STAF: Pushing the Boundaries of Test-Time Adaptation Towards Practical Noise Scenarios

Haoyu Xiong[a], **Xinchun Zhang**[a], **Leixin Yang**[a], **Yu Xiang**[a,†], **Gang fang**[a]
[a]School of Information Science and Technology, Yunnan Normal University, China
{xionghaoyu,zxc,yangleixin,xiangyu,fanggang}@ynnu.edu.cn

Abstract

Test-time adaptation (TTA) aims to adapt the neural network to the distribution of the target domain using only unlabeled test data. Most previous TTA methods have achieved success under mild conditions, such as considering only a single or multiple independent static domains. However, in real-world settings, the test data is sampled in a correlated manner and the test environments undergo continual changes over time, which may cause previous TTA methods to fail in practical noise scenarios, i.e., independent noise distribution shifts, continual noise distribution shifts, and continual mixed distribution shifts. To address these issues, we elaborate a **S**table **T**est-time **A**daptation **F**ramework, called STAF, to stabilize the adaptation process. Specifically, to boost model robustness to noise distribution shifts, we present a multi-stream perturbation consistency method, enabling weak-to-strong views to be consistent, guided by the weak view from the original sample. Meanwhile, we develop a reliable memory-based corrector which utilizes reliable snapshots between the anchor model and the adapt model to correct prediction bias. Furthermore, we propose a dynamic parameter restoration strategy to alleviate error accumulation and catastrophic forgetting that takes into account both the distribution shift and sample adaptation degree. Extensive experiments demonstrate the robustness and effectiveness of STAF, which pushes the boundaries of test-time adaptation to more realistic scenarios and paves the way for stable deployment of real-world applications.

Keywords: Test time adaptation, Offensive language detection, Text categorisation

1. Introduction

Pre-trained language models (PLMs) have demonstrated superior performance on various natural language processing (NLP) tasks (He et al., 2023; Sanh et al., 2019; Liu et al., 2019; Wang et al., 2022c). However, when the training domain and testing domain are taken from different distributions, the deployed model often violates this assumption. In the real world, environmental data are typically non-stationary and constantly changing, and the testing data unavoidably undergoes natural variations or corruption. For instance, word spelling errors, toxic comments, OCR recognition text errors, which make PLMs often suffer from severe performance degradation (Lazaridou et al., 2021; Yao et al., 2022; Zhang and Gao, 2022). And due to the ever-changing nature of language, the test input might exhibit continual distribution shift over time (Dhingra et al., 2022).

In order to address this issue, an ideal goal is to enable deployed models to achieve human-like learning capabilities, allowing them to adapt and respond to diverse environments and tasks. Specifically, these models should be capable of learning and adapting in dynamic environments while retaining previously acquired knowledge. These abilities are vital for long-term deployment in the real-world (Wang et al., 2022a). For instance, autonomous driving systems and chat assistants interact with ever-changing environments for extended periods and require rapid and effective adaptation to new circumstances. To enhance the robustness and adaptability of models in such scenarios, researchers have explored methods such as continual learning (CL) (Lesort et al., 2020; Zenke et al., 2017) and domain adaptation (DA) (You et al., 2019; Pei et al., 2018). These approaches achieve their objectives through incremental training or retraining. However, these methods often assume that the source domain is accessible, data is labeled, and require a heavier burden of backpropagation. Moreover, these methods struggle to generalize to a wide array of potential unknown data distributions during training.

Recently, test-time adaptation (TTA) methods have emerged as an alternative solution (Wang et al., 2021; Niu et al., 2022, 2023; Lee, 2013; Ravichander et al., 2021). TTA methods update the model online using only the current unlabeled test data to adapt the model to the target domain distribution. TTA has been shown to be effective in handling distribution shift (Wang et al., 2021, 2022b; Niu et al., 2022; Manli et al., 2022). However, its superior performance is usually achieved under some mild test settings, where the test samples are independently sampled from single or multiple distributions. In real-world scenarios, the test data distribution may be non-stationary. As shown in Fig. 1 (left), these scenarios may encounter: (1) independent noise distribution shifts, (2) continual noise distribution shift, (3) continual mixed distribution shifts. These are common scenarios in reality,

†Corresponding author

LREC-COLING 2024, pages 15226–15237
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

where the test data not only significantly differs from the source domain distribution but also contains noise. Meanwhile, the test data is sampled correlatively over time, which makes the pseudo labels become noisy and calibration errors, leading to unstable adaptation of existing TTA methods.

To mitigate the degradation of model performance, we elaborate a **S**table **T**est-time **A**daptation **F**ramework (STAF) to further stabilize the adaptation process in practical noise scenarios from three aspects: (1) encourage the model to be consistent with the original weak view and the weak-to-strong perturbed views to boost model robustness to noise distribution shifts by MPC, (2) take into account both the distribution shift and sample adaptation degree to dynamically restore the parameters of the model by DPR, thereby alleviating catastrophic forgetting, and (3) utilize reliable snapshots between the anchor model and the adapt model to correct prediction bias while maintaining long-term memory by RMC. Promising results demonstrate that STAF can effectively extend the capabilities of deployed PLMs, enabling them to thrive in practical noise scenarios.

Main Contributions

- We construct a more challenging robustness evaluation benchmark, called NOISE WILDS-CIVILCOMMENTS, which not only contains significant distribution shifts but also have common natural noise.
- We elaborate a stable test-time adaptation framework, which considers more practical scenarios and is easy to implement and deploy.
- Extensive experimental results demonstrate the effectiveness of our proposed STAF and outperform the traditional TTA methods in practical noise scenarios.

