Revisiting Entropy Minimization for *Long-Sequence* Continual Test-Time Adaptation

Wei Qin Chuah
wei.qin.chuah@rmit.edu.au
Ruwan Tennakoon
ruwan.tennakoon@rmit.edu.au
Ali Bab-Hadiashar
ali.bab-hadiashar@rmit.edu.au

RMIT University
128 La Trobe Street,
Melbourne, VIC, Australia

Abstract

Deep neural networks (DNNs) have revolutionized a wide range of tasks, yet they remain vulnerable to distributional shifts between training and test data, which is a common occurrence in real-world scenarios. These shifts often unfold continually over time, posing significant challenges for test-time adaptation. A critical issue we identify in this setting is **noisy label memorization**, where the model gradually overfits to its own erroneous predictions (pseudo-labels), leading to significant performance degradation. This problem is particularly severe in **long-sequence** continual test-time adaptation, where the model must adapt continuously in non-stationary environments. To address this, we propose a novel methodology that strategically constrains weight updates to prevent noisy labels from degrading the model, resulting in significant performance improvements, particularly in long-sequence continual test-time adaptation (CoTTA) scenarios. Extensive experiments on benchmarks such as CIFAR10-C, CIFAR100-C, ImageNet-C, CCC, and DomainNet demonstrate that our method not only outperforms state-of-the-art CoTTA techniques but also ensures sustained model reliability and robustness over time. Furthermore, our method exhibits greater stability across varying sample sizes and batch sizes, with reduced sensitivity to model selection—all achieved without the need for source data calibration. For details and code, please visit our GitHub repository.

1 Introduction

Deep neural networks (DNNs) have achieved remarkable success across many tasks. However, this success relies on the assumption that test data are drawn from the same distribution as training data. In real-world scenarios, this assumption is often violated due to out-of-distribution (OOD) test data caused by changes in the data or environment, leading to significant performance degradation.

To address this challenge, test-time adaptation (TTA) has emerged as a promising solution by enabling models to adapt to shifting data distributions during deployment. Its

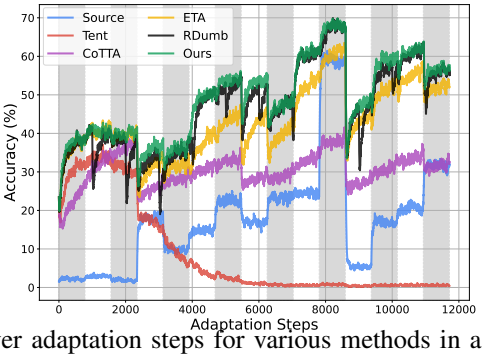

Figure 1: Accuracy over adaptation steps for various methods in a long-sequence continual test-time adaptation task. The Tent model deteriorates over time, ultimately performing worse than the source model, which does not apply any adaptation. The ETA method shows strong performance initially but declines in later stages, revealing its limitations. In contrast, our method consistently outperforms others, maintaining high accuracy throughout the sequence and effectively mitigating the negative effects of noisy label memorization.

continual variant, continual test-time adaptation (CoTTA) [21], extends this idea to handle multiple and unpredictable shifts. Yet, in long-sequence continual test-time adaptation (LoCoTTA) settings, existing methods can struggle. For example, as shown in Figure 1, methods like TENT [20] quickly degrade, and even dedicated CoTTA approaches [21] offer only marginal improvements over a non-adapted source model. RDumb [14] may provide strong initial performance, but its reliance on fixed reset intervals makes it hard to optimize without prior knowledge of future data size.

In light of these challenges, we revisit entropy minimization (EM)[1] from the perspective of learning with noisy labels [9, 17]. We focus on the early learning phenomenon (ELP) [9], where a model initially learns intrinsic patterns but eventually starts memorizing label noise, leading to performance degradation.

Our analysis in Section 3 shows that ELP, previously observed in noisy label settings, also occurs during test-time adaptation with EM. In Section 3, we demonstrate that the extent of noisy label memorization correlates with changes in the model parameters, measured via cosine similarity between their initial and adapted states. Critically, parameters important for adapting to new environments show larger updates early on than those sensitive to noise.

