Test-Time Model Adaptation for Quantized Neural Networks

Zeshuai Deng*
South China University of Technology
Guangzhou, China
sedengzeshuai@mail.scut.edu.cn

Guohao Chen*
Nanyang Technological University
Singapore, Singapore
guohao.chen@ntu.edu.sg

Shuaicheng Niu*
Nanyang Technological University
Singapore, Singapore
shuaicheng.niu@ntu.edu.sg

Hui Luo
Institute of Optics and Electronics,
Chinese Academy of Sciences
Chengdu, China
luohui19@mails.ucas.ac.cn

Shuhai Zhang
South China University of Technology
Guangzhou, China
shuhaizhangshz@gmail.com

Yifan Yang
South China University of Technology
Guangzhou, China
youngyif1@gmail.com

Renjie Chen
South China University of Technology
Guangzhou, China
202410190283@mail.scut.edu.cn

Wei Luo*
South China Agricultural University
Guangzhou, China
cswluo@scau.edu.cn

Mingkui Tan†
South China University of Technology
Guangzhou, China
mingkuitan@scut.edu.cn

Abstract

Quantizing deep models prior to deployment is a widely adopted technique to speed up inference for various real-time applications, such as autonomous driving. However, quantized models often suffer from severe performance degradation in dynamic environments with potential domain shifts and this degradation is significantly more pronounced compared with their full-precision counterparts, as shown by our theoretical and empirical illustrations. To address the domain shift problem, test-time adaptation (TTA) has emerged as an effective solution by enabling models to learn adaptively from test data. Unfortunately, existing TTA methods are often impractical for quantized models as they typically rely on gradient backpropagation—an operation that is unsupported on quantized models due to vanishing gradients, as well as memory and latency constraints. In this paper, we focus on TTA for quantized models to improve their robustness and generalization ability efficiently. We propose a continual zeroth-order adaptation (ZOA) framework that enables efficient model adaptation using only two forward passes, eliminating the computational burden of existing methods. Moreover, we propose a domain knowledge management scheme to store and reuse different domain knowledge with negligible memory consumption, reducing the interference of different domain knowledge and fostering the knowledge accumulation during long-term adaptation. Experimental results on three classical architectures, including quantized transformer-based and CNN-based models, demonstrate the superiority of our methods for quantized model adaptation. On the quantized W6A6 ViT-B model, our ZOA is able to achieve a 5.0% improvement over the state-of-the-art FOA on ImageNet-C dataset. The source code is available at https://github.com/DengZeshuai/ZOA.

*Equal contribution.
†Corresponding author.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754730

CCS Concepts

• **Computing methodologies → Online learning settings**.

Keywords

Test-Time Adaptation; Quantized Neural Network; Zeroth-Order Optimization; Domain Knowledge Management

ACM Reference Format:
Zeshuai Deng, Guohao Chen, Shuaicheng Niu, Hui Luo, Shuhai Zhang, Yifan Yang, Renjie Chen, Wei Luo, and Mingkui Tan. 2025. Test-Time Model Adaptation for Quantized Neural Networks. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3754730

1 Introduction

The common paradigm in deploying deep neural network models involves quantizing full-precision models into low-bit precision, *e.g.*, 8-bit precision, before deployment using training-aware quantization [13, 16, 25, 27] or post-training quantization [40, 54, 56, 59]. This workflow has become a standard practice in multimedia applications, aiming to accelerate inference and improve response speed across a wide range of scenarios, including those involving low-power edge devices or latency-sensitive contexts. However, after deployment, the environments may dynamically change with domain shifts, such as changes in lighting variations and sensor noise [42, 52, 53], and under such domain shifts, these models may suffer from severe performance degradation.

In this paper, we identify that quantized models are more sensitive to domain shifts compared with their full-precision counterparts, exhibiting significantly greater performance degradation when encountering such shifts. As illustrated in Fig. 1, the accuracy gap between in-distribution (ID) and out-of-distribution (OOD) data

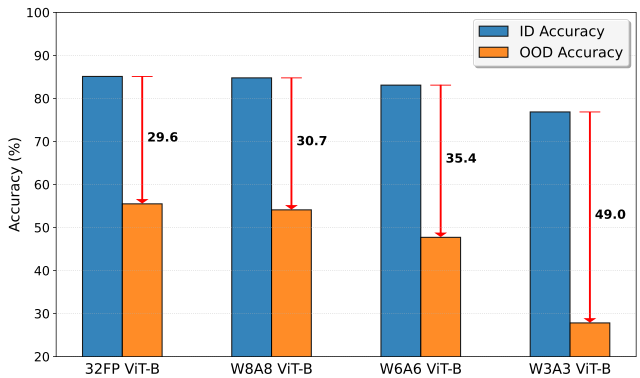

Figure 1: Quantized models often exhibit substantial performance degradation on out-of-distribution (OOD) data compared to their full-precision counterparts. The red arrows represent the accuracy gap between the results of models on in-distribution (ID) data (ImageNet) and OOD data (ImageNet-C). "32FP" denotes the 32-bit floating-point ViT-B model, and "WNAN" indicates the N-bit quantized ViT-B variant.