2. Related Work

2.1. Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) aims to alleviate distribution shift by jointly optimizing the source domain data and unlabeled target domain data. Some approaches focus on self-supervised learning (Kumar et al., 2020), contrastive learning (Kang et al., 2020), or domain discriminators (Ganin and Lempitsky, 2015) to reduce the distribution shifts. To avoid accessing source domain data, recent works utilize information maximization (Liang et al., 2020), but they often require the entire target domain dataset and are performed offline, making them challenging to deploy in practical online applications.

2.2. Test-Time Adaptation

Test-time Adaptation (TTA) focuses on more challenging settings, which only use the current unlabeled test data to adapt the model to the target domain distribution. Since the test data also provides insights into distribution shift (Schneider et al., 2020a), simply adjusting the normalization statistics (Schneider et al., 2020b) can significantly improve the model's performance. While methods based on self-training with hard pseudo-labels (Lee, 2013) or entropy minimization (Wang et al., 2021) further perform backpropagation to update the parameters of normalization during testing. In a similar vein, (Niu et al., 2022, 2023) seeks to minimize reliable samples to restrict drastic updates.

2.3. Continual Learning

Continual/lifelong Learning (CL) is designed to imbue the model with the ability to acquire new knowledge in an uninterrupted data stream, transfer knowledge from the source domain to the target domain, and retain the memory of the source knowledge without succumbing to catastrophic forgetting (Parisi et al., 2019). Consequently, several CL methods strive to alleviate catastrophic forgetting by regularizing the preservation of source knowledge (Ahn et al., 2019; Kirkpatrick et al., 2017) and employing experience replay (Rolnick et al., 2019; Rebuffi et al., 2017). In this study, our motivation aligns with CL as we emphasize that TTA methods encounter the issue of catastrophic forgetting even in practical noise scenarios, thereby rendering the deployed model unstable.

3. Problem Definition and Motivation

Problem Definition. Given a model θ_0 with parameter $\boldsymbol{\theta}_0$, the parameter $\boldsymbol{\theta}_0$ is trained on the source domain $\mathcal{D}_{\mathrm{S}} = \{(\mathcal{X}_{\mathrm{S}}, \mathcal{Y}_{\mathrm{S}})\}$. Then, we use $\mathcal{P}_S$ and $\mathcal{P}_T$ to denote the data distribution of the source domain $\mathcal{D}_{\mathrm{S}}$ and the target domain $\mathcal{D}_{\mathrm{T}}$, respectively, where $\mathcal{X}_S \sim \mathcal{P}_S, \mathcal{X}_T \sim \mathcal{P}_T$. In general, the pre-trained model f_{θ_0} performs well on $\mathcal{D}_S$. However, due to the distribution shift between the source domain and the target domain, i.e., $\mathcal{P}_S \neq \mathcal{P}_T$, f_{θ_0} suffers significant performance degradation on the target domain.

To address these issues, typical (fully) test-time adaptation method (Wang et al., 2021) seeks to update the norm layer parameters by minimizing some unsupervised objective with the current unlabeled test data $x_t \in \mathcal{D}_{\mathrm{T}}$ at time step t, which can be formulated as follows:

$$\begin{aligned} &\min_{\boldsymbol{\theta}_t} \mathbb{E}_{x_t \sim \mathcal{D}_{\mathrm{T}}} \left[\ell(f_{\boldsymbol{\theta}_t}(x_t), y_t)\right] \\ \text{s.t.} \quad &\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \left(\nabla_{\boldsymbol{\theta}_t} \ell(\boldsymbol{\theta}_t)\right). \end{aligned} \tag{1}$$

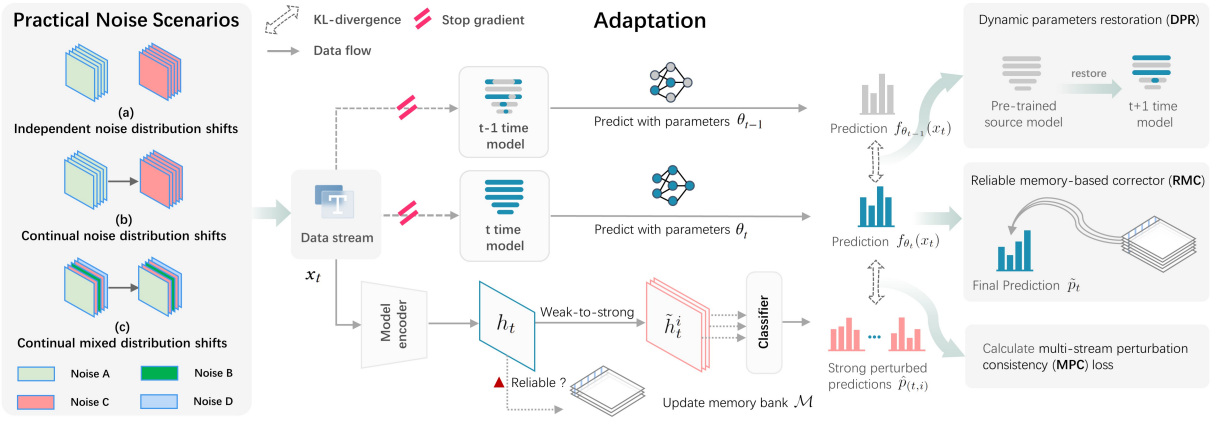


Figure 1: **Framework overview.** We mainly consider the following practical scenarios: (a) independent noise distribution shifts, (b) continual noise distribution shifts, and (c) continual mixed distribution shifts. Prior to adaptation, STAF is initialized with source pre-trained weights. During test-time, MPC in eq. (4) is designed to encourage weak-to-strong views to be consistent guided by the weak view from the original sample, thus boosting model robustness to noise distribution shifts. Meanwhile, DPR in eq. (12) is constructed to dynamically restore the parameters of the model by estimating the discrepancy between the predictions of $f_{\theta_{t-1}}(x_t)$ and $f_{\theta_t}(x_t)$ to alleviate catastrophic forgetting. Finally, RMC in eq. (9) utilizes reliable snapshots in the memory bank to correct prediction bias.