Building on these insights, we propose a method that dynamically constrains parameter changes at each layer based on cosine similarity between the current and initial parameters. This strategy preserves essential representations while ensuring effective adaptation, thereby reducing overfitting to incorrect labels. Unlike EATA [12], our approach does not require a separate validation set for calibration, and it exhibits strong stability across different batch sizes and model architectures, including ResNet and Vision Transformer. Our method achieves state-of-the-art performance on extensive benchmarks, including challenging long-sequence environments, demonstrating its effectiveness and robustness.

[1]EM is computationally efficient as it updates only the normalization layers, making it well suited for TTA. Note that our experiments reveal that student-teacher methods underperform on complex datasets like ImageNet-C in long-sequence scenarios.

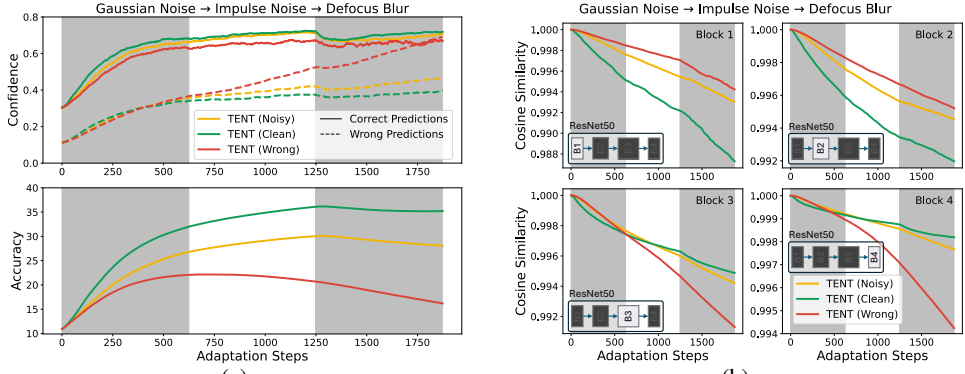

Figure 2: Analysis of noisy label memorization during LoCoTTA. (a) Confidence and accuracy plots. (b) Cosine similarity in different ResNet-50 blocks. Transitions between different corruption types are marked by the alternating grey and white background.

2 Related Workds

Test-time Adaptation (TTA) refers to the process of adapting a pre-trained model to a target domain distribution using only the test data encountered during inference, without relying on any source domain data and without access to labelled data. Recent works can be grouped into two categories: self-training [4, 10, 18] and entropy regularization [12, 13, 20]. For instance, TENT [20] updates the training parameters in batch normalization layers through entropy minimization. This method has inspired subsequent research efforts [12, 13, 14], which further investigate the robustness of normalization layers.

Continual Test-Time Adaptation (CoTTA) extends the concept of TTA to scenarios where the target domain is dynamically changing, presenting additional challenges for conventional TTA methods. The initial approach to CoTTA was introduced in [21], where a teacher-student framework was employed, integrating bi-average pseudo labels and stochastic weight reset. This approach leverages the insight that mean teacher predictions are often more robust than those of standard models [19]. Subsequent methods [3, 8] have continued to explore the teacher-student framework and self-supervised learning within the CoTTA context. However, the dynamic nature of continually changing environments poses significant challenges, particularly concerning the accumulation of noisy labels and the stability of model adaptation over time.

In this work, we demonstrate that existing approaches, including teacher-student frameworks and entropy minimization techniques, are vulnerable to noisy label memorization, leading to significant performance degradation in long-sequence CoTTA settings. To address this, we propose a novel framework using dynamic weight regularization based on cosine similarity, specifically designed to mitigate noisy label memorization and maintain stable performance in dynamically changing environments.

3 Problem Statement and Motivation

Consider a classification model $f_\theta : \mathcal{X} \to \mathcal{Y}$ pre-trained on domain $\mathcal{D}$, which maps input $\mathbf{x} \in \mathcal{X}$ to a logit $\mathbf{z} = f_\theta(\mathbf{x}) \in \mathbb{R}^K$, where K is the number of classes. The logit is then transformed into a probability distribution using the softmax function. Domain $\mathcal{D}$ is defined

as the joint distribution $P_{\mathcal{D}}(X,Y)$ on the space $\mathcal{X} \times \mathcal{Y}$. When deployed in new a domain $\bar{\mathcal{D}}$, such that $P_{\mathcal{D}}(X) \neq P_{\bar{\mathcal{D}}}(X)$, the model faces the challenge of maintaining performance.

