Navigating Continual Test-time Adaptation with Symbiosis Knowledge

Xu Yang , **Moqi Li** , **Jie Yin** , **Kun Wei** and **Cheng Deng***
Xidian University
{xuyang.xd, moqili14, weikunsk, chdeng.xd}@gmail.com, yinjie_xidian@163.com

Abstract

Continual test-time domain adaptation seeks to adapt the source pre-trained model to a continually changing target domain without incurring additional data acquisition or labeling costs. Unfortunately, existing mainstream methods may result in a detrimental cycle. This is attributed to noisy pseudo-labels caused by the domain shift, which immediately negatively impacts the model's knowledge. The long-term accumulation of these negative effects exacerbates the model's difficulty in generalizing to future domain shifts and contributes to catastrophic forgetting. To address these challenges, this paper introduces a Dual-stream Network that independently optimizes different parameters in each stream to capture symbiotic knowledge from continual domains, thereby ensuring generalization while enhancing instantaneous discrimination. Furthermore, to prevent catastrophic forgetting, a weighted soft parameter alignment method is designed to leverage knowledge from the source model. Finally, efforts are made to calibrate and explore reliable supervision signals to mitigate instantaneous negative optimization. These include label calibration with prior knowledge, label selection using self-adaptive confidence thresholds, and a soft-weighted contrastive module for capturing potential semantics. 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 testing data obey the same distribution. Such networks, however, suffer from the generalization problem due to the ubiquitous domain shift [Wang *et al.*, 2023b]. For example, a classification network pre-trained in the normal, natural images may not recognize the corrupted images. Thus, domain adaptation is essential to transfer knowledge from the source domain to the target one by reducing the shift. However, the target domain labels are usually unavailable, and the problem is primarily explored at *Unsupervised Domain Adaptation* (UDA) [Li *et al.*, 2020; Wang *et al.*, 2023c; Yang *et al.*, 2022]. More realistically, the source data is often inaccessible during test time due to privacy or business problems, making the adaptation problem more challenging. Initial approaches attempt to employ the source model and unlabeled target data for testing, such as Source-Free and Test-Time domain Adaptation (TTA) [Chen *et al.*, 2022; Yang *et al.*, 2021; Liu *et al.*, 2021].

Common techniques in Test-Time Adaptation (TTA) typically address the challenge of domain shift by updating adapted model parameters using either generated pseudo-labels or entropy regularization [Yang *et al.*, 2023]. While effective for static target distributions, these approaches exhibit instability when the target domain's distribution is in a continual state of flux [Wang *et al.*, 2022; Prabhu *et al.*, 2021]. The presence of noisy pseudo-labels, stemming from the constantly shifting distribution, significantly hampers the adaptation process [Wang *et al.*, 2022; Prabhu *et al.*, 2021]. To address this issue, CoTTA [Wang *et al.*, 2022] introduces the concept of Continual Test-Time Domain Adaptation, wherein a source pre-trained model must adapt to an evolving stream of target domains without recourse to source data. CoTTA leverages a weight-average teacher network to enhance the quality of generated pseudo-labels. Additionally, Robust Mean Teacher [Döbler *et al.*, 2023] employs a multi-viewed contrastive loss to guide test features back towards the initial source space and learns invariant features concerning the input space. Nevertheless, recent studies [Marsden *et al.*, 2023] have demonstrated that tuning network parameters based solely on the current domain may result in a loss of generalization and impair performance on subsequent domains. Some strategies [Gong *et al.*, ; Marsden *et al.*, 2023] advocate for updating only the network's normalization parameters while freezing all others, which can mitigate the rapid loss of generalization. However, this approach may lead to a lack of discriminative power in certain domains due to the constrained learning parameters.

The aforementioned methods have motivated our primary research objective: enhancing network discrimination within the current domain while preserving generalization for subsequent domains. Our focus lies in developing a dual-stream architecture leveraging distinct optimization parameters to encapsulate synergistic knowledge of

*Corresponding Author

generalization and discrimination. One stream of this dual-stream network exclusively adjusts normalization parameters to maintain generalization prowess, while the other stream harnesses all learnable parameters to enhance discriminative capacity. Furthermore, to mitigate the risk of catastrophic forgetting, we propose continual integration of source knowledge into each stream using varied strategies to fine-tune the adapted dual-stream model. Specifically, a weight-average strategy is employed in both the source and adapted models for the normalization-tuned stream. To achieve this, we introduce a weighted soft parameter alignment mechanism that encourages similarity between the adapted network and the source. Importantly, the absence of supervisory signals in generated pseudo-labels inevitably introduces noise. We propose a novel weighted guidance approach, inspired by the observation that latter layers in a network are more susceptible to label noise while former layers exhibit greater robustness [Bai *et al.*, 2021]. These weights modulate the similarity between the adapted model and the source model across layer depths, enabling noise-resistant former layers to undergo more adjustment and noise-sensitive latter layers to undergo less adjustment.