widens markedly as the bit-width of the quantized model decreases. This is because the quantization process inevitably introduces loss errors when facing out-of-distribution perturbations, and this loss increases as the bit-precision decreases, *see* our analyses in Sec. 3.1. However, existing quantization methods typically optimize the quantized models on source ID data, overlooking their performance on OOD ones [29, 57]. Therefore, there remains an urgent demand to efficiently adapt quantized models to new environments.

To address the domain shift issue, test-time adaptation (TTA) [26, 42, 43, 52, 63] has emerged as a promising paradigm by adapting deployed models to out-of-distribution data during test time. Based on whether they rely on backpropagation (BP), existing methods can be broadly categorized into two groups: (i) BP-based methods, such as entropy minimization [42, 52], prediction consistency maximization [63], and feature distribution alignment [38]; and (ii) BP-free methods like calibrating the statistics of batch normalization layers [39, 45] and correcting the predicted logits [23]. However, effectively supporting various quantized models remains challenging for these approaches, making them impractical for scenarios such as deployment on edge devices. The reasons are as follows.

BP-based TTA methods typically [42, 50, 52] focus on full-precision models and rely on BP for model adaptation, which requires the function of model inference to be continuous and differentiable. However, the forward process of quantized models exhibits discrete characteristics and lacks differentiability due to their extremely low-bitwidth parameter representations (*e.g.*, 4-bit or 8-bit precision). Besides, gradient computation through BP entails substantial memory overhead for retaining intermediate activation values. While the edge devices, such as FPGAs and smartphones, are often resource-limited, making it unaffordable to perform these computationally expensive BP-based methods. Therefore, it is non-trivial to directly use these methods for efficient quantized model adaptation.

BP-free TTA methods [2, 23, 41] adapt the deployed models using forward passes only. However, conventional methods like T3A [23] and LAME [2] are learning-free. Without updating the core parameters of models (such as parameters of normalization layers), these methods often show limited learning capabilities(see results in Tab. 2). Recently, the learning-based method FOA [41] has been proposed, demonstrating much better performance compared with learning-free approaches. However, it often requires many forward passes per test sample to achieve satisfactory results, which may be impractical for real-time applications. Moreover, FOA is specifically designed for transformer architectures, limiting its applicability to a broader range of model types. Thus, achieving real-time, resource-efficient adaptation for quantized models in long-term, dynamically changing environments remains an open and challenging problem.

In this paper, we propose a zeroth-order test-time adaptation algorithm (ZOA) that efficiently adapts quantized models using only two forward passes. Specifically, we design a continual domain knowledge learning method that accumulates knowledge across different domains. To reduce the interference between different domains, we design a domain knowledge management scheme to store and reuse the learned knowledge. With a set of learnable coefficients to aggregate the knowledge from different domains, our ZOA efficiently enhances the reliability of quantized models in long-term adaptation. We evaluate our ZOA and existing methods across three classical architectures and different quantization bit-precision. Experimental results demonstrate the effectiveness of our methods on quantized model adaptation. Overall, the contribution of our work is summarized as follows:

- We identify that the quantized models exhibited higher sensitivity to domain shift than their full-precision counterparts. To achieve reliable AI applications under dynamic conditions on edge devices, we design a continual zeroth-order adaptation framework, namely ZOA, to efficiently adapt the quantized models during testing using only two forward passes per sample.
- We introduce a zeroth-order continual domain knowledge learning scheme. This method reduces cross-domain knowledge interference and enables the accumulation and reuse of historical adaptation knowledge for more effective forward-only TTA. We further propose a domain knowledge management strategy to ensure the computational and memory efficiency of our ZOA.
- Experiments on three classical architectures across different bit-precisions demonstrate that our ZOA is able to efficiently adapt the quantized models by storing and reusing the learned domain knowledge to boost the long-term test-time adaptation.