In LoCoTTA, the model encounters a long sequence of, T, new domains $\bar{\mathcal{D}}_1, \bar{\mathcal{D}}_2, \ldots, \bar{\mathcal{D}}_T$, where $P_{\mathcal{D}}(X) \neq P_{\bar{\mathcal{D}}_i}(X) \neq P_{\bar{\mathcal{D}}_j}(X), \forall i \neq j$. Assuming a covariate shift across domains, the model must continuously adapt during inference without access to the original training data or labels from the incoming data stream.

Noisy Labels Memorization in Continual Test Time Adaptation. Noisy label memorization is a well-known challenge in the field of learning with noisy labels [9, 11, 17]. Specifically, the work by Liu *et al.* [9] revealed that models tend to be resilient to noisy labels, focusing on learning the intrinsic patterns in the data during the early stages of training. However, as training continues, models begin to memorize the noisy labels, which negatively impacts performance and reduces the ability to generalize to new data. This memorization effect is detrimental, particularly in scenarios requiring continual adaptation, such as long-sequence continual test-time adaptation (LoCoTTA)[2].

Recent work [7] has shown that memorization by deep neural networks is closely linked to prediction confidence $\text{Conf}(x) = \max_i p(y_i \mid x)$, which is defined as the maximum softmax probability that the model assigns to a predicted class, where $p(y_i|x)$ represents the predicted softmax probability for class i given the input x. When a model's prediction confidence for an instance exceeds a certain threshold, it is more likely to memorize that instance, even if the prediction is incorrect. Thus, the likelihood and strength of memorization increases with the prediction's confidence, posing a significant challenge in handling noisy pseudo-labels.

We conducted an experiment to investigate the impact of noisy label memorization in long-sequence continual learning settings. Our setup involved adapting a ResNet-50 model from ImageNet to ImageNet-C using entropy minimization (i.e., TENT). The model was sequentially exposed to three types of corruption shifts: Gaussian noise, impulse noise, and defocus blur, with 50,000 samples per corruption type. We deliberately selected two noise-based corruptions followed by a blur corruption to examine how noisy label memorization evolves as the model transitions from a similar to a significantly different domain. In this analysis, we used the class labels of the data points (which are typically unavailable in real-world scenarios) to devise three adaptation strategies: adapting the model using exclusively 1) correct data points, 2) wrong data points (determined by comparing the model's predictions to the actual labels), and 3) a mixture of both, representing a more realistic scenario.

Figure 2a presents a comparative analysis of the results, which shows that the model trained with correct data points maintains stable confidence for both correct and wrong predictions across all corruptions, effectively avoiding noisy label memorization. In contrast, the model adapted with *wrong* data points exhibits a gradual increase in confidence for *wrong* predictions, signifying the onset of noisy label memorization, which correlates with declining accuracy over time. The model trained with a mix of correct and incorrect data points displays intermediate behaviour, maintaining high confidence in correct predictions while gradually increasing confidence in incorrect ones, ultimately leading to a drop in accuracy.

Impact of Noisy Label Memorization on Weight Update Dynamics. To uncover how noisy label memorization impacts a model's ability to adapt and generalize, particularly in LoCoTTA, we investigated the weight update dynamics[3]. In our analysis, we use cosine

[2]The conventional CoTTA benchmark on ImageNet-C uses 5,000 samples, resulting in just 78 adaptation steps per domain, which appears to be too short for the model to experience the effects of noisy label memorization. Consequently, this benchmark setting fails to fully demonstrate the challenges posed by memorization.