The second objective of this paper is to uncover prior knowledge aimed at enhancing supervision signals. As noisy pseudo-labels accumulate over time, the model's discriminative capacity faces significant challenges. In practical scenarios, the distribution of features among samples often reflects their semantic properties, yet this knowledge remains largely unexplored. To address this gap, we leverage source predictions to approximate the lower and upper bounds of individual class probabilities, facilitating the calibration of pseudo-labels and averting trivial solutions. Additionally, we employ global and local strategies to independently determine thresholds for each class, thereby selecting more reliable pseudo-labels. Subsequently, we construct a soft-weighted contrastive learning module based on these dependable components, which brings potential same-class samples closer while effectively discriminating against unrelated samples. This approach harnesses various priors to enhance the model's adaptability to target domains.

Contributions. The highlights of the paper are threefold: 1) By analyzing the update properties of parameters, we design a dual-stream framework for continual test-time domain adaptation. The different learnable parameters are tuned in each stream to form a symbiotic knowledge for long-term generalization and instantaneous discrimination, while the source parameters are appropriately introduced into dual streams to alleviate catastrophic forgetting; 2) We explore reliable supervision signals with prior knowledge to guide the test time tuning. The lower and upper bounds of individual class probability are employed to calibrate the pseudo-labels, and the reliable samples are selected in a self-adaptive manner. The feature distribution of the source pre-trained model is adopted to guide a soft-weighted contrastive module for capturing potential semantics lost during adaptation; 3) Extensive experimental results demonstrate that our method achieves state-of-the-art performance on several datasets. The ablation experiments are conducted to verify the effectiveness of each module.

2 Related Work

2.1 Domain Adaptation

Domain adaptation [Cao *et al.*, 2022; Jiang *et al.*, 2021] refers to acquiring knowledge from labeled data within the source domain to achieve effective performance across diverse yet related target domains. A fundamental challenge in domain adaptation lies in the misalignment between the feature and label spaces of the source and target domains [Wang *et al.*, 2023a; Ding *et al.*, 2023; Xie *et al.*, 2022]. To tackle this challenge, certain domain adaptation techniques aim to guide deep models in learning domain-invariant representations [Sun *et al.*, 2022] and classifiers [Wang *et al.*, 2023c]. Notably, some methodologies [Ganin and Lempitsky, 2015; Tzeng *et al.*, 2017; Ganin *et al.*, 2016] employ adversarial training to align feature distributions with a domain discriminator, while others impose constraints on the cross-domain feature space, such as entropy constraint [Saito *et al.*, 2019] or maximum prediction rank [Cui *et al.*, 2020]. It's worth noting that all the aforementioned methods necessitate access to both source and target data during the adaptation process, rendering the learning transductive.

2.2 Test-Time Domain Adaptation

In recent studies on test-time domain adaptation, attention has shifted towards a more demanding scenario where solely the source model and unlabeled target samples are accessible. Certain test-time domain adaptation approaches [Li *et al.*, 2020] leverage generative models to effectuate feature alignment between the source and target domains without requiring additional source data acquisition. Additionally, some methodologies achieve test-time domain adaptation by refining the source model with the aid of target data, obviating the need for explicit domain alignment. Test Entropy Minimization (TENT) [Wang *et al.*, 2020] introduces entropy minimization as a test-time optimization objective, wherein normalization statistics are estimated, and channel-wise affine transformations are optimized online with each batch update. Source Hypothesis Transfer (SHOT) [Liang *et al.*, 2020] endeavors to learn the optimal target-specific feature learning module to align with the source hypothesis.

Most test-time adaptation methodologies primarily address the offline scenario, wherein the complete test data set is available during the training phase. However, CoTTA [Wang *et al.*, 2022] extends test-time adaptation from the offline setting to an online continual scenario. This extension tackles a more challenging yet realistic problem termed *Continual Test-Time Domain Adaptation*, wherein a source pre-trained model must adapt to a continuously evolving stream of target test data without access to any source data. [Gan *et al.*, 2022] employs visual domain prompts to dynamically update a small portion of input image pixels, thereby mitigating the issue of error accumulation. Additionally, NOTE [Gong *et al.*,] introduces instance-aware batch normalization to rectify normalization for out-of-distribution samples.

