully test-time adaptation (FTTA) algorithms have emerged, enhancing the performance of pre-trained DNNs using only testing data. These methods are particularly designed to deal with covariate distribution shift (Wang et al. 2021), label distribution shift (Wu et al. 2021), or both (Zhou et al. 2023a), adapting the model parameters (Wang et al. 2022) or optimizing the predictions (Boudiaf et al. 2022). However, they are primarily designed for image tasks and heavily rely on image augmentation strategies (Wang et al. 2022) and image-specific data assumptions (Boudiaf et al. 2022; Zhou et al. 2023a), rendering them less effective for tabular data. As a result, fully test-time adaptation for tabular data remains underexplored, despite its significance in real applications.

To this end, we study the fully test-time adaptation problem setting for tabular data, namely *AdaTab*, which holds significant practical value (Altman and Krzywinski 2017). For example, in financial applications (Kritzman, Page, and Turkington 2012), the non-stationary financial market environment can cause significant changes in the data distribution between training and testing. For instance, shifts in the stock market can significantly affect market behavior and investor sentiment. These distribution shifts degrade the model performance and seriously affect investment decision-making and risk management, thereby leading to financial losses (Guo, Hu, and Yang 2023). The goal of the *AdaTab* problem setting is to adapt the trained deep tabular model to unknown distributions using only testing data, preventing the performance degradation caused by distribution shifts in downstream tabular applications.

In this paper, we conduct an in-depth investigation into the *AdaTab* problem. Our four observations reveal three key challenges in designing FTTA methods for tabular data: (a) Covariate and label distribution shifts exist in tabular data, but they cannot be effectively addressed by existing FTTA methods; (b) Typical augmentation for test-time adaptation is often ineffective for tabular data, limiting the ability of FTTA methods to compute consistency; (c) Adaptation is sensitive to both tasks and models for tabular data. To address these challenges, we propose a novel FTTA approach, FTAT. It comprises three essential modules: *Confident Distribution Optimizer*, *Local Consistent Weighter*, and *Dynamic Model Ensembler*, which robustly track and optimize the label distribution of predictions, adapt the model
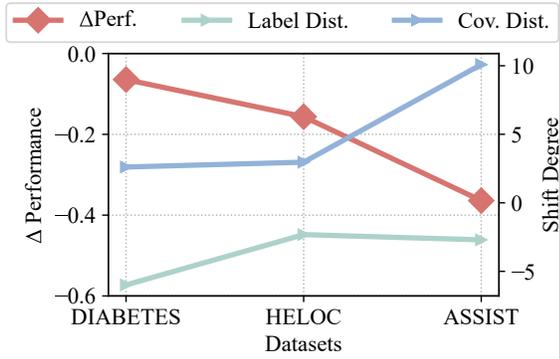
Figure 1: Label and covariate distribution shifts between training and testing in tabular data degrade model performance. Shift degree is taken logarithm for aesthetics.

to shifted covariate distribution, and dynamically adapt the model for various tasks and models. To summarize, the contributions of this paper are threefold:

(1) We investigate fully test-time adaptation for tabular data, identifying three key challenges: the existence of label and covariate distribution shifts, the lack of effective data augmentation, and the sensitivity of adaptation.

(2) We propose a novel approach, FTAT, which incorporates the *Confident Distribution Optimizer*, *Local Consistent Weighter*, and *Dynamic Model Ensembler* to address the challenges of shifted label distribution, shifted covariate distribution, and sensitivity, respectively.

(3) We evaluate the FTAT approach on six tabular benchmarks with real distribution shifts using three backbone models, demonstrating that the proposed approach significantly outperforms state-of-the-art FTTA methods.

2 Problem and Analysis

In this section, we first introduce the *AdaTab* problem setting, including the notations and problem formulation. We then present four observations of *AdaTab*, which highlight three main challenges and underscore the necessity of designing FTTA methods specifically for tabular data.

2.1 Problem Formulation

We consider the fully test-time adaptation problem setting for tabular classification problem, namely, *AdaTab*. The input space is $\mathcal{X} \in \mathbb{R}^d$, where d is the number of features. Each feature can be a continuous or discrete value. The label space is $\mathcal{Y} \in \{0,1\}^K$, where the K is the number of classes. In this setting, we are given a well-trained source tabular model $f_{\theta_0} : \mathcal{X} \mapsto \mathcal{Y}$ with the initial parameters θ_0. During the testing phase, the model is solely adapted based on the unlabeled batched testing data D_t at each timestamp t, updating its parameters from θ_t to θ_{t+1}. The goal of *AdaTab* problem is to adapt the initial given model f_{θ_0} during the testing phase, so that the adapted model f_{θ_t} can generalize better on the test data D_t at each timestamp t.
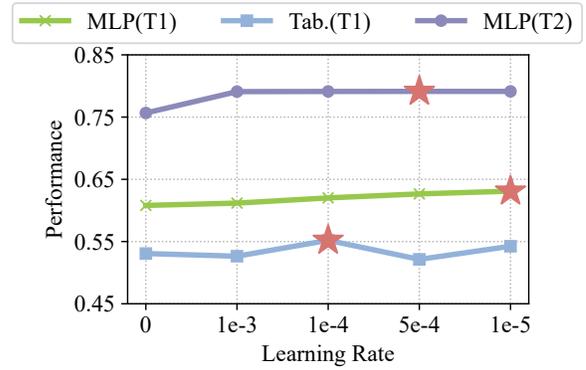

Figure 2: The performance of FTAT with different learning rates. The optimal value differs across backbones and tasks. The highest point of each line is marked by a red star.

2.2 Problem Analysis

In the context of the *AdaTab* problem, we have identified four observations that also serve as key challenges hindering FTTA methods from effectively working with tabular data, in contrast to the standard fully test-time adaptation designed for image tasks.

