Exploring Safety Supervision for Continual Test-time Domain Adaptation

Xu Yang, **Yanan Gu**, **Kun Wei** and **Cheng Deng***
Xidian University
{xuyang.xd, yanangu.xd, weikunsk, chdeng.xd}@gmail.com

Abstract

Continual test-time domain adaptation aims to adapt a source pre-trained model to a continually changing target domain without using any source data. Unfortunately, existing pseudo-label learning methods suffer from the changing target domain environment, and the quality of generated pseudo-labels is attenuated due to the domain shift, leading to instantaneous negative learning and long-term knowledge forgetting. To solve these problems, in this paper, we propose a simple yet effective framework for exploring safety supervision with three elaborate strategies: Label Safety, Sample Safety, and Parameter Safety. Firstly, to select reliable pseudo-labels, we define and adjust the confidence threshold in a self-adaptive manner according to the test-time learning status. Secondly, a soft-weighted contrastive learning module is presented to explore the highly-correlated samples and discriminate uncorrelated ones, improving the instantaneous efficiency of the model. Finally, we frame a Soft Weight Alignment strategy to normalize the distance between the parameters of the adapted model and the source pre-trained model, which alleviates the long-term problem of knowledge forgetting and significantly improves the accuracy of the adapted model in the late adaptation stage. Extensive experimental results demonstrate that our method achieves state-of-the-art performance on several benchmark datasets.

1 Introduction

Deep neural networks have achieved remarkable success in visual tasks when training and test data follow the same distribution. These networks, however, suffer from the generalization problem in the presence of domain shift. For example, a classification network pre-trained in the normal, natural images domain may not recognize the corrupted images due to domain shift [Dong *et al.*, 2020; Yang *et al.*, 2022]. Thus, an adaptation is necessary to transfer knowledge from the source domain to the target one by reducing the domain shift. However, the label of the target domain is usually unavailable, so the problem is known as *Unsupervised Domain Adaptation* (UDA) [Ganin and Lempitsky, 2015; Prabhu *et al.*, 2021; Sun *et al.*, 2019; Dong *et al.*, 2021]. In addition, the source data is often inaccessible during inference time due to privacy or business problems, making the adaptation problem more challenging but more realistic. Therefore, such adaptation problem becomes *Source-Free/Test-Time domain Adaptation* (TTA) [Chen *et al.*, 2022; Yang *et al.*, 2021; Kurmi *et al.*, 2021; Kundu *et al.*, 2020; Liu *et al.*, 2021] where only the source model and unlabeled target data are available in the adaptation process.

Existing TTA methods usually solve the domain shift problem by updating the adapted model parameters using the generated pseudo-labels or entropy regularization. These self-training-based methods are effective when the distribution of data in the target domain is fixed, but when the distribution of the target domain is constantly changing [Wang *et al.*, 2022; Prabhu *et al.*, 2021], these methods become unstable. Noise problem [Wang *et al.*, 2022] caused by the change of target domain distribution seriously affects the adaptation process. CoTTA [Wang *et al.*, 2022] first defines this kind of problem as *Continual Test-Time Domain Adaptation*, where a source pre-trained model needs to adapt to a stream of continually changing target test data without using any source data. It uses a weight-average teacher network to improve the quality of generated pseudo-labels. However, the accurate optimization of a network requires the joint action of label and loss function. The pseudo-labels generated by existing methods become noisier in the changing target domain environment because of domain shifts in the adaptation process. Thus, refining pseudo-label learning requires high-quality pseudo-labels and a reliable label selection strategy.

In this paper, we propose a simple yet effective framework for continual test-time domain adaptation, which refines the pseudo-label learning process for exploring safety supervision from three perspectives: Label Safety, Sample Safety, and Parameter Safety, to alleviate the instantaneous and long-term impact of noisy pseudo-labels. In terms of label safety, noisy pseudo-labels will immediately affect the loss calculation in network optimization. Thus, we define and adjust the confidence threshold in a self-adaptive manner to restrain the noisy pseudo-labels in the test-time learning status. Notably, instead of simply presenting a constant

*Corresponding author.

hyper-parameter and treating all classes equally, we select an independent threshold for each class through global and local strategies to choose more reliable pseudo-labels as supervision information. Based on such thresholds, we construct a soft-weighted contrastive learning module for the sample safety of contrastive learning module, which pulls the reliable same-class samples closer and discriminates against uncorrelated samples. Moreover, the Fourier transform is introduced into the contrastive learning process as a strong augmentation to explore domain-invariant predictions. In this way, we can alleviate instantaneous negative learning and improve learning efficiency.

From the perspective of long-term impact, as the model continually adapts to the target domain with changing distributions, the knowledge from the pre-trained model is constantly forgotten due to error accumulation caused by the noisy pseudo-labels. To slow down the accumulation of errors and combat knowledge forgetting, we propose a Soft Weight Alignment (SWA) strategy, which continuously distills knowledge from the source pre-trained model. In detail, we normalize the distance between the parameters of the adapted model and the pre-trained model. In doing so, we continuously distill knowledge from the source pre-trained model and can effectively alleviate the long-term knowledge-forgetting problem. Extensive experimental results demonstrate that our method achieves state-of-the-art performance on several datasets.