Our Study. Our proposed method incorporates an efficient dual-stream network to ensure long-term generalization and improve instantaneous discrimination for continual test-time

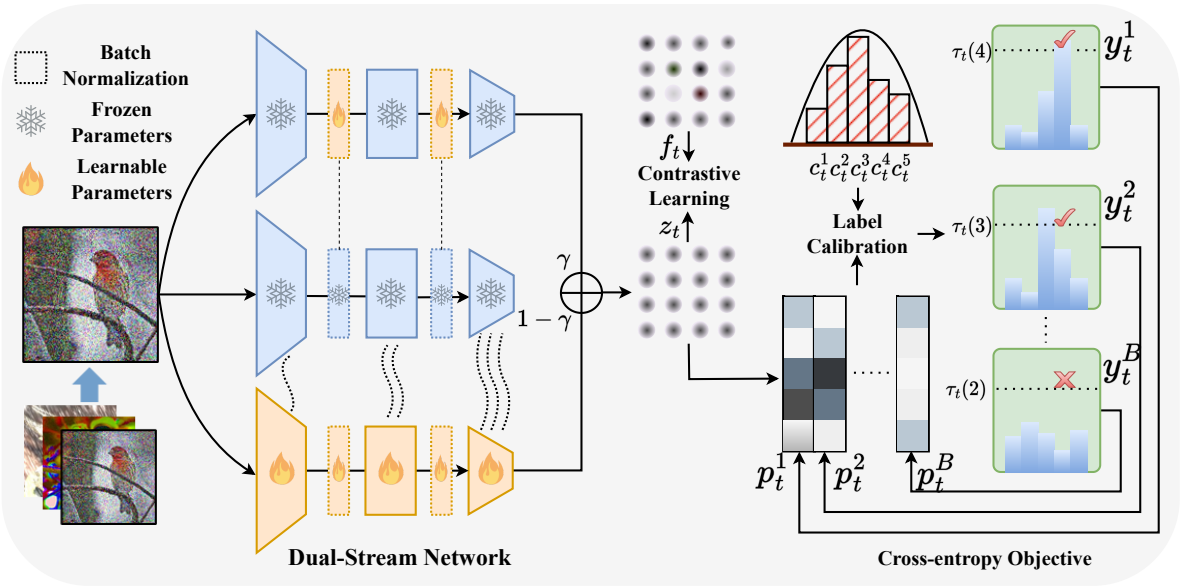

Figure 1: This is the flow of our method. We propose a Dual-stream Network to form a symbiotic knowledge with different parameters tuned in each stream, ensuring long-term generalization and instantaneous discrimination. Meanwhile, we explore various prior knowledge from the source pre-trained model to calibrate and enrich supervision signals. The lower and upper bounds of individual class probability are employed to calibrate the pseudo-labels, and a confidence threshold in a self-adaptive manner is utilized to select reliable labels. Finally, the source pre-trained model features are adopted to construct a soft-weighted contrastive module for capturing potential semantics lost during adaptation.

domain adaptation, and we first explore prior knowledge from the source pre-trained model to guide the adaptation stages.

3 Proposed Method

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 online without source data. Consider a pre-trained model $F_\theta(x)$ with parameter θ trained on the source data. Unlabeled target domain data X_t is provided sequentially, and the data distribution continually changes. At testing stage t, when the unlabeled target data $X_t = [x_t^1, ..., x_t^B]$ 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 $P_t = [p_t^1, ..., p_t^B]$ 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 of the current stage t. We design a dual-stream network, which optimizes different parameters independently in each stream, to capture knowledge from continual domains. Meanwhile, we explore prior knowledge from the source pre-trained model. The framework is shown in Figure 1.

3.1 Dual-stream Network

We first design a new dual-stream pipeline for continual test-time domain adaptation to capture symbiosis knowledge of generalization and discrimination. For the convenience of expression, θ_t in the following mainly refers to the parameters of the dual-stream network. The parameters θ_t of the dual-stream network are divided into two parts $\{\hat{\theta}_t, \bar{\theta}_t\}$, where only the batch normalization layers are tuned in $\hat{\theta}_t$ and all parameters are tuned in $\bar{\theta}_t$, and the learning process can be denoted as follows.

$$p_t^b = \texttt{Softmax}(\gamma F_{\hat{\theta}_t}(x_t^b) + (1-\gamma) F_{\bar{\theta}_t}(x_t^b)),$$
$$\mathcal{L}_{ce}(X_t) = -\frac{1}{B}\sum_{b=1}^{B}\sum_{k} y_t^{b,k} \log p_t^{b,k}, \tag{1}$$

where p_t^b represents classification result of the sample b at time t, and $k \in K$ is the k-th class. The learnable parts can improve the instantaneous discriminative adaptability and long-term generalization adaptability of the model. We use the predictions of the source model as auxiliary information to predict γ. The basic idea is that a similar prediction with the source model may represent a higher reliability stream.

$$\gamma = \frac{\text{sim}(F_{\hat{\theta}_t}(x_t^b), F_\theta(x_t^b))}{\text{sim}(F_{\hat{\theta}_t}(x_t^b), F_\theta(x_t^b)) + \text{sim}(F_{\bar{\theta}_t}(x_t^b), F_\theta(x_t^b))}, \tag{2}$$

where $\text{sim}(\cdot)$ is the cosine similarity of smaples. Previous literature [Frankle *et al.*, 2020] supports the averages of the parameters of the source model potentially has good generalization capabilities and the adapted model. Thus, we continually ensemble the parameters of the initial source model and the weights of the current model using an Exponential Moving Average of the form.