where η is the learning rate, $\ell(\cdot)$ can be formulated as the pure entropy minimization (Wang et al., 2021) or other variants (Niu et al., 2022, 2023; Lee, 2013). The model f_{θ_t} needs to update itself according to x_t and make online predictions immediately. Note that since most Transformer-based (Vaswani et al., 2017) pre-trained language models (Kenton and Toutanova, 2019; Sanh et al., 2019) do not have Batch-Norm layers, we only update the parameters of the Layer-Norm layer during the test-time adaptation process.

In our practical noise scenarios settings, as shown in Fig. 1 (left), the test scenario may meet: (a) independent noise distribution shifts, (b) continual noise distribution shifts, and (c) continual mixed distribution shifts. More challenging is that the test data distribution changes continually in scenario (b, c), i.e., $\mathcal{P}_0 \to \mathcal{P}_1 \to \cdot \to \mathcal{P}_\infty$. Furthermore, the test data x_t in scenario (c) further contains a mixture of multiple noise distribution shifts. It is important to note that the above mentioned scenarios not only have significant distribution shifts but also include common natural noise as described in Sec. 5.1.

Motivation. As a matter of fact, this setting is largely driven by the practical requirements of deploying models. Taking the chat assistants mentioned in Sec. 1 as an example, chat assistants need to interact with dynamic open environments and operate on non-static data. In addition, the constantly changing nature of language as spoken or written may be a key factor behind distribution shifts. Therefore, this degradation is also prevalent in pre-trained language models (PLMs) over time. Motivated by the fact that error accumulation and catastrophic forgetting are inevitable in practical noise scenarios, the urgent need prompts us to further propose a stable test-time adaptation framework to mitigate the degradation of model performance.

TTA considers more challenging but realistic problems and has attracted widespread attention and applications (Manli et al., 2022; Liu et al., 2022; Ma et al., 2022; Ye et al., 2022). However, it is still in its infancy in the NLP domain.

Algorithm 1 Proposed Approach STAF

Initialization: A source pre-trained model f_{θ_0};
Input: Unlabeled data stream x_t at time step t.
1: Feed forward x_t and generate the weak-to-strong perturbed predictions $\hat{p}_t^i$ by eq. (2).
2: Update model f_{θ_t} by multi-stream consistency loss in eq. (4).
3: Update memory bank $\mathcal{M}$ with more reliable snapshot in eq. (5).
4: Correct the prediction bias by eq. (9).
5: Dynamic parameters restoration by eq. (12).
Output: Predictions $\tilde{p}_t$; Updated model $f_{\theta_{t+1}}$.

4. Methods

Motivated by the fact that the error accumulation caused by noisy samples or low-quality pseudo-labels in practical noise scenarios, we propose to encourage the model to be consistent with the original weak view and the weak-to-strong perturbed views to boost model robustness to noise distribution shifts. Meanwhile, we develop a reliable

memory-based corrector, which utilizes reliable snapshots to correct prediction bias. Furthermore, to mitigate the catastrophic forgetting, we propose to dynamically restore the parameters of the model by estimating the adaptation degree of the current sample. An overview of our framework and algorithm is depicted in Fig. 1 and Algorithm 1, respectively.

Multi-stream Perturbation Consistency (MPC)

we posit that regularizing perturbed predictions to be consistent with a shared weak view from the original prediction can be regarded as enforcing consistency between these perturbed views. Although advanced methods (Bayer et al., 2022) have been proposed to generate strong views, their success heavily relies on the manual design of strong data augmentation. To break this dilemma, guided by consistency learning (Englesson and Azizpour, 2021; Wang and Shi, 2022), we propose to encourage the model to be consistent with the original weak view and the weak-to-strong perturbed views to boost model robustness to noise distribution shifts.

Specifically, let $\phi : \mathcal{X} \to \mathbb{R}^d$ be the model encoder and $g : \mathbb{R}^d \to \mathbb{R}^c$ be the classifier, where $\mathcal{X}$ is the input space, d and c are the dimension of feature space and the number of classes, respectively. For a test sample x_t appearing at the time t, we first obtain the original prediction p_t and the perturbed prediction $\hat{p}_t$ by:

$$\begin{aligned} \hat{h}_t &= \phi(x_t) \odot \xi_r, \quad \xi_r \sim \text{Bernoulli}(r), \\ \hat{p}_t &= g(\hat{h}_t), \quad p_t = g(h_t), \end{aligned} \tag{2}$$

where ξ_r is sampled from a Bernoulli distribution with a dropout rate r, and $\odot$ denotes the element-wise product. To boost the robustness to noisy samples, we propose to minimize the divergence between p_t and $\hat{p}_t$, resulting in the following single-stream perturbation consistency loss:

$$\begin{aligned} &\mathcal{L}_{\text{SPC}}(p_t, \hat{p}_t, \theta_t) = \frac{1}{2}\left(\mathcal{L}_{\text{KLD}}(p_t||\hat{p}_t) + \mathcal{L}_{\text{KLD}}(\hat{p}_t||p_t)\right), \\ &\text{with} \;\; \mathcal{L}_{\text{KLD}}(p_t||\hat{p}_t) = \sum_{i=1}^{c} p_t^i \log \frac{p_t^i}{\hat{p}_t^i}, \end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{KLD}}$ is the Kullback-Leibler Divergence (KLD). While minimizing the KLD, the divergence between the original prediction p_t and the perturbed prediction $\hat{p}_t$ is minimized, thus enhances the robustness to noisy samples.