To sum up, our contributions are as follows:

- **We design a novel yet efficient continual test-time domain adaptation method for exploring safety supervision.** The learnable thresholds are tuned in a local and global manner according to the test-time learning status to enhance label safety, sample safety, and parameter safety, further alleviating instantaneous negative learning and long-term knowledge-forgetting problems.
- **We introduce a soft-weighted contrastive learning module, which pulls the reliable same-class samples closer and alleviates the false negative samples.** The learnable thresholds and pseudo-labels are utilized to select more discriminative positive and precise negative pairs for contrastive learning. Moreover, we employ the Fourier transform to augment the positives, exploring domain-invariant predictions.
- **We propose a Soft Weight Alignment strategy, which continuously distills knowledge from the source pre-trained model.** We normalize the distance between the adapted and pre-trained models' parameters. In doing so, we continuously distill knowledge from the source pre-trained model and can effectively alleviate the long-term knowledge-forgetting problem.

2 Related Work

2.1 Domain Adaptation

Domain adaptation refers to the goal of learning a concept from labeled data in the source domain that performs well on different but related target domains. The critical problem of domain adaptation lies in the misalignment between the feature and label spaces of the source and target domains. To solve this problem, some domain adaptation methods guide the deep model to learn domain invariant representation and classifiers. Specifically, some works [Ganin and Lempitsky, 2015; Tzeng *et al.*, 2017; Ganin *et al.*, 2016] utilize adversarial training to align feature distribution with a domain discriminator, and some works constrain the cross-domain feature space by entropy constraint [Grandvalet and Bengio, 2004; Saito *et al.*, 2019] or maximum prediction rank [Cui *et al.*, 2020; Yang *et al.*, 2019].

2.2 Test-Time Domain Adaptation

Recently, some works on test-time domain adaptation focus on a more challenging setting where only the source model and unlabeled target data are available. Some test-time domain adaptation methods [Kurmi *et al.*, 2021; Li *et al.*, 2020] utilize generative models to achieve the feature alignment between the source and target domain in the absence of source data. Kurmi *et al.* adopt the trained classifier [Kurmi *et al.*, 2021] to generate samples from the source classes. This method learns the joint distribution of data by using the energy-based modeling of the trained classifier. Yeh *et al.* propose the method of Source-data-free Feature Alignment (SoFA) [Yeh *et al.*, 2021] to extract features with class semantics, thus realizing domain adaptation in the absence of the source data. In addition, some methods achieve test-time domain adaptation by finetuning the source model with the help of target data and do not require explicit domain alignment. Test entropy minimization (TENT) [Wang *et al.*, 2020] introduces entropy minimization as a test-time optimization objective, which estimates normalization statistics and optimizes channel-wise affine transformations to update online on each batch. Source HypOthesis Transfer (SHOT) [Liang *et al.*, 2020] aims to learn the optimal target-specific feature learning module to fit the source hypothesis. It is worth mentioning that SHOT requires using source data to train a specialized source model, so it cannot support an arbitrary source pre-trained model. Zhou *et al.* present Bayesian Adaptation for Covariate Shift (BACS) [Zhou and Levine, 2021] to obtain both improved accuracy and well-calibrated uncertainty estimates when faced with domain shift.

Most test-time adaptation methods only consider the offline scenario, where the full set of test data is provided during the training process. Further, CoTTA [Wang *et al.*, 2022] extends test-time adaptation from offline scenario to online continual scenario. It considers a more challenging but more realistic problem named *Continual Test-Time Domain Adaptation*, where a source pre-trained model needs to adapt to a stream of continually changing target test data without using any source data. In this paper, we also focus on such a more realistic setting than standard test-time adaptation.

3 Proposed Method

3.1 Problem Definition

Following [Wang *et al.*, 2022], we consider a continual test-time domain adaptation setting, where a pre-trained model needs to adapt to a continually changing target domain in an