2 Related Works

2.1 Test-Time Adaptation

Test-time Adaptation (TTA) aims to adapt a pre-trained model to unlabeled test data to handle the domain shift between training and testing data. According to whether the model adaptation relies on backpropagation, existing methods can be categorized into: 1) *Backpropagation-based methods*: Early TTA methods optimize a source model via backpropagation using an extra self-supervision task at testing, such as rotation prediction [49], contrastive learning [1, 32], reconstruction learning [11, 15], etc. However, self-supervised tasks are essentially proxy tasks, which usually alter the training process, and this solution is not always feasible in practice. To address this issue, fully test-time adaptation methods have been proposed, which update models using unlabeled test data through unsupervised learning objectives, including entropy minimization [42, 50, 52], energy alignment [8, 61], and prediction consistency maximization [4, 14, 53, 63], etc. Nonetheless, the aforementioned methods suffer from computational and memory inefficiency and limited application scenarios, *e.g.*, can not be applied to quantized models, due to their reliance on backpropagation. 2) *Backpropagation-free methods*: A common approach for BP-free TTA is to calibrate the statistics of batch normalization (BN) using the mean and variance computed over the test data [37, 45, 52]. Nevertheless, this method requires multiple test data to calculate statistics and assumes a balanced class distribution within a batch. To address this, later studies adopt data augmentations [63] for single-sample BN calibration or a class-wise sample bank [17] for the imbalanced data stream. However, without optimizing the core parameters of models, these methods often show limited online learning capabilities. Recently, a forward-optimization adaptation (FOA) [41] has been proposed to optimize quantized models without requiring backpropagation. However, FOA requires a long adaptation time—up to 28 forward passes per sample, which is impractical for real-time applications. Instead, our ZOA seeks to use only two forward passes per sample to adapt quantized models continually, boosting TTA's practicality in various resource-limited contexts.

2.2 Zeroth-Order Optimization

Zeroth-order Optimization (ZO) leverages forward passes to estimate gradients without backpropagation. Recent advances in ZO have significantly expanded its applications in machine learning [30] and natural language processing [36], particularly in scenarios where gradient information is unavailable or impractical to obtain [10]. Motivated by the need for black-box adaptation in NLP, methods like BBT [48] and BBTv2 [47] have employed evolutionary strategies such as CMA-ES [19] to optimize proprietary models without access to gradients, while RLPrompt [10] leverages reinforcement learning for prompt tuning. However, these approaches often face challenges with high variance and instability [30], especially in vision tasks. To address these limitations, recent work has focused on improving gradient estimation accuracy through techniques like two-sided approximations [44] and reducing variance in ZO fine-tuning of large language models by sparse parameter perturbations [33], and increased batch sizes [24] and tensorized adapters [60]. Additionally, efforts to scale up ZO optimization have included integrating historical data [6], and reusing intermediate features [3] to enhance efficiency and convergence rates. These advancements highlight the growing potential of ZO methods in handling complex, large-scale machine learning problems while maintaining computational efficiency and adaptability. In this paper, we incorporate the advantage of zeroth-order optimization to update the core parameters of the quantized neural networks for test-time adaptation.

3 Motivation and Problem Statement

In this section, we first theoretically demonstrate that QNNs are highly vulnerable to distribution shifts, and then revisit the limitations of existing TTA solutions in adapting quantized models.

3.1 Sensitivity of QNN with Distribution Shift

We theoretically analyze the impact of out-of-distribution (OOD) perturbations on QNNs and provide justification for the necessity of TTA strategies in such scenarios, especially under low-bit precision.

Proposition 1. *Considering linear models, let n denote the bit-precision of quantized models, for out-of-distribution (OOD) input perturbations δ, the quantization-induced loss difference $\Delta\mathcal{L} := \hat{\mathcal{L}}(\mathbf{x}+\delta) - \mathcal{L}(\mathbf{x}+\delta)$ satisfies:*

$$\Delta\mathcal{L} > 0, \text{ and } \Delta\mathcal{L} \propto \frac{1}{2^{2n}}, \tag{1}$$

where $\hat{\mathcal{L}}$ and $\mathcal{L}$ denote the MSE losses of quantized and full-precision models, respectively.

The proof of Proposition 1 is provided in Appendix A. This result highlights two key insights: 1) Quantization always increases the loss under OOD perturbations, *i.e.*, $\Delta\mathcal{L} > 0$; and 2) This sensitivity grows exponentially as the bit-precision n decreases, *i.e.*, $\Delta\mathcal{L} \propto \frac{1}{2^{2n}}$. Our empirical observations in Fig. 1 are also aligned with our theoretical analysis. Overall, these findings underscore a critical limitation of QNNs and motivate the need for an effective TTA solution to enhance their robustness against distribution shifts.