Observation 1: Covariate distribution and label distribution shifts in tabular data hinder performance of FTTA methods. Our first observation reveals that both covariate distribution and label distribution shifts exist in tabular data, and both contribute to performance degradation. To estimate the distribution shifts between the training and testing datasets, we use the optimal transport dataset distance with Gaussian approximation (Alvarez-Melis and Fusi 2020) to measure covariate distribution shifts and the L_2 distance (Gardner, Popovic, and Schmidt 2023) to assess label distribution shifts. As shown in Fig. 1, both increases in label distribution shift (DIABETE→HELOC) and covariate distribution shift (HELOC→ASSIST) degrade performance on the testing data with distribution shifts. However, as our experimental results reveal, existing robust FTTA methods, such as ODS (Zhou et al. 2023a), designed to address covariate and label distribution shifts in tabular data, do not perform well in practice. This observation highlights the challenges faced by FTTA methods designed for tabular data in addressing both covariate and label distribution shifts simultaneously.

Observation 2: Typical Augmentation used in test-time adaptation is ineffective for tabular data. Existing FTTA methods, such as CoTTA (Wang et al. 2022) and AdaContrast (Chen et al. 2022), rely heavily on data augmentation. However, augmentation for tabular data is not as effective as it is for images. We conduct experiments based on CoTTA methods with perturb augmentation (Fang et al. 2022) with different perturbation strengths controlled by σ. As shown in Table 1, the performance of CoTTA degrades as the augmentation strength increases, and it fails to surpasses the non-adapted source model. This observation highlights the challenges of FTTA methods designed for tabular data, particularly their inability to rely on data

	DIABETE	HELOC	ASSIST
$\sigma = 0.2$	60.46 ± 0.20	46.40 ± 3.08	54.89 ± 1.88
$\sigma = 0.4$	59.18 ± 0.42	43.36 ± 0.25	54.86 ± 3.00
$\sigma = 0.6$	57.73 ± 0.64	43.06 ± 0.07	54.51 ± 2.26
$\sigma = 0.8$	56.19 ± 0.83	43.07 ± 0.03	53.79 ± 3.80
$\sigma = 1.0$	54.74 ± 0.77	43.09 ± 0.01	54.23 ± 3.56
Source	$\mathbf{60.82 \pm 0.22}$	$\mathbf{54.37 \pm 5.35}$	$\mathbf{55.86 \pm 3.81}$

Table 1: Performance of CoTTA method with different augmentation strengths σ and the non-adapted source model using the MLP model. The best performance is in **bold**.

Method	DIABETE	HELOC	ASSIST
Source	60.82 ± 0.22	54.37 ± 5.35	55.86 ± 3.81
Opt. Parameters	61.34 ± 0.33	$\underline{54.35 \pm 5.38}$	$\underline{50.87 \pm 0.32}$
Opt. Predictions	61.47 ± 0.35	$\underline{43.10 \pm 0.00}$	$\underline{45.12 \pm 0.18}$

Table 2: Performance of the non-adapted source model and two representative FTTA methods using an MLP backbone model. Degraded performance is underlined.

augmentation method when dealing with tabular data.

Observation 3: Adaptation is sensitive to both tasks and models for tabular data. Unlike images, which exhibit strong transferability (He, Girshick, and Dollár 2019) and similar structure (Torralba and Oliva 2003), tabular data from different tasks differs significantly. Moreover, different backbone models (Gorishniy et al. 2021; Huang et al. 2020) are designed to address distinct tasks. Therefore, for specific tabular tasks and the backbone models used, FTTA methods require tuning for optimal performance. As shown in Fig. 2, the optimal learning rates for different backbone models on the same task and same backbone model on different tasks varies. This observation indicates that the *AdaTab* problem requires a model capable of dynamically tuning the learning rates, rather than relying on a fixed learning rate.

Observation 4: Existing FTTA methods degradates when dealing with tabular data. Observations 1, 2, and 3 also serve as key challenges in designing FTTA methods for tabular data, causing existing FTTA methods to fail to improve performance compared to the source model. As shown in Tab. 2, we compare performance of non-adapted source model with two representative FTTA methods which respectively optimize model parameters and predictions, i.e., TENT (Wang et al. 2021) and LAME (Boudiaf et al. 2022). As the covariate distribution and label distribution shifts become more severe (DIABETE→HELOC→ASSIST), both types of FTTA methods fail to surpass the non-adapted source model. This observation highlights the significance of developing FTTA methods specifically designed for tabular data which addressing three challenges simultaneously.

3 Methodology

In this section, we introduce our FTAT approach for *AdaTab* problem setting. As discussed in the analysis section, the

AdaTab problem encompasses three challenges:

(a) Covariate and label distribution shifts exist in tabular data, but cannot be effectively addressed by existing FTTA methods;

(b) Typical augmentation used for test-time adaptation is not very effective for tabular data, limiting the ability of FTTA methods to adopt consistency;

(c) Adaptation is sensitive to both tasks and models for tabular data.

To address the above challenges, we introduce three modules specifically designed for *AdaTab* problem, i.e., *Confident Distribution Optimizer*, *Local Consistent Weighter*, and *Dynamic Model Ensembler*. Specifically, the *Confident Distribution Optimizer* optimizes the original model predictions $f_{\theta_t}(x)$ to $\hat{f}_{\theta_t}(x)$ for a data point x at timestamp t. The *Local Consistent Weighter* affects the adaptation objective:

$$\theta_{t+1} = \arg \min_\theta \sum_{i=1}^{D_t} \mathcal{W}(x_i, D_t, \theta_t) \cdot \text{Loss}\left(\hat{f}_{\theta_t}(x_i)\right) \quad (1)$$

using a weighting function $\mathcal{W}$, where $\text{Loss}(\cdot)$ represents the unsupervised loss for test-time adaptation, and we employ entropy loss in accordance with classical methods. The *Dynamic Model Ensembler* maintains multiple models and ensembles their predictions in an online manner. We will introduce them in detail.