$$\hat{\theta}_{t+1} = \alpha\hat{\theta}_t + (1-\alpha)\theta, \tag{3}$$

where $\alpha = 0.99$ is a momentum term. For the stream that optimizes all network parameters, we hope that the objective function can be employed to directly guide the parameter transfer of the source model and the adapted one, and the

Weighted Soft Parameter Alignment can be defined as follows.

$$\mathcal{L}_{wspa}(\bar{\theta}_t) = \sum_l \mathbf{1}[l \notin \mathrm{BN}] \cdot \beta^l \left\|\bar{\theta}_t^l - \theta^l\right\|_2^2, \tag{4}$$

where l is the layer of the network and β^l repressents the similarity strength of l-th layer. BN represents the batch normalization layers. β^l is increased with the deeper layers. During this process, the accumulation of noisy labels inevitably misleads parameter learning. The following sections will detail the construction of the supervision signals.

3.2 Supervision Signals

Label Calibration with Prior Knowledge. We hope to extract prior knowledge from the source pre-training model to optimize the prediction results. The first is the rough estimation of class distribution at each stage t. To achieve this, the samples are fed into the source pre-trained network, and the occurrence probability of each class c_t^k is then calculated.

$$\begin{aligned} q_t^b &= \arg\max(\texttt{Softmax}(F_\theta(x_t^b))), \\ c_t^k &= \sum_b \mathbf{1}(q_t^b = k)/B, \end{aligned} \tag{5}$$

where q_t^b is the class prediction of the sample b. Here, we choose the source pre-trained model to estimate this probability mainly because its parameters are fixed and not affected by pseudo-labels. Admittedly, such an estimation is not entirely accurate, so we need to relax this estimation to calibrate the supervision labels. The objective is defined as follows.

$$\begin{aligned} &\hat{P}_t = \max_{H_t} < H_t, P_t >, \\ &\text{s.t.} \begin{cases} \sum_k h_t^{b,k} = 1, \forall b \in B \\ h_t^{b,k} \in \{0,1\}, \forall k \in K, b \in B, \\ (1-\delta)c_t^k \le \sum_b h_t^{b,k}/B \le (1+\delta)c_t^k, \forall k \in K \end{cases} \end{aligned} \tag{6}$$

where $<,>$ is the inner product, and δ is a relaxation factor, making the probability of each class in a limited range. $H_t = [h_t^1, ..., h_t^B]$ is the variable that needs to be solved and obeys three constraints. The first two constraints ensure that the result conforms to the one-hot distribution of groundtruth. δ is a relaxation factor, making the probability of each class in a limited range with c_t^k. The objective is a Zero-One Programming problem and can be solved with standard solvers [Wolsey, 2020]. We combine the calibrated pseudo-labels with the original predictions to ensure model stability while employing an adaptive threshold to select the final supervision labels.

$$Y_t = (\hat{P}_t + P_t)/||\hat{P}_t + P_t||. \tag{7}$$

Label Selection with Self-adaptive Thresholds. The calibration can suppress noisy labels, but cannot eliminate them. Thus, we adopt a confidence threshold to filter reliable labels. Thus, we present self-adaptive thresholding that automatically defines and adaptively adjusts the confidence threshold for each class by leveraging the current predictions during adaptation. The global threshold should represent the confidence of the model, reflecting the overall learning status. We set the global threshold τ_t as the average confidence from the model, and estimate the global confidence at each stage t. τ_t is defined and adjusted as:

$$\tau_t = \frac{1}{B}\sum_{b=1}^{B} \max(y_t^b). \tag{8}$$

Except for 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:

$$\xi_t(k) = \frac{1}{B}\sum_{b=1}^{B} y_t^{b,k}. \tag{9}$$

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

$$\tau_t(k) = \frac{\xi_t(k)}{\max\{\xi_t(k) : k \in K\}}\tau_t. \tag{10}$$

Based on such thresholds, the samples at current batch can be divided into two parts, the reliable part $N_{rel}(t) = \{b|b \in B, \max(y_t^b) \ge \tau_t(\arg\max y_t^b)\}$ and unreliable one $N_{unrel}(t) = \{b|b \in B, \max(y_t^b) < \tau_t(\arg\max y_t^b)\}$. The objective of $\mathcal{L}_{rce}(X_t)$ can be denoted as follows:

$$\mathcal{L}_{rce}(X_t) = -\frac{1}{|N_{rel}(t)|}\sum_{b\in N_{rel}(t)}\sum_k y_t^{b,k}\log p_t^{b,k} \tag{11}$$

Soft-weighted Contrastive Learning. Undeniably, the source pre-trained model is fully trained with labels, so even if the domain shift causes the classification results to be biased, it is still a suitable feature extractor. In other words, the source domain training model can still judge samples' similarity. Based on this, we design a contrastive learning framework to improve the discriminative ability of the model further. Specifically, we first exploit the source pre-trained model to extract the sample features and establish a similarity matrix.