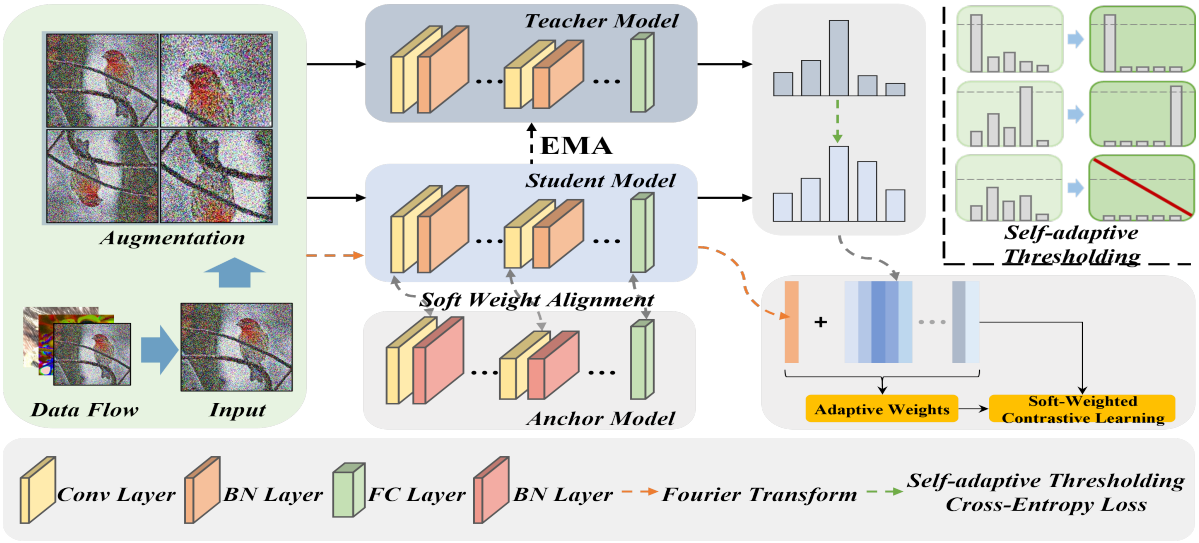


Figure 1: This is the flow of our method. At the time t, the model receives an incoming batch of images from the unlabeled target data flow. After data augmentation, these incoming images are fed into both the teacher and student networks. The average outputs of the augmented images in the teacher network are utilized as self-adaptive thresholding pseudo-labels to supervise the student network. In addition, the parameters of the student network are softly aligned with those of the anchor network to prevent knowledge forgetting in the adaptation process.

online fashion when the absence of source data. Consider a pre-trained model $F_\theta(x)$ with parameter θ trained on the source data $(\mathcal{X}^S, \mathcal{Y}^S)$. Unlabeled target domain data $\mathcal{X}^T$ is provided sequentially and the data distribution of $\mathcal{X}^T$ is continually changing. At inference time t, when the unlabeled target data $X^T(t) = [x_1^T(t), ..., x_B^T(t)]$ is sent to the model F_{θ_t}, where B is the number of samples. The model F_{θ_t} needs to make the prediction $F_{\theta_t}(X^T(t))$ and adapts itself accordingly for the next input ($\theta_t \rightarrow \theta_{t+1}$). It is worth noting that the total evaluation process is online, and the model only has access to the data $X^T(t)$ of the current time step t. The framework is shown in Fig. 1.

3.2 Self-adaptive Thresholding Cross-Entropy Loss

In this section, we detail the self-adaptive thresholding cross-entropy loss. In the process of adapting to a continually changing target domain, the quality of pseudo-labels decreases significantly because of the distribution shift. To improve the quality of pseudo-labels, following [Wang *et al.*, 2022], we use a weight-averaged teacher model $F_{\hat{\theta}}$ to generate the pseudo-labels. At the time $t = 0$, the teacher network is initialized by the pre-trained source model. Then, at time t, the teacher model $F_{\hat{\theta}}$ generates the pseudo-label $\hat{y}^T(t)$ to help the learning process of the student model (adapted model) F_{θ_t}. Specifically, we compute the cross-entropy loss between the output of the student model and the pseudo-label $\hat{y}^T(t)$. The learning process can be denoted as follows:

$$\begin{aligned} \hat{y}_b^T(t) &= \texttt{Softmax}(F_{\hat{\theta}}(x_b^T(t))), \\ \mathcal{L}_{ce} &= -\frac{1}{B}\sum_{b=1}^{B}\sum_{k}\hat{y}_{bk}^T(t)\log y_{bk}^T(t). \end{aligned} \tag{1}$$

$\hat{y}_{bk}^T(t)$ is the soft predication of the teacher model of class $k \in K$, and $y_{bk}^T(t)$ is the prediction of the student model.

During this process, the accumulation of erroneous pseudo-labels can severely disturb the model's predictive performance. This work focuses on pseudo-labeling using cross-entropy loss with a confidence threshold. Instead of simply presenting a constant hyperparameter and treating all classes equally, we advocate that the key to determining thresholds is that thresholds should reflect the test learning status. Thus, we present self-adaptive thresholding that automatically defines and adaptively adjusts the confidence threshold for each class by leveraging the current predictions during test-time training.

The global threshold should represent the model's confidence in the test data, reflecting the overall learning status. We set the global threshold τ_t as the average confidence from the model on test data, and estimate the global confidence as the exponential moving average (EMA) at each time step, where t represents the t-th time step. The global threshold τ_t is defined and adjusted as:

$$\tau_t = \frac{1}{B}\sum_{b=1}^{B}\max(\hat{y}_b^T(t)). \tag{2}$$

In addition to the global threshold, the local threshold is utilized to modulate the global threshold in a class-specific fashion to account for the intra-class diversity and the possible class adjacency. We compute the expectation of the model's predictions on each class k to estimate the class-specific learning status:

$$p_t(k) = \frac{1}{B}\sum_{b=1}^{B}\hat{y}_{bk}^T(t). \tag{3}$$

After integrating the global and local thresholds, we can obtain the final self-adaptive threshold of each class k.

$$\tau_t(k) = \frac{p_t(k)}{\max\{p_t(k) : k \in [K]\}}\tau_t. \tag{4}$$

Finally, we adopt the self-adaptive threshold to select reliable samples for supervision, and the training objective of Eq. 1 can be denoted as follows:

$$\mathcal{L}_{sat} = -\frac{1}{B}\sum_{b=1}^{B}\mathbf{1}(\max(\hat{y}_b^T(t)) > \tau_t(\arg\max \hat{y}_b^T(t))) \cdot \sum_k \hat{y}_{bk}^T(t)\log y_{bk}^T(t). \tag{5}$$