3.1 Confident Distribution Optimizer

First, we aim to optimize the model predictions to align with the current shifted label distribution. The existing solution (Zhou et al. 2023a) fails because the challenging nature of tabular data prevents the model from making accurate predictions, which in turn hinders the estimation of the label distribution. Therefore, the key challenge is how to robustly track the shifted label distribution $\hat{P}_t$ at each timestamp t. With original label distribution P_0 and esitmated label distribution $\hat{P}_t$, the optimized model prediction $\hat{f}_{\theta_{t+1}}(x_k)$ for next timestamp on data point x_k is

$$\hat{f}_{\theta_{t+1}}(x_k) = \frac{f_{\theta_{t+1}}(x_k) \circ \hat{P}_t}{P_0} \quad (2)$$

Motivated by our observations, we recognize that we can estimate the label distribution $\widetilde{P}_t$ with bias from model $\hat{f}_{\theta_t}$ at each timestamp t using only data with low-entropy predictions (i.e., data with confident predictions):

$$\widetilde{P}_t = \frac{\sum_{i=1}^{|D_t|} \mathbb{I}\left[\text{Entropy}\left(\hat{f}_{\theta_t}(x_i)\right) < \epsilon\right] \cdot \hat{f}_{\theta_t}(x_i)}{\sum_{i=1}^{|D_t|} \mathbb{I}\left[\text{Entropy}\left(\hat{f}_{\theta_t}(x_i)\right) < \epsilon\right]} \quad (3)$$

where D_t is current data batch, ϵ is a threshold, and $\text{Entropy}(\cdot)$ is the function for computing entropy of predictions. Note that there exists bias in esitmated $\widetilde{P}_t$ as the model predictions may contains errors. To address this issue, we compute the covariate matrix $\hat{C}_t$ at the current timestamp t,
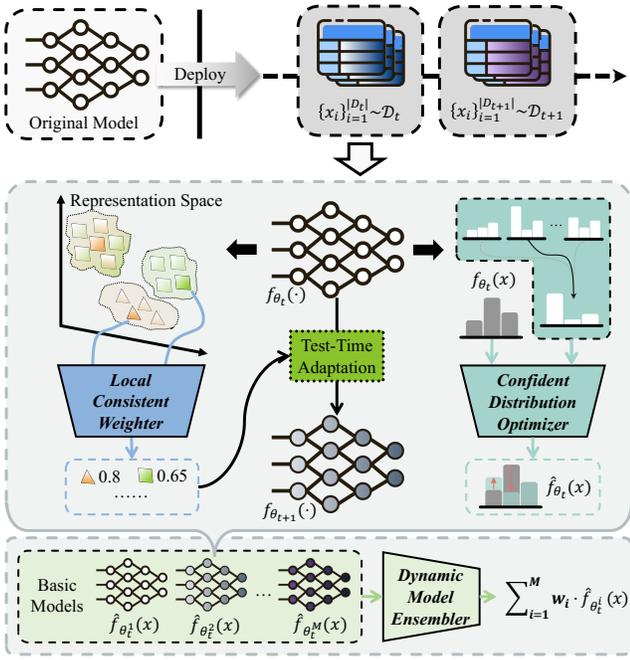
Figure 3: The overall illustation of FTAT approach.

where its k-th row is equal to

$$\frac{\sum_{i=1}^{|D_t|} \mathbb{I}\left[\arg\max_j \widehat{f}_{\theta_t}(x_i)_j = k\right] \cdot \widehat{f}_{\theta_t}(x_i)}{\sum_{i=1}^{|D_t|} \mathbb{I}\left[\arg\max_j \widehat{f}_{\theta_t}(x_i)_j = k\right]} \quad (4)$$

Then, the unbiased label distribution is $\widehat{C}_t^{-1}\widetilde{P}_t$. We additionally adopt a temperal ensemble method (Laine and Aila 2017) to robustly track estimated label distribution smoothly with a factor α and previous estimated shift $\widehat{P}_{t-1}$:

$$\widehat{P}_t = \text{Norm}\left(\widehat{P}_{t-1} - \alpha \cdot \widehat{C}_t^{-1}\widetilde{P}_t\right) \quad (5)$$

where $\text{Norm}(\cdot)$ normalizes the distribution to sum to one, and we use the Softmax function for this purpose.

3.2 Local Consistent Weighter

Second, to mitigate the adverse effects of shifted covariate distribution, we propose filtering testing data with low-quality predictions to ensure robust test-time adaptation and avoiding error accumulation. However, for tabular data, computing consistency through data augmentation is nontrivial because our observation indicates that augmentation for tabular data is not as reliable as it is for image data.

To address this issue, we propose replacing the consistency between a data point and its augmentations with the consistency between a data point and its neighborhood, under the inspiration of one existing tablar study (Gorishniy et al. 2024). Specifically, we define the neighborhood set $N(x_k, D_t)$ of each data point x_k in current batch D_t measured by one distance function $\text{Dist}(\cdot, \cdot)$:

$$N(x_k, D_t) = \left\{x \mid \text{Dist}(x, x_k) < \overline{\text{Dist}_t}, x \in D_t\right\} \quad (6)$$

where $\overline{\text{Dist}_t} = \frac{2}{|D_t|(|D_t|-1)}\sum_{i=1}^{|D_t|}\sum_{j=i+1}^{|D_t|}\text{Dist}(x_i, x_j)$ is the average pair-wise distance in D_t and we adopt L_2 distance as the distance function. Next, we define the prediction of one data point x_k is consistent if its soft pseudo-label vector is close to the average soft pseudo-label vectors in its neighborhood $N(x_k, D_t)$. Then, we define the indication function $\mathcal{I}(x_k, D_t, \theta_t)$ to decide whether one data point x_k in current batch D_t is consistent:

$$\mathcal{I}(x_k, D_t, \theta_t) = \begin{cases} 1, & \left\|f_{\theta_t}(x_k) - \frac{\sum_{x \in N(x_k, D_t)} f_{\theta_t}(x)}{|N(x_k, D_t)|}\right\| < \beta, \\ 0, & Otherwise. \end{cases}$$
$$(7)$$

where θ_t is the parameters of model at timestamp t, $f_{\theta_t}(\cdot)$ predicts pseudo-label of data point, and β is a hyperparameter to control the degree of consistency. To additionally ensure the robustness of adaptation, we compute the uncertainty of each data point using margin of prediction (Helton and Johnson 2011) as $\max \hat{f}_{\theta_t}(x_k) - \min \hat{f}_{\theta_t}(x_k)$. Finally, our proposed local consistent weighter $\mathcal{W}(x_k, D_t, \theta_t)$ is formulated as follows:

$$\left[\max \hat{f}_{\theta_t}(x_k) - \min \hat{f}_{\theta_t}(x_k)\right] \cdot \mathcal{I}(x_k, D_t, \theta_t) \quad (8)$$