$$f_t^b = F_\theta(x_t^b), w_t^{b,d} = \mathrm{sim}(f_t^b, f_t^d), \tag{12}$$

where d represents the dth sample at time step t. Such a matrix can be further utilized to promise the model more substantial representation power, while previous methods have not achieved it. Moreover, in 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. We adopt the weighted similarity matrix w_t to guide the traditional contrastive loss, which can be rewritten as follows,

$$\mathcal{L}_{swcl}(X_t) = -\frac{1}{B}\sum_{b=1}^{B}\log\frac{\sum_{d\in N_{pos}(b)} w_t^{b,d}\exp(z_t^b \cdot z_t^d)}{\sum_{d\in N_{neg}(b)}\exp(z_t^b \cdot z_t^d)}, \tag{13}$$

where $z_t^b = \gamma F_{\hat{\theta}_t}(x_t^b) + (1-\gamma)F_{\bar{\theta}_t}(x_t^b)$. We then introduce the components of the objective function in detail.

Positives. We attempt to present more potential positive samples by utilizing the correlation between samples during the instantaneous learning process. We select the samples of the same class with b as the positive sample set from the calibrated labels Y_t.

$$N_{pos}(b) = \{d | d \in N_{rel}(t), \arg\max y_t^d = \arg\max y_t^b\}. \tag{14}$$

Negtatives. The traditional contrastive loss strives to maximize the cosine distances between b and every d 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 reliable same-class pairs from all negative pairs:

$$\begin{aligned} N_{neg}(b) = &\{d | d \in N_{rel}(t), \arg\max y_t^d \neq \arg\max y_t^b\} \cup \\ &\{d | d \in N_{unrel}(t), \arg\max y_t^d \neq \arg\max y_t^b\}. \end{aligned} \tag{15}$$

Here, we believe that the confidence of the selected same-class samples is higher, so these samples are excluded from the negative samples. The same-class samples with lower confidence are considered potentially similar samples. Therefore, we remove them from the positives and negatives, expecting to optimize their distribution using the soft weights transfer.

3.3 Overall

The overall objective of our method is as follows.

$$\begin{aligned} \mathcal{L}(X_t) &= \mathcal{L}_{rce}(X_t) + \lambda_1 \mathcal{L}_{wspa}(\bar{\theta}_t) + \lambda_2 \mathcal{L}_{swcl}(X_t), \\ \hat{\theta}_{t+1} &= \alpha \hat{\theta}_t + (1-\alpha)\theta, \end{aligned} \tag{16}$$

where λ_1 and λ_2 are hyperparameters. In general, we do not directly use the results of pre-trained models as supervision signals, but apply them as prior knowledge to calibrate pseudo-labels, and design a soft-weighted contrastive learning method. In order to prevent the influence of noisy labels, adaptive thresholds are devised to select reliable samples.

4 Experiments

In this section, we evaluate the effectiveness of the proposed method on three benchmark datasets in terms of 1) whether our dual-stream network learns meaningful results, 2) whether the proposed label selection and correction strategies can improve the discrimination, and 3) the parameters analysis of the proposed method.

4.1 Datasets

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

4.2 Implementation Details

Following [Wang *et al.*, 2022], the corrupted images are provided to the network online, which means these images can be utilized to update the model only once in the adaptation process. In addition, unlike 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 at 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. For CIFAR100-to-CIFAR100C, we use a pre-trained ResNeXt-29 [Xie *et al.*, 2017] from [Hendrycks *et al.*, 2019]. For ImageNet-to-ImageNet-C, we use the standard pre-trained ResNet-50 from RobustBench [Croce *et al.*, 2020]. The experiments on ImageNet-to-ImageNet-C are performed under ten diverse corruption orders. The relaxation factor δ is set as 0.2, $\lambda_1 = 0.1$ and $\lambda_2 = 1$ in our experiments. We set $\beta^l = \frac{1-e^{-5l}}{1+e^{-5l}}$ and l is the number of layers.

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** directly uses the pre-trained model for adaptation without any specific method for domain adaptation; 2) **BN Stats Adapt** 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] picks up the class which has the maximum predicted probability as the pseudo-labels to update the model; 4) **TENT** [Wang *et al.*, 2020] reduces generalization error by reducing the entropy of model predictions on test data, **TENT-continual** is a continual learning version of TENT; 5) **CoTTA** [Wang *et al.*, 2022] 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; 6) **NOTE** [Gong *et al.*,] adopts an Instance-Aware Batch Normalization to correct normalization for out-of-distribution samples; 7) **RoTTA** [Yuan *et al.*, 2023] presents a robust batch normalization scheme to estimate the normalization statistics; 8) **RMT** [Döbler *et al.*, 2023] uses symmetric cross-entropy and contrastive learning to pull the test feature space closer to the source domain; 9) **ROID** [Marsden *et al.*, 2023] proposes to continually weight-average the source and adapted model, and an adaptive additive prior correction scheme.