However, the severity of noise varies from sample to sample, and the model may not exhibit consistent stability across different severity levels of noise. Guided by weak-to-strong consistency (Yang et al., 2023) and the smoothing assumption (Wagner et al., 2018), we propose to gradually increase the dropout rate r ($r : 0.1 \to 0.2 \cdots$), resulting in weak-to-strong perturbed predictions $\hat{p}_t^i$. This allows us to enable weak-to-strong views to be consistent guided by the weak view from original sample and multiple views can be complementary to each other. Overall, the multi-stream perturbation consistency loss is given by:

$$\mathcal{L}_{\text{MPC}}(x_t, \theta_t) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}_{\text{PC}}(p_t, \hat{p}_t^i, \theta_t). \tag{4}$$

Here, the additional costs are negligible since only optimize the parameters in Layer-Norm and a single forward pass is required to perform multiple perturbations.

Reliable memory-based corrector (RMC)

Motivated by the fact that the continually changing environments, the pseudo-labels tend to become noisier and miscalibrated over time. An ideal solution is to maintain a memory bank $\mathcal{M}$, which can be used to correct the prediction bias. However, updating $\mathcal{M}$ may contain unreliable snapshots, which may distract the model from the correct direction.

To address this issue, we propose a reliable memory-based corrector (RMC) to correct the prediction bias. Specifically, we propose to update the memory bank $\mathcal{M}$ by maintaining more reliable key-value $\{q_t : \phi_\Theta(x_t), v_t : f_\Theta(x_t)\}$ pairs between the anchor model f_{θ_0} with parameter θ_0 and the adapted model f_{θ_t} with parameter θ_t, which is defined as follows:

$$\mathcal{M} \leftarrow \mathcal{M} \cup \{q_t, v_t\} \cdot \mathbb{I}(x_t; \theta_0, \theta_t,) \tag{5}$$

where $\mathbb{I}(\cdot)$ is an indicator function to determine whether the snapshot is reliable or not, is defined as:

$$\mathbb{I}(\cdot) = \begin{cases} \phi_{\theta_t}(x_t), & \text{if } \max\{f_{\theta_0}(x_t)\} \geq \max\{f_{\theta_t}(x_t)\} \\ \phi_{\theta_0}(x_t), & \text{otherwise} \end{cases} \tag{6}$$

For a test sample x_t appering at the time t, we initially retrieve a support set $\mathcal{S}(\phi(x_t)) = \{(h_j, v_j)\}_{j=1}^{k}$ from $\mathcal{M}$, where k is the number of retrieved samples. Guided by the smoothing assumption (Zhang et al., 2019), let $c_{i,j} = \frac{h_i \cdot h_j}{\|h_i\| \cdot \|h_j\|}$ be the cosine similarity between feature h_i and h_j. Then, we can use the cosine similarity distance to assign the attention weight $w_{t,k}$ to each corresponding sample in $\mathcal{S}$ as:

$$w_{t,k} = \frac{\exp(c_{t,k})}{\sum_{j \in k} \exp(c_{t,k})} \tag{7}$$

where cosine similarity $c_{t,k}$ is then computed between k selected samples and x_t.

We adopt ensemble strategy (Dong et al., 2020) to take into account the intermediate result $w_{t,k}$ in eq. (7), which is ensembled as follows:

$$\tilde{v}_{t,k} = \sum_{j \in k} w_{t,j} \cdot v_j \tag{8}$$

Then, the final prediction can be corrected as follows:

$$\tilde{p}_t = \left(f_{\theta_t}(x_t) + \tilde{v}_{t,k}\right)/2 \tag{9}$$

Furthermore, the estimates are not stable within a single mini-batch, and the model may not exhibit consistent stability across different mini-batches. Therefore, we use a fixed-length FIFO (first-in, first-out) queue to cache the most recent key-value pairs. We discuss the computational cost of the RMC module in Sec. 6.

Dynamic Parameters Restoration (DPR)

To reduce the long-term error accumulation and catastrophic forgetting in lifelong TTA, (Wang et al., 2022b) proposed to further update the parameters by randomly restoring a small number of tensor elements in the trainable weights after the gradient update at time step t:

$$\begin{aligned} \mathbf{M}_t &= \text{Bernoulli}(\rho_0) \\ \theta_{t+1} &= \mathbf{M}_t \odot \theta_0 + (1 - \mathbf{M}_t) \odot \theta_{t+1}. \end{aligned} \tag{10}$$

where $\mathbf{M}$ is a mask matrix that determines which parameters within θ_{t+1} need to be restored to the initial weights θ_0, $\rho_0 = 0.1$ is stochastic restore probability, and $\odot$ denotes element-wise product. Note that (Wang et al., 2022b) is not suitable for the NLP domain.