3.3 Dynamic Model Ensembler

Third, to address the sensitivity issue of adaptation, we employ the online ensemble learning paradigm (Bai et al. 2022) to optimize multiple models with different learning rates and ensemble their outputs through weighted averaging to obtain the overall robust prediction. Specifically, we maintain M models using different learning rates during the testing, denoted as the set of base models $\left\{\hat{f}_{\theta_t^i}\right\}_{i=1}^M$. Then, model predictions are weighted according to corresponding loss values $w_i \propto 1 - R_t^i(D_t)$, where $R_t^i(D_t)$ is the loss value of i-th model $\hat{f}_{\theta_t^i}$ evaluated on current batched data D_t and satisfies the constraint $\sum_{i=1}^M w_i = 1$. The final prediction of the FTAT approach is obtained by weighted ensemble for a data point x, that is, $\sum_{i=1}^M w_i \cdot \hat{f}_{\theta_t^i}(x)$.

4 Experiments

In this section, we first introduce the experimental setup. Next, we present our empirical results, comparing our FTAT approach with existing FTTA methods across four benchmark datasets. Finally, we conduct an ablation study and provide further analysis for our proposed method.

4.1 Experimental Setup

Evaluation Protocol. In our experiments on tabular tasks, we follow the fully test-time adaptation setting, where the source model is trained on training data and adapted to shifted test data without any access to the source training data. Specifically, we train the source model on training data and select the best model based on the validation set following the TableShift benchmark (Gardner, Popovic, and Schmidt 2023). Then, FTAT approach and existing FTTA methods are evaluated on the shifted test set. We select six common tabular benchmark datasets from the TableShift

Method	MLP			TabTransformer			FT-Transformer		
	Acc	BAcc	F1	Acc	BAcc	F1	Acc	BAcc	F1
Source	62.5	64.6	60.6	60.9	63.1	58.3	59.7	**62.5**	54.3
TENT	58.4	61.6	51.0	58.3	61.4	51.7	51.6	55.4	36.3
EATA	61.4	63.7	60.1	60.3	62.4	60.1	56.0	59.0	44.6
LAME	59.5	62.3	58.5	59.2	62.5	58.4	58.9	62.0	51.9
CoTTA	61.6	63.8	60.6	60.4	62.8	59.8	59.6	62.4	53.4
ODS	59.2	62.2	57.8	59.2	62.0	58.5	59.1	61.7	51.4
SAR	61.2	63.5	59.2	60.3	62.8	59.7	59.3	62.1	57.0
FTAT	**66.8**	**65.0**	**72.0**	**66.1**	**64.4**	**69.0**	**64.0**	**62.5**	**69.6**

Table 3: The average performance of FTAT approach and comparison methods using three backbone models. The best performance is in **bold**. Our FTAT approach achieves the best performance across all three backbone models.

benchmark, which exhibit significant performance gaps under distribution shifts. Therfore, these datasets contains samples from 10K to 5M and features from 26 to 365, which can cover a wide range of tabular scenarios under distribution shifts. All experiments are repeated with different random seeds, and the mean and standard deviation are reported.

Comparison Methods. We first compare our approach with the non-adapted source model, referred to as Source. To compare our FTAT apporach with various FTTA methods, including typical FTTA methods (i.e., TENT (Wang et al. 2021) and EATA (Niu et al. 2022)), continual FTTA methods (i.e., CoTTA (Wang et al. 2022)), and recently proposed robust FTTA methods (i.e., SAR (Niu et al. 2023), LAME (Boudiaf et al. 2022), and ODS (Zhou et al. 2023a)).

4.2 Main Results

To evaluate the effectiveness of FTAT, we report the detailed experimental results using three backbone models in Tab. 3. The performance are measured by three metrics including accuracy, balanced accuracy, and F1 score. The experimental results show our FTAT approach outperforms existing methods by a margin on all metrics.

Moreover, we report the detailed results on each dataset using MLP backbone model in Tab. 4. FTAT achieves the best performance on major cases and give competitive performance on the resting cases, demonstrating the effectiveness of FTAT on various tabular datasets with different backbone models. ODS (Zhou et al. 2023a) addresses both the covariate and label distribution shifts, however, it underperforms FTAT on most cases, demonstrating the effectiveness of FTAT in handling distribution shifts for tabular data. The detailed results using FT-Transformer (Gorishniy et al. 2021) and TabTransformer (Huang et al. 2020) are included in our supplementary material due to the space limits.

Our experimental results confirm our last observation, showing that existing FTTA methods face performance degradation on tabular data with distribution shifts, thereby demonstrating the necessity to study the *AdaTab* problem. Our FTAT approach consistently outperforms non-adapted baseline and existing FTTA methods in most cases, offering insights into this challenging problem.
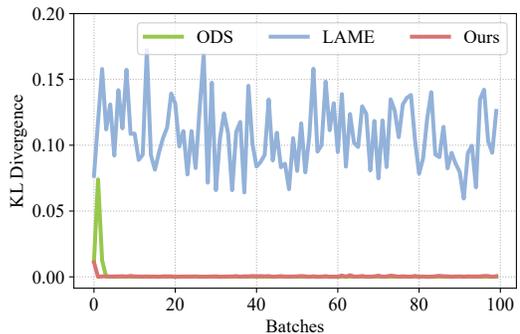

Figure 4: The performance of LAME, ODS, and FTAT in estimating label distribution evaluated using KL divergence.