4.4 Performance Evaluation

CIFAR10-to-CIFAR10C. Table 1 shows the classification error rate for the standard CIFAR10-to-CIFAR10C task. We

Time	t ⟶																
Method	*Gaussian*	*Shot*	*Impulse*	*Defocus*	*Glass*	*Motion*	*Zoom*	*Snow*	*Frost*	*Fog*	*Brightness*	*Contrast*	*Elastic_trans*	*Pixelate*	*Jpeg*	Mean	Gain
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	+23.1
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	+23.7
TENT-continual [ICLR'21]	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	+23.1
CoTTA [CVPR'22]	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	+27.0
NOTE [NeurIPS'22]	**7.3**	7.4	**12.5**	20.9	**13.8**	15.5	34.2	34.2	39.6	25.0	11.6	24.2	29.9	14.1	**12.7**	20.1	+23.4
RoTTA [CVPR'23]	30.3	25.4	34.6	18.3	34.0	14.7	11.0	16.4	14.6	14.0	8.0	12.4	20.3	16.8	19.4	19.3	+24.2
RMT [CVPR'23]	24.1	20.2	25.7	13.2	25.5	14.7	12.8	16.2	15.4	14.6	10.8	14.0	18.0	14.1	16.6	17.0	+26.5
ROID [2023.6.1]	23.7	18.7	26.4	11.5	28.1	12.4	10.1	14.7	14.3	12.0	7.5	9.3	19.8	14.5	20.3	16.2	+27.3
ViDA [ICLR'24]	52.9	47.9	19.4	**11.4**	31.3	13.3	**7.6**	**7.6**	**9.9**	12.5	**3.8**	26.3	**14.4**	33.9	18.2	20.7	+22.8
Ours	19.7	15.7	19.6	12.6	23.8	**11.6**	10.3	12.8	11.8	**9.8**	7.8	**9.3**	16.9	**11.2**	15.8	**13.9**	**+29.6**

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. Blue is the suboptimal solution.

Time	t ⟶																
Method	*Gaussian*	*Shot*	*Impulse*	*Defocus*	*Glass*	*Motion*	*Zoom*	*Snow*	*Frost*	*Fog*	*Brightness*	*Contrast*	*Elastic_trans*	*Pixelate*	*Jpeg*	Mean	Gain
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	+11.0
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	+0.2
TENT-continual [ICLR'21]	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	-15.2
CoTTA [CVPR'22]	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	+13.9
NOTE [NeurIPS'22]	**28.4**	32.7	36.4	44.4	42.9	42.2	65.8	61.1	70.8	51.6	34.4	45.4	62.7	39.9	36.4	43.3	+3.1
RoTTA [CVPR'23]	49.1	44.9	45.5	30.2	42.7	29.5	26.1	32.2	30.7	37.5	24.7	29.1	32.6	30.4	36.7	34.8	+11.6
RMT [CVPR'23]	40.2	36.2	36.0	27.9	33.9	28.4	26.4	28.7	28.8	31.1	25.5	27.1	28.0	26.6	29.0	30.2	+16.2
ROID [2023.6.1]	36.5	31.9	33.2	24.9	34.9	26.8	24.3	28.9	28.5	31.1	22.8	24.2	30.7	26.5	34.4	29.3	+17.1
ViDA [ICLR'24]	50.1	40.7	**22.0**	**21.2**	45.2	**21.6**	**16.5**	**17.9**	**16.6**	**25.6**	**11.5**	29.0	29.6	34.7	27.1	**27.3**	**+19.1**
Ours	33.8	**31.8**	30.5	25.5	**30.9**	25.5	25.7	27.0	27.3	30.6	25.9	**22.9**	**26.6**	**26.0**	**26.9**	27.8	+18.4

Table 2: 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. Blue is the suboptimal solution.

Avg. Error (%)	Source	BN Adapt	TENT-continual	CoTTA	RoTTA	RMT	ROID	Ours
ImageNet-C	82.4	72.1	66.5	63.0	67.3	59.9	54.5	**50.3**

Table 3: 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. **Bold** text indicates the best performance. Blue is the suboptimal solution.

compare our method with the seven baseline methods. 'Gain' represents the percentage of improvement in model accuracy compared with the source method. *CoTTA* considers the error accumulation to improve performance further. As the latest proposed methods, *NOTE* attempts to improve the performance of the model in different domains from the distribution with BN. Although it performs well in domains such as Gaussian and shot, it performs poorly in some simple domains, such as Brightness and Contrast. *ROID* has dramatically improved the overall performance of the model. However, the model does not perform well in some difficult domains due to the limited parameters that can be learned. Compared with all the previous methods, our method achieves the best results in the average error value and most of the corruption-type data.

CIFAR100-to-CIFAR100C. Table 2 shows the classification error rate for the standard CIFAR100-to-CIFAR100C task. *BN Stats Adapt* and *NOTE* do not bring error accumulation, but there is little room for improvement. *CoTTA* considers the error accumulation problem and reduces the error to 32.5%. Similarly, Visual Domain Prompt performs well in some domains, but in some relatively complex domains, the limited learnable parameters lead to a limited upper bound of the model. Further, the performance of our method is better than *RMT* and *ROID* on several corruption types of data, and the average error value is reduced to 27.8%.