However, due to the change of model parameters over time, even for samples with similar distribution shifts but different arrival times, the demand for adaptation degree should be different. If the parameters of samples with slight distribution shifts are restored drastically, it will lead to the degradation of the model's ability to adapt to new samples. Therefore, it is necessary to dynamically adjust the probability of parameter restoration according to the adaptation degree required by each sample, so as to reduce the long-term error accumulation while maintaining the ability to adapt to new samples.

Specifically, for a test sample x_t appearing at time t, we aim to estimate the adaptation degree τ_t of the model to the current sample by capturing the distribution shift before and after adaptation. Let p_t be the prediction of the model with parameters θ_t, then the adaptation degree τ_t is estimated as follows:

$$\tau_t = \frac{1}{2}\left(\mathcal{L}_{\text{KLD}}(f_{\theta_{t-1}}(x_t)||p_t) + \mathcal{L}_{\text{KLD}}(p_t||f_{\theta_{t-1}}(x_t))\right) \tag{11}$$

Thus, the stochastic restore probability in eq. (10) can be dynamically adjusted over time, which is defined as follows:

$$\rho_t = \exp\left(\tau_t\right) \cdot \rho_0 \ , \tag{12}$$

satisfying constraints $\rho_t \in [0, 1]$. Accordingly, the elements in θ_{t+1} are restored to the initial weights θ_0 with a probability of ρ_t.

5. Experiments

5.1. Setup

Dataset. To evaluate our method, we selected a dataset with a significant distribution shift between the train and test distributions, i.e. WILDS-CIVILCOMMENTS (Koh et al., 2021), which is a modification of the original CivilComments dataset (Borkan et al., 2019). This dataset comprises 269,038 training samples and 133,782 test samples. Each comment text is associated with metadata indicating membership in one or more of eight sensitive groups, and is labeled as toxic or non-toxic using a binary indicator[1].

Challenge Settings of Dataset. Due to the lack of datasets with relevant distribution shift and noise in the NLP field, we have constructed a more challenging robustness evaluation benchmark by processing WILDS-CIVILCOMMENTS, called "Noise WILDS-CivilComments". This dataset not only contains significant distribution shifts but also have common natural noise, i.e. recognition errors by mimicking optical character recognition (OCR) engines (Ma, 2019), keyboard errors (keyboard) (Belinkov and Bisk, 2018), machine translation errors (backtranslate) (Jörg Tiedemann, 2020), synonym replacer (Pavlick et al., 2015), and spelling errors (Coulombe, 2018). Tab. 1 shows examples of NOISE WILDS-CIVILCOMMENTS dataset.

Baselines and Models. All experiments were conducted on the pre-trained DistilBERT (Sanh et al., 2019) network and fine-tuned on the WILDS-CIVILCOMMENTS dataset followed by (Koh et al., 2021). During test time, the *Baseline* represents the pre-trained model directly evaluated on the target domain without any adaptation. Apart from the baseline, we compare with the following typical and strong baselines to verify the effectiveness of STAF: (1) *TENT* (Wang et al., 2021) minimizes entropy to update norm layer parameters. (2) *PL* (Lee, 2013) updates norm layer parameters with hard pseudo-labels. (3) *LN* (Schneider et al., 2020b)

[1]Our source code, dataset, and pre-trained models are available at https://anonymous.4open.science/r/coling-tta-D527/.

Comment	...	Operation	Conditions
Only an idiot would believe that.	...	Original comment	Shift
$\underline{On.j}$ an idiot $\underline{eoupd}$ $\underline{F2lisve}$ that.	...	Keyboard	Shift + Noise
Only an idiot would believe $\underline{it}$.	...	Backtranslate	Shift + Noise
Only an 1diot would 6e1ieve that.	...	OCR	Shift + Noise
Only $\underline{at}$ idiot $\underline{wood}$ believe that.	...	Spelling	Shift + Noise
Only an idiot would $\underline{understood}$ that.	...	Synonym	Shift + Noise

Table 1: Examples of NOISE WILDS-CIVILCOMMENTS challenge sets from 5 types of natural noise: keyboard error, machine translation error, OCR engines recognition error, synonym, and spelling error. The underline indicates the operation part. Our NOISE WILDS-CIVILCOMMENTS not only have significant distribution shift but also have common natural noise.

only utilizes layer normalization statistics from the test input and keep frozen model parameters. (4) *EATA* (Niu et al., 2022) seeks to minimizes reliable and non-redundant samples, and use the fisher regularizer to restrict model updates. (5) *SAR* (Niu et al., 2023) seeks to find falt minimum (Foret et al., 2021) and minimizes reliable samples, also further restoring the model by recording a moving average of loss.

Implementation Details. In our experiments, we adopt Adam optimizer (Kingma and Ba, 2015) with learning rate $1e^{-5}$, the memory queue length is set to $c \times 100$ (c is the number of classes), the retrieval size $k = 6$ in eq. (7), $m = 5$ in eq. (4), and default values for all other hyperparameters. For a fair comparison, we set batch size to 8 and keep the default settings of other TTA methods. **Note** that we do not perform any tuning during training, we only conduct adaptation on the pre-trained model.

5.2. Results for Independent Noise Distribution Shifts

To evaluate the effectiveness of STAF, we first consider the independent noise distribution shifts scenario in Fig. 1, scenario (a), where the target domain is exposed to distribution shifts and noise independently. From Tab. 2, it it obvious that the performance of the model after adaptation has been improved to varying degrees, which highlights the indispensability of adaptation.

Notably, *TENT* and *PL* can achieve significant improvements under mild conditions (e.g. *Backtranslation*, *Spelling*, *Source*) and even comparable to our method. However, since the EATA and SAR methods stabilize the adaptation by restricting the update, but in some cases it hinders the model's adaptation ability, resulting in poor gains.