4.3 Further Analysis

Ablation Study. We analyze the effectiveness of the *Confident Distribution Optimizer* (denoted as CDO) and the *Local Consistent Weighter* (denoted as LCW) in Tab. 5 on DIABETE and HELOC datasets using the MLP backbone model. The experimental results of FTAT apporach without CDO and LCW are reported measured by three metrics. Without CDO, the performance of FTAT approach improves marginally, which indicates that the label distribution shift hinders the performance and CDO addressing the label distribution shift plays a more important role in the FTAT approach. Without LCW, the performance of FTAT cannot achieve the optimal level, demonstrating the essential role of LCW to robustly update the model. Overall, our FTAT approach achieves the best performance when both CDO and LCW are employed, demonstrating their effectiveness in addressing challenges of FTTA for tabular data.

Estimation of Label Distribution. We compare the performance in estimating label distribution on the DIABETE dataset using the MLP backbone model. We adopt KL divergence to measure the distance between the ground-truth label distribution and its estimation. As shown in Fig. 4, the LAME method cannot accurately estimate the label distribution. While the ODS method can robustly track the label distribution, it requires several iterations to converge to an accurate estimation. In contrast, our FTAT approach achieves accurate label distribution estimation at a much faster speed. This result demonstrates the superiority of FTAT approach.

Effects of Dynamic Model Ensembler. To validate the effectiveness of *Dynamic Model Ensembler*, we conduct experiments running with base models with different learning rates, ensemble baseline, and FTAT approach. Here, we compare with four base models with different learning rates $\{1e-3, 1e-4, 5e-4, 1e-5\}$. The ensemble baseline are the direct ensemble of these base models. 6 presents the average performance using MLP model on DIABETE dataset. The results show that our *Dynamic Model Ensembler* module consistently outperforms the average ensemble baseline, demonstrating the its effectiveness. Moreover, the FTAT approach can achieve the best performance or competitive performance when compared to base learners with the optimal learning rate without requiring tuning, indicating the advan-

Method	HELOC			ANES			Health Ins.		
	Acc	BAcc	F1	Acc	BAcc	F1	Acc	BAcc	F1
Source	54.4 ± 5.4	58.3 ± 3.6	40.0 ± 16.8	79.1 ± 0.3	75.7 ± 0.5	$\mathbf{84.2 \pm 0.2}$	65.8 ± 0.6	$\mathbf{70.7 \pm 0.4}$	66.2 ± 0.9
TENT	54.4 ± 5.4	58.2 ± 3.6	40.0 ± 16.9	78.1 ± 0.4	74.1 ± 0.7	83.8 ± 0.1	64.3 ± 0.7	69.8 ± 0.5	63.9 ± 1.1
EATA	54.4 ± 5.4	58.3 ± 3.6	40.0 ± 16.8	78.1 ± 0.3	74.2 ± 0.6	83.8 ± 0.1	65.8 ± 0.6	70.7 ± 0.4	66.2 ± 0.9
LAME	43.1 ± 0.0	50.0 ± 0.0	30.1 ± 0.0	63.5 ± 0.0	54.6 ± 0.0	46.8 ± 0.0	63.4 ± 1.7	69.1 ± 1.1	62.6 ± 2.7
CoTTA	54.4 ± 5.4	58.3 ± 3.6	40.0 ± 16.8	78.1 ± 0.3	74.2 ± 0.6	83.8 ± 0.1	65.8 ± 0.6	$\mathbf{70.7 \pm 0.4}$	66.2 ± 0.9
ODS	43.1 ± 0.0	50.0 ± 0.0	30.1 ± 0.0	63.5 ± 0.0	54.6 ± 0.0	46.8 ± 0.0	63.5 ± 1.7	69.1 ± 1.1	62.6 ± 2.7
SAR	52.3 ± 6.1	56.7 ± 4.0	33.2 ± 19.0	78.1 ± 0.3	74.2 ± 0.6	83.8 ± 0.1	65.8 ± 0.6	$\mathbf{70.7 \pm 0.4}$	66.2 ± 0.9
FTAT	$\mathbf{64.1 \pm 1.1}$	$\mathbf{63.6 \pm 0.9}$	$\mathbf{67.8 \pm 2.7}$	$\mathbf{80.1 \pm 0.2}$	$\mathbf{79.1 \pm 0.2}$	83.4 ± 0.3	$\mathbf{72.4 \pm 0.2}$	65.3 ± 0.6	$\mathbf{80.8 \pm 0.2}$

Method	ASSIST			DIABETE			Hypertension		
	Acc	BAcc	F1	Acc	BAcc	F1	Acc	BAcc	F1
Source	55.9 ± 3.8	60.8 ± 3.4	66.4 ± 1.9	60.8 ± 0.2	60.6 ± 0.2	51.2 ± 1.7	58.8 ± 1.7	$\mathbf{61.7 \pm 1.0}$	55.5 ± 4.0
TENT	50.9 ± 0.3	56.4 ± 0.3	64.0 ± 0.2	61.3 ± 0.3	61.2 ± 0.3	53.8 ± 1.0	41.7 ± 0.1	50.1 ± 0.1	0.5 ± 0.4
EATA	55.9 ± 0.2	60.8 ± 0.2	66.4 ± 0.1	61.4 ± 0.3	61.2 ± 0.3	53.7 ± 1.1	57.8 ± 2.3	61.2 ± 1.4	52.9 ± 5.8
LAME	45.1 ± 0.2	51.3 ± 0.2	61.4 ± 0.2	61.5 ± 0.4	61.3 ± 0.4	54.7 ± 1.5	58.6 ± 1.6	61.6 ± 0.9	55.1 ± 3.8
CoTTA	55.9 ± 0.2	60.8 ± 0.2	66.4 ± 0.1	61.4 ± 0.3	61.2 ± 0.3	53.8 ± 1.1	58.8 ± 1.7	$\mathbf{61.7 \pm 1.0}$	55.5 ± 4.0
ODS	45.1 ± 0.2	51.3 ± 0.2	61.4 ± 0.2	61.5 ± 0.4	61.3 ± 0.4	54.7 ± 1.4	57.1 ± 1.5	60.8 ± 0.9	51.4 ± 3.4
SAR	55.9 ± 0.2	60.8 ± 0.2	66.4 ± 0.1	61.4 ± 0.3	61.2 ± 0.3	54.0 ± 0.9	58.2 ± 1.5	61.5 ± 0.8	53.8 ± 4.1
FTAT	$\mathbf{60.2 \pm 2.9}$	$\mathbf{63.8 \pm 2.2}$	$\mathbf{66.9 \pm 1.2}$	$\mathbf{61.7 \pm 0.3}$	$\mathbf{61.5 \pm 0.3}$	$\mathbf{59.3 \pm 1.0}$	$\mathbf{62.2 \pm 0.9}$	56.4 ± 1.6	$\mathbf{73.8 \pm 0.1}$

Table 4: Performance of FTAT approach and comparison methods on 6 datasets using MLP. The best performance is in **bold**.