3.3 Soft-weighted Contrastive Learning

Such pseudo-labels can be further utilized to promise the model more substantial representation power, while previous methods have not achieved it. Under the context of contrastive learning, in particular, these semantic class structures can give helpful guidance in selecting contrastive pairs with similar semantics to improve training efficiency. Specifically, the similarity of the samples b and q can be calculated with cosine similarity using pseudo-labels.

$$w_{b,q}(t) = \frac{\hat{y}_b^T(t)(\hat{y}_q^T(t))^\top}{\max(w(t))}. \tag{6}$$

We adopt the weighted similarity matrix w_t to guide the traditional contrastive loss, which can be written as follows:

$$\mathcal{L}_{swc}(x_b^T(t)) = -\log\frac{\exp(z_b^T(t)\cdot\tilde{z}_b^T(t))}{\sum_{q\in N_{neg}(b)}\exp(z_b^T(t)\cdot z_q^T(t))}, \\ -\log\frac{\sum_{q\in N_{pos}(b)} w_{b,q}(t)\exp(z_b^T(t)\cdot z_q^T(t))}{\sum_{q\in N_{neg}(b)}\exp(z_b^T(t)\cdot z_q^T(t))}, \tag{7}$$

where $z_b^T(t) = F_\theta(x_b^T(t))$. We then introduce the components of the objective in detail.

Fourier Augmentation. Some work [Xu *et al.*, 2021; Lu *et al.*, 2022] have demonstrated that Fourier phase information contained high-level semantics and was not easily affected by domain shifts, which may be a kind of universal domain-invariant features. Thus, we introduce the Fourier transform as a new augmentation and attempt to learn domain-invariant representations. The Fourier transformation $\texttt{FFT}(x)$ for a single-channel two-dimensional data x is formulated as:

$$\texttt{FFT}(x)(u,v) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} x(h,w)e^{-2\pi j(\frac{h}{H}u+\frac{w}{W}v)}, \tag{8}$$

where u and v are indices, H and W are the height and the width, respectively. For data $x_b^T(t)$ with several channels, the Fourier transformation for each channel is computed independently to obtain the corresponding phase information $\tilde{x}_b^T(t)$.

$$\tilde{z}_b^T(t) = F_\theta(\tilde{x}_b^T(t)). \tag{9}$$

Positives. Except for introducing Fourier transform as a positive sample, we attempt to present more positive samples by utilizing the correlation between samples during the instantaneous learning process. According to the pseudo-labels output by the teacher model, we select reliable samples of the same class with b as the positive sample set through the learned threshold.

$$N_{pos}(b) = \{q|q\in B, \max(\hat{y}_q^T(t)) > \tau_t(\arg\max\hat{y}_q^T(t)), \arg\max\hat{y}_q^T(t) = \arg\max\hat{y}_b^T(t)\}. \tag{10}$$

Negatives. The traditional contrastive loss strives to maximize the cosine distances between b and every q in the batch. Instead, we argue that not pushing away same-class pairs helps learn better semantically meaningful clusters. Specifically, we adopt the labels to exclude same-class pairs from all negative pairs:

$$N_{neg}(b) = \{q|q\in B, \arg\max\hat{y}_q^T(t) \neq \arg\max\hat{y}_b^T(t)\}. \tag{11}$$

The objective of the contrastive loss can be written as:

$$\mathcal{L}_{swc} = \frac{1}{B}\sum_{b=1}^{B}\mathcal{L}_{swc}(x_b^T(t)). \tag{12}$$

Similar to [Wang *et al.*, 2022], we also use the data augmentation strategy to refine the pseudo-label learning process. Specifically, we combine the original images with its N augmented images into a batch and send it to the student network. Large batch size is conducive to the update of the batch normalization layer and the convergence of the network, especially for the student model in an online update fashion. We use the average predictions for all augmented samples as the pseudo-labels to supervise the learning process of the student model. After the update of student model ($\theta_t \rightarrow \theta_{t+1}$). The teacher model is updated by exponential moving average [Polyak and Juditsky, 1992; Tarvainen and Valpola, 2017] using the weights of the student model:

$$\hat{\theta}_{t+1} = \alpha\hat{\theta}_t + (1-\alpha)\hat{\theta}_{t+1}, \tag{13}$$

where α represents the smoothing factor.

3.4 Soft Weight Alignment

With the help of the pseudo-labels, the pre-trained model can quickly adapt to the target domain and make a more accurate prediction of the target data. However, due to the difference in the data distribution of the source and target domains, there are noisy pseudo-labels in the adaptation process, which leads to instantaneous negative learning. Furthermore, as the model continually adapts to the target domain with changing distributions, the instantaneous errors accumulate, and the knowledge from the pre-trained model is constantly forgotten. For example, the model may not recover after encountering some hard samples, which is detrimental to the test of subsequent data, even if the latter data is not severely shifted. To slow down the accumulation of errors and combat knowledge forgetting, we propose a Soft Weight Alignment (SWA) strategy, which continuously distills knowledge from the source pre-trained model.