[3]Weight update dynamics can be assessed by examining changes in both the angle and magnitude of the weight

similarities between the initial and adapted weights to measure these dynamics. Given that entropy minimization methods typically optimize only the trainable parameters of batch normalization layers, our analysis focuses specifically on these parameters. Let $\mathbf{W}_l^0$ denote the initial ImageNet-pretrained weights of a layer l, and $\mathbf{W}_l$ denote the weights after adapting for t steps. The cosine similarity for that layer is then computed as:

$$\psi_l = \cos(\mathbf{W}_l, \mathbf{W}_l^0) = \frac{\mathbf{W}_l \cdot \mathbf{W}_l^0}{\|\mathbf{W}_l\| \|\mathbf{W}_l^0\|}. \tag{1}$$

Higher cosine similarity suggests that the weights remain closer to their initial values, indicating less significant updates. Figure 2b illustrates the cosine similarities for the ResNet model. For clarity in visualization, we group the layers into four blocks, $\{B_1, \ldots, B_4\}$. As each block comprises multiple layers, the cosine similarity for a block, b is computed as $\Psi_b = \frac{1}{|B_b|} \sum_{l \in B_b} \psi_l$.

In our analysis, we observe distinct patterns in parameter updates across network layers based on label quality during test-time adaptation. For blocks closer to the network input (e.g., blocks 1 and 2), models adapted with correct labels display a noticeable decrease in cosine similarity, indicating substantial weight updates. This adaptive behavior reflects the model's ability to adjust early layers, which are crucial for fundamental feature extraction and promoting robust adaptation. In contrast, models adapted with incorrect or noisy labels demonstrate more conservative updates in these critical early layers, suggesting limited capacity to generalize to new data.

The blocks closer to the network output (e.g., blocks 3 and 4) that are adapted with incorrect labels show a sharp decline in cosine similarity, indicating aggressive parameter updates in these deeper layers to fit and memorize patterns associated with noisy labels. This aggressive adaptation reflects the network's struggle to handle conflicting signals, as it overfits to noise rather than learning generalizable features. Conversely, when adapted with correct labels, these blocks maintain more stable cosine similarity, facilitating balanced adaptation without overfitting.

In a realistic scenario, where the model is adapted with a mix of correct and incorrect labels, the model initially shows conservative updates in the deeper layers. However, as adaptation progresses, these updates deviate from the adaptation pattern seen with correct labels alone and leading to a gradual decline in accuracy. This shift indicates that while noisy labels may support some initial adaptation, the accumulation of noisy signals ultimately destabilizes the model in long-sequence adaptation scenarios.

Overall, these results highlight the impact of noisy label memorization on weight update dynamics in continual test-time adaptation (CoTTA). Correct labels facilitate stable and effective updates across layers crucial for feature extraction, which supports robust adaptation. In contrast, incorrect labels introduce erratic adjustments, especially in the layers redundant to robust adaptation, leading the model to memorize noise rather than adapt meaningfully.

These findings underscore the necessity of reliable label quality to maintain stable adaptation and sustained performance in CoTTA, as inconsistent label quality disrupts the model's capacity to generalize across diverse data distributions. While existing methods, such as sample selection [12] and pseudo-label correction [3, 21], can mitigate the presence of incorrect

vectors. Due to space constraints, our main manuscript focuses on the changes in angles; however, our analysis reveals that weight magnitude exhibits similar trends. We have included a detailed analysis of weight magnitude in the supp. document. Thus, we emphasize the importance of aligning both the direction and magnitude of weight updates to effectively mitigate the impact of noisy label memorization.

labels to some degree, yet they remain vulnerable to the detrimental effects of memorization, especially in longer sequences, leading to overfitting and reduced generalization. Addressing this issue is crucial for developing robust models capable of continual adaptation in dynamic, real-world environments.

4 Proposed Method: Strategic Memorization Regularization (SMR)

In the previous section, we identified noisy label memorization as a significant challenge in long-sequence continual test-time adaptation (LoCoTTA), contributing to performance loss over time. Furthermore, we observed that while the model initially leverages correct signals, it progressively deteriorates as noisy labels accumulate, with specific layers undergoing substantial changes to accommodate this noise, ultimately undermining robustness.