On the contrary, STAF attains superior results on most conditions compared to previous methods, and significantly outperforms the baseline by 2.18%, verifying the effectiveness of our method to boost robustness to noisy samples.

5.3. Results for Continual Noise Distribution Shifts

Moreover, real-world applications are running in practical noise scenarios, where the environment is non-stationary and continually changing, and the test data is sampled correlatively over time. As shown in eq. (13), it is necessary to further evaluate on continual noise distribution shifts scenario, where the target data arrives continually from dif-

Method	Source	Keyboard	Backtranslation	OCR	Spelling	Synonym	Avg.
Baseline	90.06	88.99	89.67	90.11	89.13	90.71	89.78±0.30
PL	2.37	1.77	<u>2.42</u>	<u>1.60</u>	<u>2.92</u>	1.44	2.09±0.38
TENT	<u>2.39</u>	<u>1.78</u>	**2.44**	1.59	<u>2.92</u>	<u>1.46</u>	<u>2.10±0.39</u>
EATA	0.62	0.51	0.53	0.43	0.65	0.40	0.52±0.08
LN	0.60	0.50	0.56	0.44	0.69	0.38	0.53±0.08
SAR	0.54	0.65	0.51	0.57	0.64	0.75	0.61±0.21
STAF (Ours)	**2.61**	**1.98**	2.18	**1.74**	**2.94**	**1.65**	**2.18**±0.35

Table 2: **Independent noise distribution shifts scenario** (Fig. 1, scenario-a). Percentage difference in accuracy (%) over 5 runs. The number in brackets represents the standard deviation and underline indicates the second best result.

Time	$t \longrightarrow$							
Method	Source	Keyboard	Backtranslation	OCR	Spelling	Synonym	Source*	Avg.
Baseline	90.06	88.99	89.67	90.11	89.13	90.71	90.06	89.82±0.27
PL	2.37	-0.01	-0.07	-1.37	-0.31	-1.90	-1.27	-0.37±0.45
TENT	2.39	0.01	-0.14	-1.33	-0.38	-1.96	-1.22	-0.38±0.46
EATA	0.62	0.58	0.55	0.41	0.68	0.38	0.63	0.55±0.08
LN	0.60	0.50	0.52	0.45	0.69	0.38	0.60	0.53±0.09
SAR	0.59	0.47	0.62	0.41	0.59	0.43	0.65	0.54±0.12
STAF (Ours)	**2.40**	**1.72**	**1.91**	**1.46**	**2.48**	**1.19**	**2.45**	**1.94**±0.47

Table 3: **Continual noise distribution shifts scenario** (Fig. 1, scenario-b). Percentage difference in accuracy (%) over 5 runs. The test inputs t from different target domains arrive continually. Here, *Source** indicates returning to the source domain to re-adapt, the red color indicates the results which lower than the *Baseline*, the underline indicates the second best result, and the **bold** indicates the best performance.

ferent target domain distributions as:

$$\ldots \underbrace{\mathcal{D}_{t-1}}_{\mathcal{P}_{t-1}} \xrightarrow[\text{noise distribution}]{\text{change}} \underbrace{\mathcal{D}_{t}}_{\mathcal{P}_{t}} \xrightarrow[\text{noise distribution}]{\text{change}} \underbrace{\mathcal{D}_{t+1}}_{\mathcal{P}_{t+1}} \ldots \tag{13}$$

From Tab. 3, we can observe that *PL* and *TENT* are particularly prone to occur degradation, especially on more challenging scenarios (e.g., *OCR*), resulting in a significant decline of -1.37% and -1.33%. Furthermore, we find that although *SAR* benefits from record a moving average of entropy loss values to reset the model to prevent model collapse, however, when the loss fluctuates greatly, it is easy to trigger the model reset condition frequently, resulting in performance similar to *LN* which only uses layer normalization statistics. It is worth to note that *EATA* achieves a gain of 0.55% by restricting model updates, but it also requires access to the source data, which defeats the whole purpose of the TTA paradigm.

Conversely, STAF achieves better and more robust results, and significantly outperforming the second-best method by 1.39%. Moreover, STAF is the best result in all conditions, verifying the effectiveness in the continual adaptation process.

Comparison of Continual and Independent Noise Test-time Adaptation. In Fig. 2, although most methods perform well on fixed domains, the performance of Tent and PL has declined to varying degrees due to the lack of effective measures to deal with distribution shifts. On the contrary, our method is label-independent, which is not susceptible to noisy pseudo-labels, and thus can stably adapt in continual noise distribution shifts.

5.4. Results for Continual Mixed Distribution Shifts

Under practical noise scenarios, the data distribution types may be arbitrary. Therefore, we additionally evaluate our method on continual mixed

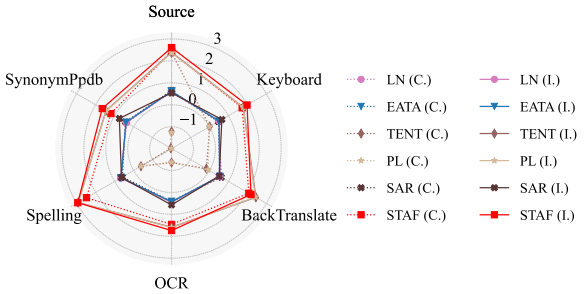


Figure 2: **Comparison of continual (C.) and independent (I.) noise distribution shifts.** Each vertex represents a type of corruption, and the farther the vertex is from the center, the better the performance.