Given a pre-trained model F_θ and its two derivative models: anchor model F_{θ_a} and student model F_{θ_t}, and both F_{θ_t}

and F_{θ_a} are initialized by F_θ. In the continual test-time domain adaptation process, the parameters of the anchor model are fixed, and the parameters of the student network are constantly updated. In fact, the anchor model F_{θ_a} is the pre-trained model F_{θ_0} at time $t = 0$. To maintain the knowledge gained from the pre-training model, we soft-align the parameters of the student network F_{θ_t} and the anchor network F_{θ_a}:

$$\mathcal{L}_{swa} = \sum_l \mathbf{1}[l \notin \mathrm{BN}] \cdot \left\|\theta_a^l - \theta_t^l\right\|_2^2, \tag{14}$$

where θ_a^l and θ_t^l are parameters of anchor model F_{θ_a} and student model F_{θ_t} at layer l, respectively. BN represents the Batch Normalization (BN) layer. SWA forces the parameters of the student model to be soft consistent with those of the anchor model, thereby enhancing the anti-forgetting ability of the student model. Specifically, the indicator function $l \notin \mathrm{BN}$ equals 1 if layer l of the anchor or student network is not the BN layer, and 0 otherwise.

Therefore, we use SWA to constrain the update of the student network. The total loss function $\mathcal{L}_T$ can be formulated as follows:

$$\mathcal{L}_T = \mathcal{L}_{sat} + \lambda_1 \mathcal{L}_{swc} + \lambda_2 \mathcal{L}_{swa}, \tag{15}$$

where β_1 and β_2 are the balance hyper-parameters of the total loss function.

4 Experiments

In this section, we review the proposed method on several benchmark tasks: CIFAR10-to-CIFAR10C (Standard and Gradual), CIFAR100-to-CIFAR100C, and ImageNet-to-ImageNet-C. We first introduce several commonly used datasets and our method's implementation details; then, we present several commonly used comparison methods; finally, we report and analyze the results to validate the effectiveness of our approach.

4.1 Datasets

We use CIFAR10, CIFAR100, and ImageNet as the source domain datasets, and CIFAR10C, CIFAR100C, and ImageNet-C as the corresponding target domain datasets, respectively. The target domain datasets were originally created to evaluate the robustness of classification networks [Hendrycks and Dietterich, 2019]. Each target domain dataset contains 15 types of corruption with 5 levels of severity. Following [Wang *et al.*, 2022], for each corruption, we use 10000 images for both CIFAR10C and CIFAR100C datasets and use 5000 images for ImageNet-C.

4.2 Implementation Detail

Following [Wang *et al.*, 2022], the corrupted images are provided to the network online, which means these images can be used to update the model only once in the adaptation process. In addition, different from traditional test-time adaptation methods, which adapt to each corruption type data individually, we adjust the source model to each corruption type sequentially. We evaluate the adaptation performance immediately after encountering each corruption type data. The total type of corruption is set as 15, and the corruption level is set to the highest level of 5 (except for the gradual experiments on CIFAR10-to-CIFAR10C). For CIFAR10-to-CIFAR10C, we use a pre-trained WideResNet-28 [Zagoruyko and Komodakis, 2016] model from the RobustBench benchmark[Croce *et al.*, 2020]. We use Adam to optimize the network and set the learning rate to 1e-3. The data augmentation strategy is the same as [Wang *et al.*, 2022], including color jitter, gaussian blur, gaussian noise, random affine, and random horizontal flip, $N = 8$. CIFAR100-to-CIFAR100C, we use a pre-trained ResNeXt-29 [Xie *et al.*, 2017] from [Hendrycks *et al.*, 2019], $N = 4$. For ImageNet-to-ImageNet-C, we use the standard pre-trained ResNet-50 from RobustBench [Croce *et al.*, 2020], $N = 4$. The experiments on ImageNet-to-ImageNet-C are performed under ten diverse corruption orders. The smoothing factor α is set as 0.99.

4.3 Baselines

We compare our method with several state-of-the-art continual test-time adaptation algorithms, the details of these methods are as follows: 1) **Source:** It directly uses the pre-trained model for adaptation without any specific method for domain adaptation; 2) **BN Stats Adapt:** Batch Normalization Statistics Adaptation method keeps the pre-trained model weights and uses the Batch Normalization statistics from the input data of the input batch for the prediction [Li *et al.*, 2016; Schneider *et al.*, 2020]; 3) **Pseudo-Label [Lee and others, 2013]:** This method picks up the class which has the maximum predicted probability as the pseudo-labels to update the model; 4) **TENT [Wang *et al.*, 2020]:** Test entropy minimization, a test-time entropy minimization scheme to reduce generalization error by reducing the entropy of model predictions on test data; 5) **TENT-continual** is a continual learning version of TENT; 6) **CoTTA [Wang *et al.*, 2022]:** Continual Test-Time Adaptation, which reduces the error accumulation by using weight-averaged and augmentation-averaged predictions and avoids catastrophic forgetting by stochastically restoring a small part of the source pre-trained weights.

4.4 CIFAR10-to-CIFAR10C

Performance Evaluation. Table 1 shows the classification error rate for the standard CIFAR10-to-CIFAR10 task. *Source* method shows the highest average error rate. It depends mainly on the distance between the current corruption-type data distribution and the source domain distribution. *BN Stats Adapt* method is simple and effective, dramatically reducing the evaluation error compared with the *Source* method. Compared with *BN Stats Adapt* and *Pseudo-Label*, the performance of *TENT-continual* is not improved or even decreased. It is mentioned that *TENT-continual* outperforms the *BN Stats Adapt* in earlier stages of the adaptation. However, after adapting to several types of corruption, the performance of *TENT-continual* decreases rapidly. This shows that the *TENT-continual* can not deal with the error accumulation or forgetting problem in continuous adaptation, and may be unstable under long-term continuous adaptation. *CoTTA* takes the error accumulation into account to further improve the performance. Our method achieves the best results in the average error value and most of the corruption-type data.