Building on these insights, we propose a novel methodology that constrains weight updates to address noisy label memorization in LoCoTTA. By integrating weight similarity into the commonly used entropy minimization (EM) strategy, this approach enhances the model's robustness throughout extended adaptation sequences. Below, we provide a brief overview of EM, followed by an introduction to our proposed method. The EM method minimizes the entropy loss to encourage the model to produce confident predictions, thereby reducing uncertainty. For a given input batch, $\mathcal{B}$, the entropy loss is computed as:

$$\mathcal{L}_{\text{entropy}} = -\frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \sum_{i=1}^{C} p(y_i \mid x) \log p(y_i \mid x), \tag{2}$$

where C is the number of classes, and $p(y_i|x)$ is the predicted probability for class i. In the context of continual test-time adaptation, EM is applied iteratively as the model encounters new, unseen data, allowing it to adapt dynamically to changes in data distribution.

Building on the foundation of EM, our method strategically constrains weight updates at each layer based on their similarity to the initial state. Specifically, throughout the adaptation process, we impose restrictions on weights that remain close to their initial values during the early adaptation steps, preventing them from undergoing significant changes.

This constraint is supported by our observations discussed in Section 3, which are consistent with findings in ELP [9, 17]; that is, during early adaptation steps, the model is naturally more resistant to fitting noisy labels, as it is focused on adjusting parameters essential for adapting to new environments. Constraining certain weights near their initial values leverages this early-stage robustness, allowing the model to prioritize stable adaptation. Weights less likely to memorize noise are updated more freely, promoting flexible but controlled adaptation.

This regularization, which we termed Strategic Memorization Regularization (SMR), is incorporated into the overall loss function as an additional term that penalizes significant deviations from the initial weights, weighted by their cosine similarity, ψ_l ,which was defined in Equation (1):

$$\mathcal{L}_{\text{smr}} = \sum_{l=1}^{L} \psi_l^2 \left\| \mathbf{W}_l - \mathbf{W}_l^0 \right\|_2^2. \tag{3}$$

Here, L is the total number of layer in the neural network. While our proposed method shares some similarities with the Fisher regularizer introduced in EATA [12], it differs fundamentally in approach, as we discuss in the supplementary document due to space constraints.

The combined loss function can be expressed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{entropy}} + \lambda \mathcal{L}_{\text{smr}}. \tag{4}$$

Here, the hyperparameter λ controls the strength of this regularization.

By integrating this regularization into the adaptation process, our method mitigates the risk of noisy label memorization, ensuring the model focuses on accurate signals and avoids overfitting. This approach enhances model robustness and promotes stable performance across extended adaptation in dynamic environments.

5 Experimental Results and Discussion

In the following sections, we first provide detailed descriptions of the datasets and experimental setup. We then compare our proposed method against state-of-the-art approaches across a variety of benchmarks and architectures. Following that, we present a comprehensive discussion of our findings. Due to space constraints, the hyperparameter sensitivity analysis is included in the supplementary material.

5.1 Datasets

We evaluate our approach on the CIFAR10-C, CIFAR100-C, ImageNet-C, DomainNet-126 [15] datasets, and Continually Changing Corruptions benchmark (CCC) [14]. The CIFAR-C and ImageNet-C are standard benchmarks for assessing the robustness of classification networks [8] and are widely used in test-time adaptation evaluations. These datasets contain 15 types of corruption, each applied at 5 levels of severity. Our evaluation considers two continual adaptation settings: *abrupt* changes and *gradual* changes. In the *abrupt* setting, the model adapts to a sequence of test domains, each presenting the highest severity level (level 5) of all 15 corruptions, simulating sudden and dramatic environmental changes. In the *gradual* setting, corruption severity increases and then decreases incrementally through the sequence, simulating more gradual environmental shifts. To challenge the model further on ImageNet-C, we extended the abrupt and gradual benchmarks by increasing the number of samples from 5,000 per domain to 50,000, thereby using the entire dataset for long-sequence adaptation.

We also evaluated our model on the extremely challenging Continually Changing Corruptions benchmark [14], which spans 7.5 million images—ten times larger than the ImageNet-C benchmark (750k images).

In addition to corruption shifts, we evaluated our method on style shifts using the DomainNet126 dataset [15], which has 126 classes across four domains: real, clipart, painting, and sketch. This dataset alters visual styles rather than introducing traditional corruptions, allowing us to thoroughly assess our method's robustness across diverse domain shifts. However, due to space constraints, results for this benchmark are included in the appendix.