distribution shifts scenario, as depicted in Fig. 1 (left-c). In this scenario, each time step t contains a mixture of noise distribution types with a random shuffle order, which are sampled from a uniform distribution over each test input and share semantic category labels with *Source*. In order to simulate scenarios in real-life situations that may be revisited, and evaluate the forgetting effect of our approach, we repeat the same target sequence group eight times followed by (Wang et al., 2022b) as:

$$\ldots \quad \underbrace{x_{t-1}}_{\ldots,\mathcal{P}_0,\mathcal{P}_2,\mathcal{P}_2,\ldots} \quad \rightarrow \quad \underbrace{x_{t}}_{\ldots,\mathcal{P}_1,\mathcal{P}_0,\mathcal{P}_2,\ldots} \quad \rightarrow \quad \underbrace{x_{t+1}}_{\ldots,\mathcal{P}_3,\mathcal{P}_1,\mathcal{P}_3,\ldots} \quad \ldots \tag{14}$$
$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{8\times\mathcal{D}_T}$$

where each x_t contains a mixture of different noise distribution types.

From Tab. 4, we can observe that although *SAR* and *EATA* have limited gains in the independent noise distribution shifts scenario, but they can maintain consistent positive performance, resulting in improvements of 0.55% and 0.61%, respectively. Moreover, *LN* also exhibits similar behavior. On the other hand, *TENT* and *PL* exhibit a rapid degra-

Time	t ⟶								
Round	1	2	3	4	5	6	7	8	Avg.
Baseline	89.70	89.70	89.70	89.70	89.70	89.70	89.70	89.70	89.70
PL	2.33	-0.01	-0.61	-0.80	-0.85	-0.88	-0.92	-0.96	-0.34±0.38
TENT	2.35	0.02	-0.58	-0.76	-0.83	-0.92	-0.96	-0.98	-0.33±0.37
EATA	0.61	0.63	0.60	0.59	0.57	0.60	0.62	0.63	0.61±0.07
LN	0.61	0.59	0.58	0.61	0.58	0.59	0.61	0.58	0.59±0.07
SAR	0.51	0.58	0.57	0.58	0.60	0.49	0.51	0.57	0.55±0.13
STAF (Ours)	**2.48**	**2.32**	**2.05**	**1.89**	**2.02**	**1.86**	**1.75**	**1.87**	**2.03**±0.34

Table 4: **Continual mixed distribution shifts scenario** (Fig. 1, scenario-c). Percentage difference in accuracy (%) over 5 runs. The test inputs arrive continually while contain a mixture of multiple noise distribution shifts. Here, the red color indicates the results which lower than the *Baseline*, the underline indicates the second best result, and the **bold** indicates the best performance.

dation in performance after the second round, with a decrease of 0.58% and 0.61%, which gradually intensifies and becomes unavoidable. Moreover, expanding the gains in such a challenging dynamic scenario is difficult, but STAF consistently achieves the best results in all rounds, and leverages an average improvement of 2.03%, which again demonstrates the effectiveness of STAF.

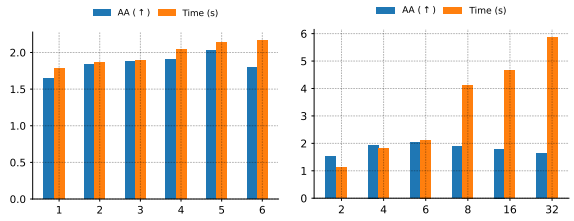
6. Ablation Studies

Effect of Each Component From Tab. 5, compared with pure entropy minimization, our MPC in eq. (4) significantly improves the model performance, i.e., -0.38% → 1.49%. DPR in eq. (12), further improves the average classification accuracy by 1.88%. Meanwhile, the RMC in eq. (9) is also effective, increasing accuracy from 1.88% → 2.03%. If any component is removed from STAF, the performance will decline, thus confirming the indispensability of each component.

Method	AA ↑
Baseline	89.82
Entropy (Wang et al., 2021)	-0.38
MPC	1.49
✢ DPR	1.88
✢ DPR + RMC	2.03

Table 5: Average accuracy (AA ↑) difference of each component on the continual noise distribution shifts scenario.

Ablations Fig. 3 shows the sensitivity analysis of different perturbation number m and different retrieval size k and the time calculation overhead. We observe that $m = 5$ provides a good balance between performance and computational overhead. We limit the maximum to 5 because the dropout rate over 0.5 easily cuts off too many connections between layers and limits the learning ability of the network. When the retrieval size k is 6, it can earn more profits, while when k is 8, the performance begins to decline.



(a) Perturbation number m (b) Retrieval size k

Figure 3: Average accuracy (AA ↑) difference and time (s) for different perturbation number m and different retrieval size k on continual noise distribution shifts scenario.

7. Conclusion

In this work, we introduce several practical settings for test-time adaptation, i.e., independent noise distribution shifts, continual noise distribution shifts, continual mixed distribution shifts. To stabilize the adaptation process in practical noise scenarios, we elaborate a stable test-time adaptation framework (STAF). Motivated by the fact that the error accumulation in practical noise scenarios, we present a multi-stream perturbation consistency method (MPC), which enables multiple perturbed views to be consistent guided by the weak view from original sample to boost noise distribution shifts. Meanwhile, we develop a reliable memory-based corrector, which utilizes reliable snapshots to correct prediction bias. Furthermore, we propose a dynamic parameter restoration strategy that takes into account both the distribution shift and sample adaptation degree, thus mitigating catastrophic forgetting. Extensive experimental results demonstrate the stability and effectiveness of STAF, which pushes the boundaries of test-time adaptation towards practical noise scenarios and paves the way

for stable deployment of real-world applications.