Time	t ⟶															
Method	*Gaussian*	*shot*	*impulse*	*defocus*	*glass*	*motion*	*zoom*	*snow*	*frost*	*fog*	*brightness*	*contrast*	*elastic_trans*	*pixelate*	*jpeg*	Mean
Source	72.3	65.7	72.9	46.9	54.3	34.8	42.0	25.1	41.3	26.0	9.3	46.7	26.6	58.5	30.3	43.5
BN Stats Adapt	28.1	26.1	36.3	12.8	35.3	14.2	12.1	17.3	17.4	15.3	8.4	12.6	23.8	19.7	27.3	20.4
Pseudo-Label	26.7	22.1	32.0	13.8	32.2	15.3	12.7	17.3	17.3	16.5	10.1	13.4	22.4	18.9	25.9	19.8
TENT-continual [Wang *et al.*, 2020]	24.8	**20.5**	28.5	14.5	31.7	16.2	15.0	19.2	17.6	17.4	11.4	16.3	24.9	21.6	26.0	20.4
CoTTA [Wang *et al.*, 2022]	24.6	21.9	26.5	11.9	27.8	12.4	10.6	15.2	14.4	12.8	**7.4**	11.1	18.7	13.6	17.8	16.5
Ours	**23.9**	**20.5**	**24.5**	**11.2**	**26.3**	**11.8**	**10.1**	**14.0**	**12.7**	**11.5**	7.6	**9.5**	**17.6**	**12.0**	**15.8**	**15.2**

Table 1: Classification error rate (%) for the standard CIFAR10-to-CIFAR10C continual test-time adaptation task. All results are evaluated with the largest corruption severity level 5 in an online fashion. **Bold** text indicates the best performance.

Time	t ⟶															
Method	*Gaussian*	*shot*	*impulse*	*defocus*	*glass*	*motion*	*zoom*	*snow*	*frost*	*fog*	*brightness*	*contrast*	*elastic_trans*	*pixelate*	*jpeg*	Mean
baseline	27.3	23.9	32.1	12.7	30.2	14.1	12.3	17.4	17.0	15.5	9.5	15.0	20.4	15.9	20.9	19.0
+AugTeach	25.6	22.5	27.8	12.6	29.5	14.3	12.5	17.3	16.7	15.4	9.5	15.1	20.9	15.9	19.7	18.4
+AugTeach+SAT	24.5	20.5	25.2	12.0	26.8	13.1	11.5	16.0	14.6	13.2	8.3	11.8	19.3	14.1	16.4	16.5
+AugTeach+SAT+SWC w/o FFT	24.4	20.4	25.2	11.4	26.0	12.5	10.8	15.2	13.6	12.7	8.1	11.4	18.5	13.5	16.0	15.9
+AugTeach+SAT+SWC	23.5	20.5	24.4	11.2	26.2	12.1	10.3	14.3	13.1	12.0	7.5	9.3	17.9	12.9	16.3	15.4
+AugTeach+SAT+SWC+SWA	23.9	20.5	24.5	11.2	26.3	11.8	10.1	14.0	12.7	11.5	7.6	9.3	17.6	12.0	15.8	15.2

Table 2: Ablation experiments for the CIFAR10-to-CIFAR10C task. 'AugTeach' means the pseudo-labels are generated by the teacher model. 'STA' means the self-weighted thresholding cross-entropy loss. 'SWC' means Soft-weighted contrastive learning, and 'SWA' means Soft Weight Alignment strategy.

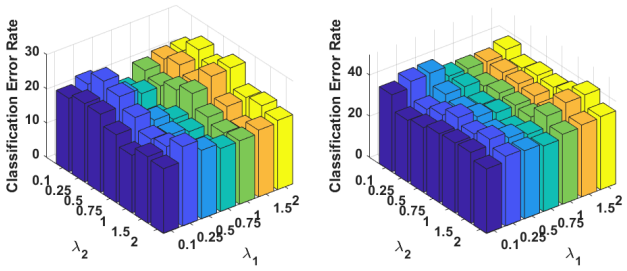
Avg. Error (%)	Source	BN Adapt	TENT-continual [Wang *et al.*, 2020]	CoTTA [Wang *et al.*, 2022]	Ours
CIFAR10C	24.8	13.7	29.2	10.4 ± 0.3	$\mathbf{8.0 \pm 0.5}$

Table 3: Gradually changing setup results on CIFAR10-to-CIFAR10C. The severity level changes gradually between the lowest and the highest. Results are the mean over ten diverse corruption-type sequences. **Bold** text indicates the best performance.

Compared with the strongest baseline *CoTTA*, the average error value of our method is reduced from 16.5% to 15.2%.

Ablation Studies. In addition, we also conducted ablation experiments to prove the effectiveness of each module of our method. The results are shown in Table 2. All modules we propose are helpful for performance gains.