5.2 Implementation details

For the CIFAR10-C and CIFAR100-C benchmarks, we used the pre-trained WideResNet-28 [23] and ResNeXt-29 [22], respectively. Both pre-trained models are sourced from the RobustBench framework [2]. For ImageNet-C and DomainNet benchmarks, we utilized a pre-trained ResNet-50 from the Torchvision library and included ViT for ImageNet-C. We

Method	Source	BN-1	ETA	TENT	CoTTA	EATA	RDumb	Ours
Avg. Error	82.7	72.1	98.9	98.6	92.3	64.6	61.1	**57.3**

Table 1: Classification error rates (%) on the **Continually Changing Corruptions benchmark**, averaged across all corruption transition speeds for the medium difficulty level [4]. Results for methods other than ours are obtained from the official benchmark.

Dataset	Network	Levels	Source	TENT	ETA	Rdumb	EATA	Ours
CIFAR100-C	ResNext29	level 1-5	33.6	86.7	32.1	32.1	31.7	**26.3**
		level 5	46.4	87.9	34.5	34.5	34.0	**30.5**
ImN-C (5K)	ReNet50	level 1-5	58.4	46.1	44.0	39.8	38.9	**38.3**
		level 5	82.0	58.7	51.4	50.7	49.5	**48.9**
	ViT_B_16	level 1-5	39.0	36.6	33.9	32.1	32.7	**32.0**
		level 5	60.2	54.4	46.6	45.0	45.4	**44.3**
ImN-C (50K)	ReNet50	level 1-5	58.5	97.8	42.4	38.9	38.2	**37.9**
		level 5	82.0	98.1	50.9	50.6	49.4	**49.3**
	ViT_B_16	level 1-5	38.9	97.8	37.0	30.9	32.1	**30.7**
		level 5	60.2	99.0	50.1	42.9	43.3	**42.2**

(a)

Method	Category	CF10-C	CF100-C	ImN-C	
				5K	50K
Source	-	43.5	46.4	82.0	82.0
BN-1 [14]	N	20.4	35.4	68.6	68.5
AdaContrast [3]	PL	18.5	33.5	65.5	92.4
CoTTA [34]	PL	**16.5**	32.8	63.1	69.7
RMT [3]	PL	*17.0*	**30.2**	59.9	58.2
TENT [33]	EM	20.0	62.2	62.6	91.3
DeYO [1]	EM	18.7	33.8	60.0	91.3
ETA [24]	EM	17.9	32.2	60.2	57.1
Rdumb [24]	EM	17.8	32.2	60.0	58.9
SAR [24]	EM	20.4	32.0	61.9	60.2
EATA [24]	EM	17.9	32.2	*58.0*	*52.0*
Ours	EM	**16.5**	*30.3*	**56.9**	**51.0**

(b)

Table 2: Performance comparison with the state-of-the-art methods for CoTTA under (a) "**gradual changes**" and (b) "**abrupt changes**", with classification error rates (%) reported. (a) We report the average error over all corruption severity levels (level 1–5) as well as the error at the highest severity level (level 5 only), averaged across 15 corruption types. (b) Methods are grouped by: N (normalization), PL (pseudo-label refinement), and EM (entropy minimization). *italicized* indicates second-best results, **bolded** *italicized* indicates best.

set learning rates at 0.001 for CIFAR datasets and 0.00025 for both ImageNet and DomainNet, with batch sizes of 200 for CIFAR, 64 for ImageNet, and 128 for DomainNet. Optimization was done using ADAM for CIFAR datasets and SGD for ImageNet-C and DomainNet. Additionally, we incorporated the sample selection procedure from Niu et al. [12] to further enhance our method.

5.3 Experimental Results on Continual Test-time Adaptation

Continually Changing Corruptions. In Table 1, we show the classification error rates on the extremely challenging Continually Changing Corruptions benchmark [14]. Despite the extended adaptation process, our method achieves the lowest error rate (57.3%), significantly outperforming other approaches, including RDumb (61.1%) and EATA (64.6%). In contrast, entropy minimization-based methods like TENT and ETA exhibit severe performance degradation, with error rates of 98.6% and 98.9%, respectively. These results underscore our method's robustness in maintaining strong performance and effectively mitigating noisy label memorization, even under such prolonged and challenging conditions.