DIABETE			
Method	Acc	BAcc	F1
Source	60.8 ± 0.2	60.6 ± 0.2	51.2 ± 1.7
FTAT w/o CDO	60.9 ± 0.2	60.6 ± 0.2	51.3 ± 1.7
FTAT w/o LCW	61.4 ± 0.2	61.3 ± 0.2	55.6 ± 1.7
FTAT	$\mathbf{61.7 \pm 0.3}$	$\mathbf{61.5 \pm 0.3}$	$\mathbf{59.3 \pm 1.0}$

HELOC			
Method	Acc	BAcc	F1
Source	54.4 ± 5.4	58.3 ± 3.6	40.0 ± 16.8
FTAT w/o CDO	62.7 ± 0.2	62.6 ± 0.5	66.5 ± 0.9
FTAT w/o LCW	62.5 ± 0.9	61.7 ± 1.1	67.0 ± 1.2
FTAT	$\mathbf{64.1 \pm 1.1}$	$\mathbf{63.6 \pm 0.9}$	$\mathbf{67.8 \pm 2.7}$

Table 5: Ablation study of FTAT approach on DIABETE dataset using MLP backbone model. The best performance is in **bold**. The results show that both *Confident Distribution Optimizer* (denoted as CDO) and the *Local Consistent Weighter* (denoted as LCW) are essential for FTAT approach.

Method	Acc	BAcc	F1
Lr=1e-3	61.49 ± 0.46	61.47 ± 0.42	$\mathbf{60.39 \pm 0.73}$
Lr=1e-4	61.58 ± 0.30	61.53 ± 0.26	59.37 ± 0.95
Lr=5e-4	61.56 ± 0.33	61.53 ± 0.30	59.93 ± 0.90
Lr=1e-5	61.58 ± 0.30	61.52 ± 0.26	59.23 ± 0.97
Avg. Ens.	61.57 ± 0.31	61.51 ± 0.27	59.24 ± 0.95
FTAT	$\mathbf{61.60 \pm 0.31}$	$\mathbf{61.53 \pm 0.27}$	59.27 ± 0.96

Table 6: Performance of base models with different learning rates, directly ensemble of base models, and the FTAT approach on DIABETE dataset using the MLP backbone model. The best performance is in **bold**.

tage of our *Dynamic Model Ensembler* module.

Robustness of Batch Size. In the main experiments, the batch size of the data stream is set to 512 due to the large quantity of testing data. A natural question that arises is how the batch size affects the performance of the proposed method. We conduct experiments on the DIABETE dataset using an MLP backbone model, with batch sizes set to $\{64, 128, 256, 512, 1024\}$. As shown in Fig. 5, the results indicate that the accuracy and balanced accuracy metrics are robust across different batch sizes. Regarding the F1 score, it decreases as the batch size increases. Nevertheless, FTAT consistently outperforms existing methods by a margin.

Robustness of Hyperparameters. To validate whether our proposed FTAT approach is robust to the choices of hyperparameters, we conduct hyperparameter robustness experiments on DIABETE dataset evaluated by three metrics using MLP backbone model. Specifically, FTAT contains three hyperparameters, i.e., ϵ, α and β. The hyperparameter α controls the rate at which the estimated label distribution is updated, enhancing the robustness of the FTAT to estimation errors in certain batches. The hyperparameter ϵ governs the entropy-based confident samples selection to accurately estimate the label distribution. β determines the construction of the neighbor set for entropy minimization, contributing to robust model adaptation. We conducted three runs of experiments for each set of hyperparameters with α in
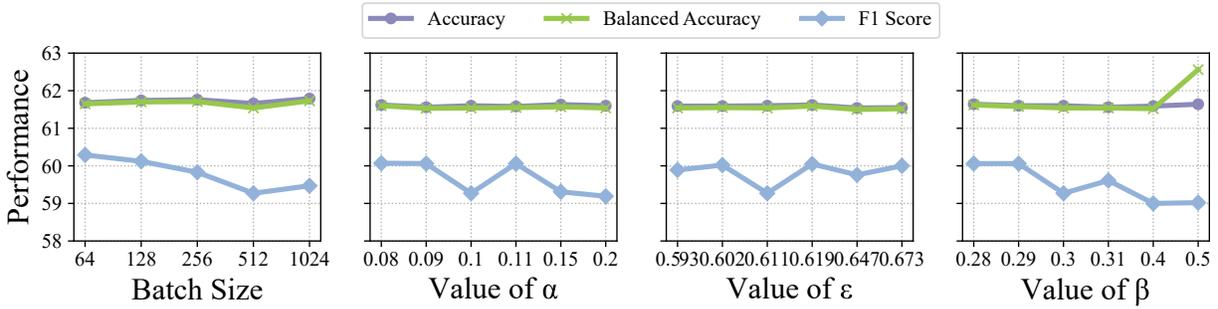
Figure 5: Robustness of batch size and hyperparameters α, ϵ, β on DIABETE dataset using MLP backbone model. The results indicate that minor perturbations to the hyperparameters of FTAT do not significantly affect its performance, demonstrating the practical robustness of FTAT.