Following [Wang *et al.*, 2022], *baseline* is a combination of *BN Stats Adapt* and *Pseudo-Label*. Then we add the 'AugTeach' strategy to the baseline model, and the average error value dropped from 19.0% to 18.4%. Further, the addition of the 'STA' strategy reduced the model's average error rate from 18.4% to 16.5%, and 'SWC' loss further reduces the average error to 15.5%. It is worth mentioning that after adding FFT augmentations, the test model achieved better results. Next, the addition of 'SWA' has led to an increase in overall performance and a significant reduction in the error of subsequent adaptations. Inevitably, adding this additional constraint will slow down the adaptation, for example, the addition of 'SWA' makes the performance of the first three corrupted data worse than not adding. Finally, We randomly selected a batch of data in the middle and late stages, and the visualization results are shown in Fig. 3. The results demonstrate that the proposed label selection strategy is effective for instantaneous parameter learning.



(a) CIFAR10-to-CIFAR10C (b) CIFAR100-to-CIFAR100C

Figure 2: The influence of λ_1 and λ_2

Parameters Analysis. We also investigate the parameter sensitivity in the proposed method. Fig. 2 represents the change of accuracies with different loss weights, which indicates that our method is insensitive to the parameters λ_1 and λ_2 in the range of [0.1,2]. **Gradually Changing Setup.** Following [Wang *et al.*, 2022], we also consider a gradually changing setup. For the standard setup, corruption types change abruptly in the highest severity. For the gradually changing setup, the corruption types change is gradual.

Time	t →															
Method	*Gaussian*	*shot*	*impulse*	*defocus*	*glass*	*motion*	*zoom*	*snow*	*frost*	*fog*	*brightness*	*contrast*	*elastic_trans*	*pixelate*	*jpeg*	Mean
Source	73.0	68.0	39.4	29.3	54.1	30.8	28.8	39.5	45.8	50.3	29.5	55.1	37.2	74.7	41.2	46.4
BN Stats Adapt	42.1	40.7	42.7	27.6	41.9	29.7	27.9	34.9	35.0	41.5	26.5	30.3	35.7	32.9	41.2	35.4
Pseudo-Label	38.1	36.1	40.7	33.2	45.9	38.3	36.4	44.0	45.6	52.8	45.2	53.5	60.1	58.1	64.5	46.2
TENT-continual [Wang *et al.*, 2020]	**37.2**	**35.8**	41.7	37.7	50.9	48.5	48.5	58.2	63.2	71.4	72.0	83.1	88.6	91.6	95.1	61.6
CoTTA [Wang *et al.*, 2022]	40.1	37.7	39.7	26.8	38.0	27.9	26.5	32.9	31.7	40.4	24.6	26.8	32.5	28.1	33.8	32.5
Ours	39.4	36.4	**37.4**	**25.0**	**36.0**	**26.6**	**25.0**	**29.1**	**28.4**	**35.0**	**23.5**	**25.1**	**28.5**	**25.8**	**29.6**	**30.0**

Table 4: Classification error rate (%) for the standard CIFAR100-to-CIFAR100C continual test-time adaptation task. All results are evaluated with the largest corruption severity level 5 in an online fashion. **Bold** text indicates the best performance.

Avg. Error (%)	Source	BN Adapt	TENT [Wang *et al.*, 2020]	CoTTA [Wang *et al.*, 2022]	Ours
ImageNet-C	82.4	72.1	66.5	63.0 ± 2.3	**62.1 ± 2.3**

Table 5: Average error of standard ImageNet-to-ImageNet-C experiments over 10 diverse corruption sequences. All results are evaluated with the largest corruption severity level 5 in an online fashion.

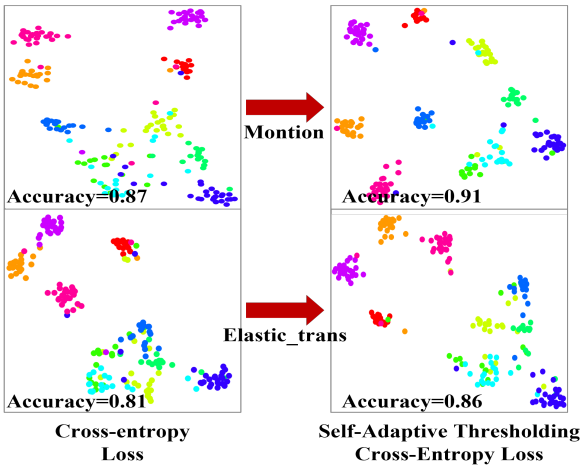


Figure 3: Visualization of the discriminative capability of the sample features on CIFAR10C. Colors represent sample classes.

Specifically, the change process can be expressed as follows:

$$\underbrace{\ldots 2\to 1}_{\text{t-1 and before}} \xrightarrow[type]{change} \underbrace{1\to 2\to 3\to 4\to 5\to 4\to 3\to 2\to 1}_{\text{corruption type t, gradually changing severity}} \xrightarrow[type]{change} \underbrace{1\to 2\ldots}_{\text{t+1 and after}}$$

, where the number represents the corruption severity. The corruption type changes when the severity level is the lowest. In addition, the severity level changes within each type are also gradual.