Gradual Image Corruption. Table 2a showcases the classification error rates for various methods on the CIFAR100-C and ImageNet-C benchmarks under gradual adaptation, where corruption severity changes incrementally. Our proposed method consistently outperforms state-of-the-art approaches in these settings.

On CIFAR100-C, our method achieves an average error rate of 26.3% across all severity levels (1-5), significantly outperforming TENT (86.7%) and ETA (32.1%). Even at the highest severity level (level 5), our method remains robust, with an error rate of 30.5%, surpassing EATA (34.0%) and other competitors.

[4] Results for the easy and difficult levels of the CCC dataset are not reported due to issues with data download accessibility.

In the more challenging ImageNet-C benchmark, our method excels, particularly as the number of adaptation samples increases. With 5,000 samples, it achieves an error rate of 38.3% across all levels, outperforming ETA (44.0%) and TENT (46.1%). When scaled to 50,000 samples, our method further improves, achieving 37.9% across all levels and 49.3% at level 5, demonstrating its scalability. Moreover, our method is not limited to ResNet architectures; it also shows significant improvements with Vision Transformers (ViT). When tested with the ViT_B_16 network on ImageNet-C, our method achieves 30.7% across all levels and 42.2% at the highest severity, outperforming all other methods. These results confirm our method's versatility and superiority in managing gradual corruption, providing a distinct advantage in real-world scenarios where conditions evolve incrementally.

Abrupt Image Corruption. Table 2b presents the error rates for various methods on the CIFAR10-C, CIFAR100-C, and ImageNet-C (5,000 and 50,000 samples) datasets under the highest corruption severity. Our method consistently achieves state-of-the-art performance among entropy minimization (EM) approaches across all benchmarks. On CIFAR10-C, it records an error rate of 16.5%, matching the leading pseudo-label refinement (PL) methods. Similarly, on CIFAR100-C, it achieves an error rate of 30.3%, outperforming all other EM-based methods and remaining highly competitive with PL methods like RMT [3].

Our approach particularly excels in the more challenging ImageNet-C benchmark, especially when the number of adaptation samples increases from 5,000 to 50,000. In this extended setting, our method reduces the error rate to 51.0%, significantly outperforming other EM-based methods, such as EATA [12] (52.0%) and TENT [20] (91.3%). This improvement underscores the robustness and scalability of our method, particularly in long-sequence adaptation scenarios with large data volumes. Additionally, while PL methods like CoTTA [21] and RMT excel on CIFAR datasets, our method outperforms them on the more challenging ImageNet-C benchmark, effectively mitigating noisy label memorization and offering a distinct advantage in diverse and challenging test-time adaptation scenarios.

5.4 Discussion

Ablation Study. The ablation study in Table 3 evaluates the effectiveness of our proposed regularization method ($\mathcal{L}_{\text{reg}}$) in combination with entropy minimization for continual test-time adaptation. We also explore the impact of integrating active sample selection from Niu *et al*. [12] and freezing layers in the final block, which we identified as sensitive to label noise memorization (see Section 3). Freezing the final block helps mitigate label noise memorization but is insufficient on its own, particularly in challenging benchmarks like ImageNet-C. This is likely because earlier layers may also be sensitive to noise. While sample selection reduces label noise, it does not eliminate it, leaving the model at risk of memorizing noisy labels. Our regularization method, however, consistently enhances performance across all datasets, especially on the more challenging ImageNet-C with 50,000 samples. When combined with sample selection and block freezing, our approach achieves the lowest error rates, highlighting the critical role of regularizing weight updates and strategically stabilizing the model during adaptation.