8. Limitations

Potential limitations of our method are that it requires maintaining a memory bank for bias correction, and the parameter restoration has uncertainty. In future work, we will explore more efficient memory algorithms (Ming et al., 2022; Johnson et al., 2021; Sun et al., 2022) to reduce computational costs, and only restore irrelevant parameters to maintain learned knowledge (Brahma and Rai, 2023; Kirkpatrick et al., 2017). Moreover, test-time adaptation may lead to carbon emission issues due to the need to adapt to all samples. Therefore, we will explore how to reduce computational costs during test-time adaptation to better deploy in real-world applications.

9. Bibliographical References

Hongjoon Ahn, Sungmin Cha, Donggyu Lee, and Taesup Moon. 2019. Uncertainty-based continual learning with adaptive regularization. *Advances in neural information processing systems*, 32.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A survey on data augmentation for text classification. *ACM Computing Surveys*, 55(7):1–39.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

Dhanajit Brahma and Piyush Rai. 2023. A probabilistic framework for lifelong test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3582–3591.

Claude Coulombe. 2018. Text data augmentation made simple by leveraging nlp cloud apis. *ArXiv*, abs/1812.04718.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. 2020. A survey on ensemble learning. *Frontiers of Computer Science*, 14:241–258.

Erik Englesson and Hossein Azizpour. 2021. Generalized jensen-shannon divergence loss for learning with noisy labels. In *Advances in Neural Information Processing Systems*.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2021. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1180–1189. JMLR.org.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

J. Johnson, M. Douze, and H. Jegou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(03):535–547.

Santhosh Thottingal Jörg Tiedemann. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Guoliang Kang, Lu Jiang, Yunchao Wei, Yi Yang, and Alexander G Hauptmann. 2020. Contrastive adaptation network for single-and multi-source domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR.

Ananya Kumar, Tengyu Ma, and Percy Liang. 2020. Understanding self-training for gradual domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5468–5479. PMLR.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, et al. 2021. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363.

Dong-Hyun Lee. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 2.

Timothée Lesort, Vincenzo Lomonaco, Andrei Stoian, Davide Maltoni, David Filliat, and Natalia Díaz-Rodríguez. 2020. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information fusion*, 58:52–68.

Jian Liang, Dapeng Hu, and Jiashi Feng. 2020. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6028–6039. PMLR.

Quande Liu, Cheng Chen, Qi Dou, and Pheng-Ann Heng. 2022. Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1756–1764.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Wenao Ma, Cheng Chen, Shuang Zheng, Jing Qin, Huimao Zhang, and Qi Dou. 2022. Test-time adaptation with calibration of medical image classification nets for label distribution shift. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 313–323. Springer.

Shu Manli, Nie Weili, Huang De-An, Yu Zhiding, Goldstein Tom, Anandkumar Anima, and Xiao Chaowei. 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*.

Yifei Ming, Ying Fan, and Yixuan Li. 2022. Poem: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning*, pages 15650–15665. PMLR.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. 2022. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pages 16888–16905. PMLR.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. 2023. Towards stable test-time adaptation in dynamic wild world. In *Internetional Conference on Learning Representations*.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430.

Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. 2018. Multi-adversarial domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Abhilasha Ravichander, Siddharth Dalmia, Maria Ryskina, Florian Metze, Eduard Hovy, and

Alan W Black. 2021. NoiseQA: Challenge set evaluation for user-centric question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2976–2992, Online. Association for Computational Linguistics.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. 2019. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. 2020a. Improving robustness against common corruptions by covariate shift adaptation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. 2020b. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33:11539–11551.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. 2022. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20827–20840. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tal Wagner, Sudipto Guha, Shiva Kasiviswanathan, and Nina Mishra. 2018. Semi-supervised learning on data streams via temporal label propagation. In *International Conference on Machine Learning*, pages 5095–5104. PMLR.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. 2022a. Generalizing to unseen domains: A survey on domain generalization. *IEEE Trans. Knowl. Data Eng.*

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. 2022b. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211.

Xin Wang and Hongbin Shi. 2022. Leveraging perturbation consistency to improve multi-hop knowledge base question answering. In *2022 IEEE Smartworld, Ubiquitous Intelligence & Computing, Scalable Computing & Communications, Digital Twin, Privacy Computing, Metaverse, Autonomous & Trusted Vehicles (SmartWorld/UIC/ScalCom/DigitalTwin/PriComp/Meta)*, pages 1360–1365. IEEE.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022c. Measure and improve robustness in NLP models: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4569–4586, Seattle, United States. Association for Computational Linguistics.

Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. 2023. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7236–7246.

Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei W Koh, and Chelsea Finn. 2022. Wild-time: A benchmark of in-the-wild distribution shift over time. *Advances in Neural Information Processing Systems*, 35:10309–10324.

Hai Ye, Yuyang Ding, Juntao Li, and Hwee Tou Ng. 2022. Robust question answering against distribution shifts with test-time adaption: An empirical study. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6179–6192, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2019. Universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2720–2729.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR.

Lei Zhang and Xinbo Gao. 2022. Transfer adaptation learning: A decade survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhong-Liang Zhang, Yu-Yu Chen, Jing Li, and Xing-Gang Luo. 2019. A distance-based weighting framework for boosting the performance of dynamic ensemble selection. *Information Processing & Management*, 56(4):1300–1316.