4.5 CIFAR100-to-CIFAR100C

As shown in Table 4, *TENT-continual* shows the highest average error rate, and its performance is even worse than the *Source* model of doing nothing. *Pseudo-Label* also offers a similar performance degradation curve. In the early stage of adaptation, the performance of *Pseudo-Label* is better than *BN stats adapt* and *Source*, but lower than these two methods in the later stage. This phenomenon is also caused by error accumulation. *CoTTA* considers the problem of error accumulation and reduces the error to 32.5%. Further, we propose a self-adaptive threshold loss to mitigate the instantaneous impact of noisy pseudo-labels on optimization objective, and offer a SWA strategy to alleviate the long-term knowledge forgetting problem in continuous adaptation. The performance of our method is better than *CoTTA* on all corruption types of data, and the average error value is reduced to 30.0%.

4.6 ImageNet-to-ImageNet-C

We also make experiments on ImageNet dataset. Following [Wang *et al.*, 2022], we conduct ImageNet-to-ImageNet-C experiments over ten diverse corruption type sequences in severity level 5. The average result of ten experiments is shown in Table 5. ImageNet is more complex than CIFAR-100 and CIFAR-10, and the overall average test error is also greater. Our method outperforms other competing methods and reduces the average test error to 62.1%.

5 Conclusion

In this paper, we propose a simple yet effective framework for continual test-time domain adaptation, which refines the pseudo-label learning process from the perspective of the instantaneous and long-term impact of noisy pseudo-labels. Firstly, we propose a self-adaptive thresholding Cross-Entropy loss to optimize the adaptation process, which facilitates learning the adapted model. Secondly, the learned thresholds and pseudo-labels are utilized to select more discriminative positive and precise negative pairs for contrastive learning. Finally, we propose a Soft Weight Alignment strategy to normalize the distance between the parameters of the adapted model and the source pre-trained model, which improves the classification accuracy in the late stage of adaptation. Extensive experimental results demonstrate that our method achieves state-of-the-art performance on various benchmark datasets.

Acknowledgements

Our work was supported by Joint Fund of Ministry of Education of China (8091B022149), Key Research and Development Program of Shaanxi (2021ZDLGY01-03), National

Natural Science Foundation of China (62132016, 62171343, 62071361 and 62201436), and Fundamental Research Funds for the Central Universities (ZDRC2102 and ZYTS23135).

Contribution Statements

Xu Yang and Yanan Gu contribute equally. They jointly designed the experiments and wrote the manuscript. Kun Wei and Cheng Deng contributed to the writing of the manuscript. Cheng Deng supervise the project.

References

[Chen *et al.*, 2022] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. *arXiv preprint arXiv:2204.10377*, 2022.

[Croce *et al.*, 2020] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

[Cui *et al.*, 2020] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3941–3950, 2020.

[Dong *et al.*, 2020] Jiahua Dong, Yang Cong, Gan Sun, Bineng Zhong, and Xiaowei Xu. What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4022–4031, June 2020.

[Dong *et al.*, 2021] Jiahua Dong, Yang Cong, Gan Sun, Zhen Fang, and Zhengming Ding. Where and how to transfer: Knowledge aggregation-induced transferability perception for unsupervised domain adaptation. pages 1–1, 2021.

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[Grandvalet and Bengio, 2004] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

[Hendrycks and Dietterich, 2019] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[Hendrycks *et al.*, 2019] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

[Kundu *et al.*, 2020] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4544–4553, 2020.

[Kurmi *et al.*, 2021] Vinod K Kurmi, Venkatesh K Subramanian, and Vinay P Namboodiri. Domain impression: A source data free domain adaptation method. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 615–625, 2021.

[Lee and others, 2013] Dong-Hyun Lee et al. 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 896, 2013.

[Li *et al.*, 2016] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016.

[Li *et al.*, 2020] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9641–9650, 2020.

[Liang *et al.*, 2020] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020.

[Liu *et al.*, 2021] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1215–1224, 2021.

[Lu *et al.*, 2022] Wang Lu, Jindong Wang, Haoliang Li, Yiqiang Chen, and Xing Xie. Domain-invariant feature exploration for domain generalization. *arXiv preprint arXiv:2207.12020*, 2022.

[Polyak and Juditsky, 1992] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[Prabhu *et al.*, 2021] Viraj Prabhu, Shivam Khare, Deeksha Kartik, and Judy Hoffman. Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8558–8567, 2021.

[Saito *et al.*, 2019] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019.

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

[Sun *et al.*, 2019] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[Tzeng *et al.*, 2017] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[Wang *et al.*, 2020] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

[Wang *et al.*, 2022] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. *arXiv preprint arXiv:2203.13591*, 2022.

[Xie *et al.*, 2017] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[Xu *et al.*, 2021] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021.

[Yang *et al.*, 2019] Xu Yang, Cheng Deng, Feng Zheng, Junchi Yan, and Wei Liu. Deep spectral clustering using dual autoencoder network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[Yang *et al.*, 2021] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8978–8987, 2021.

[Yang *et al.*, 2022] Xu Yang, Cheng Deng, Tongliang Liu, and Dacheng Tao. Heterogeneous graph attention network for unsupervised multiple-target domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1992–2003, 2022.

[Yeh *et al.*, 2021] Hao-Wei Yeh, Baoyao Yang, Pong C Yuen, and Tatsuya Harada. Sofa: Source-data-free feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 474–483, 2021.

[Zagoruyko and Komodakis, 2016] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[Zhou and Levine, 2021] Aurick Zhou and Sergey Levine. Training on test data with bayesian adaptation for covariate shift. *arXiv preprint arXiv:2109.12746*, 2021.