$\{0.08, 0.09, 0.10, 0.11, 0.15, 0.20\}$, $\epsilon = \text{Entropy}([p, 1 - p])$ where p was set to $\{0.72, 0.71, 0.70, 0.69, 0.65, 0.60\}$, and β in $\{0.28, 0.29, 0.30, 0.31, 0.40, 0.50\}$. Fig. 5 reports the average performance of each hyperparameter evaluated using three metrics on DIABETE dataset. The results demonstrate that FTAT is robust to slight changes in all hyperparameters.

5 Related Work

In this section, we mainly discuss two lines of related work, including test-time adaptation and deep tabular learning.

Test-time Adaptation. Test-time adaptation (Wang et al. 2021; Zhou et al. 2023b; Zhang, Zhou, and Li 2024) aims to adapt a source model to the distribution shift in testing data without using any source data. Previously, test-time training studies, such as TTT (Sun et al. 2020) and TTT+ (Liu et al. 2021), manipulated the model in both the training and testing phases. However, when training data is inaccessible and model training cannot be controlled, test-time training paradigms become ineffective. Fully test-time adaptation aims to tackle this limitation by adapting the model without assumptions on the source model. Tent (Wang et al. 2021) updates the parameters of the BN layer at test time using entropy minimization (Li and Liang 2019). EATA (Niu et al. 2022) additionally conducts active sample selection and weighting strategies for efficiency. Other studies (Gong et al. 2022; Goyal et al. 2022) also propose diverse methods to adapt the BN layer to the test data distribution to ensure performance. In practice, SAR (Niu et al. 2023) introduces a flat minimum optimization method to ensure generalization performance when the test batch size varies. CoTTA (Wang et al. 2022) works on continually non-i.i.d. scenarios using weight-averaged models, augmentation-averaged predictions, and stochastically restoring. LAME (Boudiaf et al. 2022) proposes a conservative approach to revise the model's predictions instead of model parameters. ODS (Zhou et al. 2023a) focuses on test-time adaptation settings where covariate and label distributions change together. Recent TTA studies mainly focus on images and natural language, paying little attention to tabular data. AdapTable (Kim et al. 2024) studies test-time adaptation for tabular data, effectively designing a graph-based module to address label shifts and providing insightful theoretical analyses. TabLog (Ren et al. 2024) is the first to examine the structure of invariant rules for tabular data in the context of test-time adaptation. However, these stduies require the training data to be available, which cannot be applied in our *AdaTab* problem setting and is not practical in real-world scenarios. Therefore, our paper focuses on the fully test-time adaptation problem for tabular data, an area that remains underexplored.

Deep Tabular Learning. Deep Tabular Learning aims to model tabular data for tasks such as classification and regression through deep learning methods. Unlike image and language data, the heterogeneity and high dimensionality make it difficult for models to extract spatial and semantic information. Recently, attention-based architectures have been introduced to the tabular data domain. FT-Transformer (Gorishniy et al. 2021) applies a feature tokenizer to heterogeneous feature columns and learns an optimal representation in embedding space. Additionally, TabTransformer (Huang et al. 2020), TabNet (Arik and Pfister 2021), and other deep tabular models (Badirli et al. 2020; Klambauer et al. 2017; Gorishniy, Rubachev, and Babenko 2022; Grinsztajn, Oyallon, and Varoquaux 2022) are proposed for better representation of tabular data. However, these methods typically works well in an i.i.d. setting, and may suffer from performance degradation when the test data distribution shifts.

6 Conclusion

In this paper, we investigate the problem of fully test-time adaptation for tabular data (*AdaTab*), an important and practically valuable issue that remains underexplored. Our observations highlight three key challenges in the *AdaTab* problem: the existence of label and covariate distribution shifts, the lack of effective data augmentation, and the sensitivity of model adaptation. To address these challenges, we propose the FTAT approach, which includes three novel modules: *Confident Distribution Optimizer*, *Local Consistent Weighter*, and *Dynamic Model Ensembler*. Our experimental results demonstrate that the FTAT approach outperforms existing FTTA methods, demonstrating its effectiveness in addressing tabular tasks.

One limitation of this paper is that the design of FTAT approach lacks deep theoretical understanding and we will explore in this direction in the future to provide deep insights for the following researchers.

Acknowledgements

References

Altman, N.; and Krzywinski, M. 2017. Tabular data. *Nature Methods*, 14(4): 329–330.

Alvarez-Melis, D.; and Fusi, N. 2020. Geometric dataset distances via optimal transport. In *Advances in Neural Information Processing Systems*, 21428–21439.

Arik, S. Ö.; and Pfister, T. 2021. TabNet: Attentive Interpretable Tabular Learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 6679–6687.

Badirli, S.; Liu, X.; Xing, Z.; Bhowmik, A.; and Keerthi, S. S. 2020. Gradient Boosting Neural Networks: GrowNet. *CoRR*, abs/2002.07971.

Bai, Y.; Zhang, Y.-J.; Zhao, P.; Sugiyama, M.; and Zhou, Z.-H. 2022. Adapting to Online Label Shift with Provable Guarantees. In *Advances in Neural Information Processing Systems*.

Boudiaf, M.; Mueller, R.; Ben Ayed, I.; and Bertinetto, L. 2022. Parameter-free Online Test-time Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8344–8353.

Chen, D.; Wang, D.; Darrell, T.; and Ebrahimi, S. 2022. Contrastive Test-Time Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 295–305.

Ching, T.; Himmelstein, D. S.; Beaulieu-Jones, B. K.; Kalinin, A. A.; Do, B. T.; Way, G. P.; Ferrero, E.; Agapow, P.-M.; Zietz, M.; Hoffman, M. M.; et al. 2018. Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface*, 15(141): 20170387.

Fang, J.; Tang, C.; Cui, Q.; Zhu, F.; Li, L.; Zhou, J.; and Zhu, W. 2022. Semi-Supervised Learning with Data Augmentation for Tabular Data. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 3928–3932.

Gardner, J.; Popovic, Z.; and Schmidt, L. 2023. Benchmarking Distribution Shift in Tabular Data with TableShift. In *Advances in Neural Information Processing Systems*.