ImageNet-to-ImageNet-C. We also make experiments on the 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 3. ImageNet is more complex than CIFAR-100 and CIFAR-10, and the overall average test error is more significant. Our method outperforms other competing methods and reduces the average test error to 50.3%.

Time	$t \longrightarrow$															
Method	*Gaussian*	*Shot*	*Impulse*	*Defocus*	*Glass*	*Motion*	*Zoom*	*Snow*	*Frost*	*Fog*	*Brightness*	*Contrast*	*Elastic_trans*	*Pixelate*	*Jpeg*	Mean
Stream 1 w/o EMA	27.5	24.8	28.9	12.0	32.8	13.6	11.2	16.9	12.8	10.2	7.9	12.2	18.5	13.8	17.5	17.4
Stream 1	25.8	22.2	27.0	11.3	29.5	13.1	10.6	15.8	12.0	10.1	7.8	12.0	17.6	11.5	15.5	16.1
Stream 2 w/o WSPA	23.3	20.4	25.0	13.8	30.5	13.9	12.8	15.5	14.6	15.4	8.0	12.4	22.4	18.2	19.4	17.7
Stream 2	21.3	17.8	22.7	13.2	26.8	13.2	11.5	14.7	13.2	10.9	8.0	10.2	18.8	14.5	16.8	15.6
0.5*Stream 1+ 0.5*Stream 2	20.5	17.3	21.2	12.6	24.5	11.9	10.9	13.6	11.8	10.6	8.5	10.6	17.5	11.0	15.3	15.0
γ*Stream 1+ $(1-\gamma)$*Stream 2	19.7	15.7	19.6	12.6	23.8	11.6	10.3	12.8	11.8	9.8	7.8	9.3	16.9	11.2	15.8	13.9

Table 4: Ablation experiments of the Dual-Stream Nerwotk for the CIFAR10-to-CIFAR10C task.'Stream 1' represents the stream in that only the batch normalization layers are tuned, and 'Stream 2' is the stream in which all parameters are tuned. EMA represents the Exponential Moving Average, and WSPA is Weighted Soft Parameter Alignment.

Method	Mean	Gain
CE	16.5	-
CE w/ CAL	16.0	+0.5
CE w/ SEL	16.2	+0.3
CE w/ SEL(CAL)	15.3	+1.2
CE w/ SEL(CAL)+CL	14.7	+1.8
CE w/ SEL(CAL)+SwCL	13.9	+2.6

Table 5: Ablation experiments of the supervision signals for the CIFAR10-to-CIFAR10C task. 'CAL' is the calibration, and 'SEL' represents the Label Selection. 'CL' is traditional contrastive learning, and 'SwCL' is the proposed Soft-weighted Contrastive Learning.

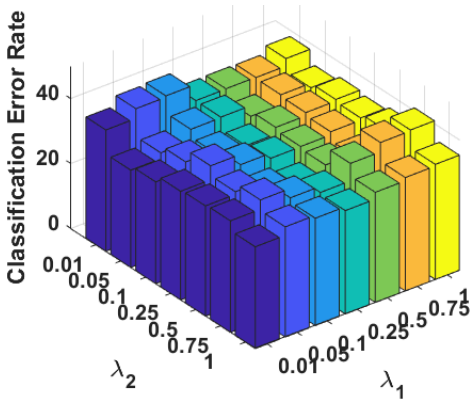

Figure 2: Parameters Analysis of λ_1 and λ_2 on CIFAR10-CIFAR10C dataset.

4.5 Ablation Studies

In addition, we first conduct ablation experiments with the same supervision signals to prove the effectiveness of the dual-stream network. For the convenience of expression, 'Stream 1' represents the stream in that only the batch normalization layers are tuned, and 'Stream 2' is the stream in which all parameters are tuned. The results are shown in Table 4, where EMA represents the Exponential Moving Average, and WSPA is Weighted Soft Parameter Alignment. The results demonstrate that the proposed modules are helpful for performance gains. Subsequently, we focus on validating the proposed supervision signals module. It can be seen that the pseudo-label after prior calibration can effectively improve the performance of the model. In addition, the proposed label selection strategy can also effectively suppress noisy labels. Finally, the proposed contrastive learning strategy effectively optimizes the sample distribution.

4.6 Parameters Analysis

We explored how the model varies with the parameter λ, and the results are shown in Figure 2. The results demonstrate that our method is not sensitive to λ_1 and λ_2 at range $[0.01, 1]$.