Impact of Sample Size on Continual Adaptation Performance. As the number of adaptation steps increases with the sample size, the extended adaptation process enters later phases where the model becomes more prone to label noise memorization [9]. To this end, we assess how effectively each method mitigates label noise under extended adaptation scenarios. Figure 3 compares the error rates of various methods as the number of adaptation samples increase in the ImageNet-C dataset. The results show that without regularization techniques,

Cosine Weighting	$\mathcal{L}_{reg}$	Freeze	Sample Selection	CF100-C	ImN (5K)	ImN (50K)
✗	✗	✗	✗	62.2	62.6	94.5
✗	✗	✓	✗	31.1	62.4	70.9
✗	✓	✗	✗	32.9	62.4	59.3
✓	✓	✗	✗	32.1	61.9	57.2
✓	✓	✓	✗	30.3	61.7	56.6
✗	✗	✗	✓	32.2	60.2	57.1
✓	✓	✓	✓	30.3	56.9	51.0

Table 3: Results of the ablation study. We also examine the effects of incorporating active sample selection [12] and freezing the final block, which has shown sensitivity to label noise memorization.

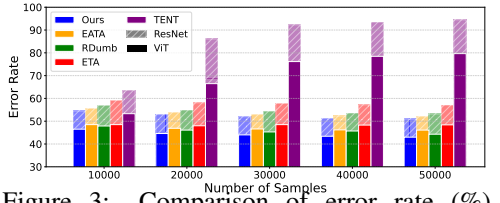

Figure 3: Comparison of error rate (%) with different numbers of samples included for continual adaptation, using ImageNet-C dataset on ResNet50 and vision transformer (ViT) models.

the performance of the TENT model deteriorates significantly as the number of samples increases, likely due to severe label noise memorization. While sample selection in ETA helps alleviate this issue and improves performance, it falls short because it cannot entirely eliminate incorrect pseudo-labels, leading to increasing error rates as more samples are added. Interval resetting and the Fisher regularizer perform better than sample selection, but our proposed method consistently achieves the best overall performance across both ResNet and ViT models, effectively mitigating label noise even under prolonged adaptation conditions.

In addition to sample size, we also examine the impact of reducing batch size, which increases the number of adaptation steps. Due to space constraints, detailed results are provided in the appendix. Briefly, most methods exhibit rising error rates as batch size decreases, highlighting increased vulnerability to label noise. By contrast, our proposed method consistently maintains robust performance even with batch sizes as small as 8, demonstrating notable resilience to label noise under extended adaptation conditions.

6 Conclusion

In this work, we have highlighted the critical challenge of noisy label memorization in long-sequence continual test-time adaptation (CoTTA), demonstrating how it leads to significant performance degradation over time. Drawing on insights from the Early Learning Phenomenon (ELP) observed in CoTTA, we proposed a novel methodology that strategically constrains weight updates based on cosine similarity between current and initial weights, effectively mitigating the impact of noisy label memorization. Our approach not only outperforms existing state-of-the-art methods across standard benchmarks but also exhibits greater stability across varying sample sizes, batch sizes, and models without requiring source data for calibration. These results underscore the effectiveness and robustness of our method in maintaining model reliability in dynamic, non-stationary environments, making it a promising solution for real-world CoTTA challenges.

References

[1] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022.

[2] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench:

a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://openreview.net/forum?id=SSKZPJCt7B.

[3] Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7704–7714, 2023.

[4] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-time training with masked autoencoders. *Advances in Neural Information Processing Systems*, 35:29374–29385, 2022.

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

[6] Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=9w3iw8wDuE.

[7] Yifan Li, Hu Han, Shiguang Shan, and Xilin Chen. Disc: Learning from noisy labels via dynamic instance-specific selection and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24070–24079, 2023.

[8] Jiaming Liu, Senqiao Yang, Peidong Jia, Renrui Zhang, Ming Lu, Yandong Guo, Wei Xue, and Shanghang Zhang. ViDA: Homeostatic visual domain adapter for continual test time adaptation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=sJ88Wg5Bp5.

[9] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

[10] Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 34:21808–21820, 2021.

[11] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

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

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

[14] Ori Press, Steffen Schneider, Matthias Kuemmerer, and Matthias Bethge. RDumb: A simple approach that questions our progress in continual test-time adaptation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=VfP6VTVsHc.

[15] 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.

[16] 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.

[17] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11):8135–8153, 2022.

[18] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020.

[19] 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.

[20] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

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

[22] 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.

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