Gong, T.; Jeong, J.; Kim, T.; Kim, Y.; Shin, J.; and Lee, S.-J. 2022. NOTE: Robust Continual Test-time Adaptation Against Temporal Correlation. In *Advances in Neural Information Processing Systems*, 27253–27266.

Gorishniy, Y.; Rubachev, I.; and Babenko, A. 2022. On Embeddings for Numerical Features in Tabular Deep Learning. In *Advances in Neural Information Processing Systems*.

Gorishniy, Y.; Rubachev, I.; Kartashev, N.; Shlenskii, D.; Kotelnikov, A.; and Babenko, A. 2024. TabR: Tabular Deep Learning Meets Nearest Neighbors. In *Proceedings of th 12th International Conference on Learning Representations*.

Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; and Babenko, A. 2021. Revisiting Deep Learning Models for Tabular Data. In *Advances in Neural Information Processing Systems*, 18932–18943.

Goyal, S.; Sun, M.; Raghunathan, A.; and Kolter, J. Z. 2022. Test Time Adaptation via Conjugate Pseudo-labels. In *Advances in Neural Information Processing Systems*, 6204–6218.

Grinsztajn, L.; Oyallon, E.; and Varoquaux, G. 2022. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems*.

Guo, Y.; Hu, C.; and Yang, Y. 2023. Predict the Future from the Past? On the Temporal Data Distribution Shift in Financial Sentiment Classifications. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 1029–1038.

He, K.; Girshick, R. B.; and Dollár, P. 2019. Rethinking ImageNet Pre-Training. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, 4917–4926.

Hein, D.; Depeweg, S.; Tokic, M.; Udluft, S.; Hentschel, A.; Runkler, T. A.; and Sterzing, V. 2017. A benchmark environment motivated by industrial control problems. In *Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence*, 1–8.

Helton, J. C.; and Johnson, J. D. 2011. Quantification of margins and uncertainties: Alternative representations of epistemic uncertainty. *Reliability Engineering & System Safety*, 96(9): 1034–1052.

Huang, X.; Khetan, A.; Cvitkovic, M.; and Karnin, Z. S. 2020. TabTransformer: Tabular Data Modeling Using Contextual Embeddings. *CoRR*, abs/2012.06678.

Kim, C.; Kim, T.; Woo, S.; Yang, J. Y.; and Yang, E. 2024. AdapTable: Test-Time Adaptation for Tabular Data via Shift-Aware Uncertainty Calibrator and Label Distribution Handler.

Klambauer, G.; Unterthiner, T.; Mayr, A.; and Hochreiter, S. 2017. Self-Normalizing Neural Networks. In *Advances in Neural Information Processing Systems*, 971–980.

Kolesnikov, S. 2023. Wild-Tab: A Benchmark for Out-Of-Distribution Generalization in Tabular Regression. *CoRR*, abs/2312.01792.

Kritzman, M.; Page, S.; and Turkington, D. 2012. Regime shifts: Implications for dynamic strategies. *Financial Analysts Journal*, 68(3): 22–39.

Laine, S.; and Aila, T. 2017. Temporal Ensembling for Semi-Supervised Learning. In *Proceeding of the 5th International Conference on Learning Representations*.

Li, Y.-F.; and Liang, D.-M. 2019. Safe semi-supervised learning: a brief introduction. *Frontiers of Computer Science*, 13(4): 669–676.

Liu, Y.; Kothari, P.; van Delft, B.; Bellot-Gurlet, B.; Mordan, T.; and Alahi, A. 2021. TTT++: When Does Self-Supervised Test-Time Training Fail or Thrive? In *Advances in Neural Information Processing Systems*, 21808–21820.

Niu, S.; Wu, J.; Zhang, Y.; Chen, Y.; Zheng, S.; Zhao, P.; and Tan, M. 2022. Efficient Test-Time Model Adaptation without Forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, 16888–16905.

Niu, S.; Wu, J.; Zhang, Y.; Wen, Z.; Chen, Y.; Zhao, P.; and Tan, M. 2023. Towards Stable Test-time Adaptation in Dynamic Wild World. In *Proceedings of the 11th International Conference on Learning Representations*.

Ozbayoglu, A. M.; Gudelek, M. U.; and Sezer, O. B. 2020. Deep learning for financial applications: A survey. *Applied Soft Computing*, 93: 106384.

Ren, W.; Li, X.; Chen, H.; Rakesh, V.; Wang, Z.; Das, M.; and Honavar, V. G. 2024. TabLog: Test-Time Adaptation for Tabular Data Using Logic Rules. In *Proceedings of the 41st International Conference on Machine Learning*.

Salehpour, A.; and Samadzamini, K. 2024. A bibliometric analysis on the application of deep learning in economics, econometrics, and finance. *International Journal of Computer Sciences and Engineering*, 27(2): 167–181.

Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A. A.; and Hardt, M. 2020. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In *Proceedings of the 37th International Conference on Machine Learning*, 9229–9248.

Torralba, A.; and Oliva, A. 2003. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14(3): 391.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B. A.; and Darrell, T. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *Proceedings of the 9th International Conference on Learning Representations*.

Wang, Q.; Fink, O.; Gool, L. V.; and Dai, D. 2022. Continual Test-Time Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7191–7201.

Wu, R.; Guo, C.; Su, Y.; and Weinberger, K. Q. 2021. Online Adaptation to Label Distribution Shift. In *Advances in Neural Information Processing Systems*, 11340–11351.

Zhang, D.-C.; Zhou, Z.; and Li, Y.-F. 2024. Robust Test-Time Adaptation for Zero-Shot Prompt Tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16714–16722.

Zhou, Z.; Guo, L.; Jia, L.; Zhang, D.; and Li, Y. 2023a. ODS: Test-Time Adaptation in the Presence of Open-World Data Shift. In *Proceedings of the 40th International Conference on Machine Learning*, 42574–42588.

Zhou, Z.; Zhang, D.-C.; Li, Y.-F.; and Zhang, M.-L. 2023b. Towards Robust Test-Time Adaptation for Open-Set Recognition. *Journal of Software*, 35(4): 1667–1681.