5 Conclusion

This paper first proposes a dual-stream structure to capture the discriminative ability, maintain generalization, and prevent catastrophic forgetting. We propose continually capturing source knowledge using different strategies in each stream to calibrate the adapted model. Then, we adopt the source predictions to rough calculate the lower and upper bounds of individual class probability, which can calibrate the pseudo-labels and avoid a trivial solution. Moreover, we select an independent threshold for each class through global and local strategies to choose reliable pseudo-labels. Based on such reliable parts, we construct a soft-weighted contrastive learning module, which pulls the potential same-class samples closer and discriminates against uncorrelated samples. Finally, we evaluate the proposed method on several benchmarks and prove its superiority.

Acknowledgements

This work is supported in part by the National Key Research and Development Program of China (No. 2023YFC3305600), Joint Fund of Ministry of Education of China (8091B022149, 8091B02072404), National Natural Science Foundation of China (62132016, 62171343, and 62201436), Key Research and Development Program of Shaanxi (2024GX-YBXM-127) and Fundamental Research Funds for the Central Universities (ZDRC2102).

References

[Bai *et al.*, 2021] Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. *NeurIPS*, 34:24392–24403, 2021.

[Cao *et al.*, 2022] Zhangjie Cao, Kaichao You, Ziyang Zhang, Jianmin Wang, and Mingsheng Long. From big to small: adaptive learning to partial-set domains. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(2):1766–1780, 2022.

[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 *CVPR*, pages 3941–3950, 2020.

[Ding *et al.*, 2023] Yifei Ding, Minping Jia, Jichao Zhuang, Yudong Cao, Xiaoli Zhao, and Chi-Guhn Lee. Deep imbalanced domain adaptation for transfer learning fault diagnosis of bearings under multiple working conditions. *Reliability Engineering & System Safety*, 230:108890, 2023.

[Döbler *et al.*, 2023] Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *CVPR*, pages 7704–7714, 2023.

[Frankle *et al.*, 2020] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *ICML*, pages 3259–3269. PMLR, 2020.

[Gan *et al.*, 2022] Yulu Gan, Xianzheng Ma, Yihang Lou, Yan Bai, Renrui Zhang, Nian Shi, and Lin Luo. Decorate the newcomers: Visual domain prompt for continual test time adaptation. *arXiv preprint arXiv:2212.04145*, 2022.

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 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. *Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[Gong *et al.*,] Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. Note: Robust continual test-time adaptation against temporal correlation. In *NeurIPS*.

[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.

[Jiang *et al.*, 2021] Junguang Jiang, Baixu Chen, Jianmin Wang, and Mingsheng Long. Decoupled adaptation for cross-domain object detection. *arXiv preprint arXiv:2110.02578*, 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 *ICMLW*, 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 *CVPR*, 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 *ICML*, pages 6028–6039. PMLR, 2020.

[Liu *et al.*, 2021] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *CVPR*, pages 1215–1224, 2021.

[Marsden *et al.*, 2023] Robert A Marsden, Mario Döbler, and Bin Yang. Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction. *arXiv preprint arXiv:2306.00650*, 2023.

[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 *ICCV*, 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 *ICCV*, 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. *NeurIPS*, 33:11539–11551, 2020.

[Sun *et al.*, 2022] Tao Sun, Cheng Lu, and Haibin Ling. Prior knowledge guided unsupervised domain adaptation. In *ECCV*, pages 639–655. Springer, 2022.

[Tzeng *et al.*, 2017] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 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.

[Wang *et al.*, 2023a] Mengzhu Wang, Shanshan Wang, Wei Wang, Li Shen, Xiang Zhang, Long Lan, and Zhigang Luo. Reducing bi-level feature redundancy for unsupervised domain adaptation. *Pattern Recognit.*, page 109319, 2023.

[Wang *et al.*, 2023b] Xu Wang, Dezhong Peng, Peng Hu, Yunhong Gong, and Yong Chen. Cross-domain alignment for zero-shot sketch-based image retrieval. *IEEE Trans. Circuits Syst. Video Technol.*, 2023.

[Wang *et al.*, 2023c] Xu Wang, Dezhong Peng, Ming Yan, and Peng Hu. Correspondence-free domain alignment for unsupervised cross-domain image retrieval. *arXiv preprint arXiv:2302.06081*, 2023.

[Wolsey, 2020] Laurence A Wolsey. *Integer programming*. John Wiley & Sons, 2020.

[Xie *et al.*, 2017] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500, 2017.

[Xie *et al.*, 2022] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, Xinjing Cheng, and Guoren Wang. Active learning for domain adaptation: An energy-based approach. In *AAAI*, volume 36, pages 8708–8716, 2022.

[Yang *et al.*, 2021] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *CVPR*, 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 Trans. Pattern Anal. Mach. Intell.*, 44(4):1992–2003, 2022.

[Yang *et al.*, 2023] Xu Yang, Yanan Gu, Kun Wei, and Cheng Deng. Exploring safety supervision for continual test-time domain adaptation. In *IJCAI*, pages 1649–1657, 2023.

[Yuan *et al.*, 2023] Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *CVPR*, pages 15922–15932, 2023.

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