3.2 Test-Time Adaptation

Let $f_{\boldsymbol{\theta}}(\cdot)$ be the model trained on the source training dataset $D_s = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N_s}$ and $\mathbf{x}_i \sim P(\mathbf{x})$. During testing, test samples can be drawn from a shifted and dynamically changing distribution $Q(\mathbf{x})$, where $Q(\mathbf{x}) \neq P(\mathbf{x})$, causing performance degradation in $f_{\boldsymbol{\theta}}(\cdot)$ predictions. To adapt the pre-trained model $f_{\boldsymbol{\theta}}(\cdot)$ to the target domain, conventional TTA methods [42, 52] seek to optimize the model using some un/self-supervised learning objective on test samples:

$$\min_{\tilde{\boldsymbol{\theta}}} \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}),\ \mathbf{x} \sim Q(\mathbf{x}), \tag{2}$$

where $\tilde{\boldsymbol{\theta}} \subseteq \boldsymbol{\theta}$ denotes the learnable parameters involved in TTA, and the test-time objective $\mathcal{L}(\cdot)$ is often formulated as entropy minimization [42, 52, 53], activation alignment [38, 41], and *etc.*

Nevertheless, existing TTA methods typically require gradient-based updates through backpropagation, posing a fundamental limitation when applied to quantized models that do not support backpropagation. Recently, FOA [41] circumvents backpropagation by employing an evolutionary strategy to optimize input prompts. However, it 1) incurs substantial computational cost—requiring up to 28 forward passes per sample—and 2) assumes a low-dimensional adaptation space, making it less flexible for adapting deep quantized models with high-dimensional parameters. Motivated by adaptation efficiency and applicability on various QNNs, in this paper, we seek to propose a novel TTA approach that effectively adapts a quantized model using only *two forward passes* per sample—one forward pass for standard inference and a single extra forward pass for TTA—without assuming a low-dimensional adaptation space.

4 Continual Zeroth-Order Adaptation for QNNs

In this paper, we propose a novel continual Zeroth-Order Adaptation (ZOA) framework, which is designed to efficiently adapt a quantized model with only two forward passes per sample to boost

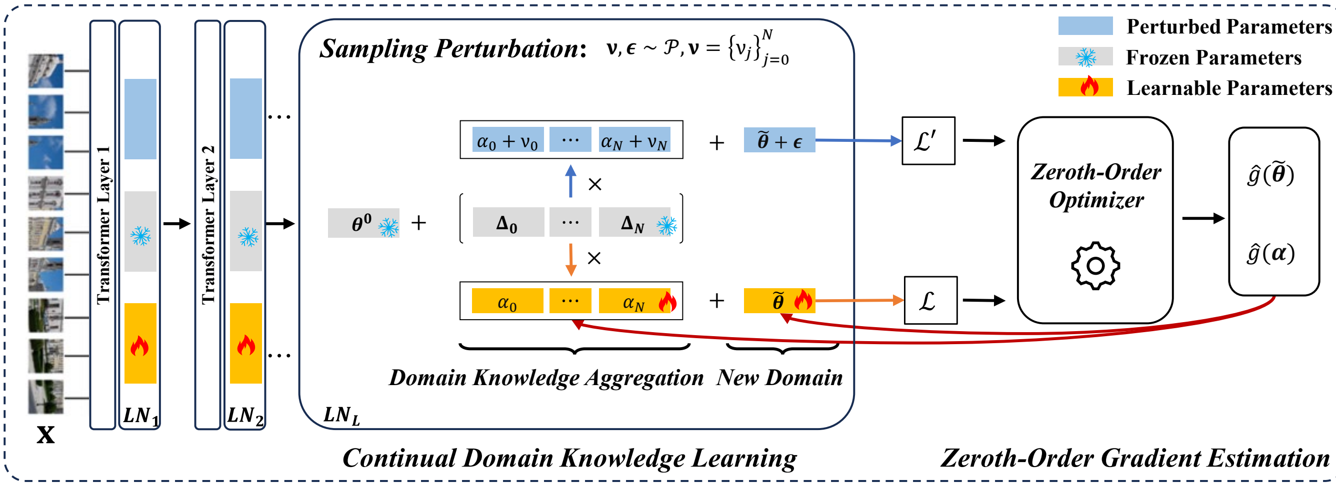

Figure 2: The overall framework of ZOA. We perturb the learnable parameters and the aggregation coefficients using the random perturbation vectors ϵ and ν, which are sampled from a predefined distribution $\mathcal{P}$. The gradient is estimated based on the loss $\mathcal{L}'$ and $\mathcal{L}$, where $\mathcal{L}' = \mathcal{L}(\mathbf{x}; \tilde{\theta} + c\epsilon, \alpha + c\nu)$ is the loss related to perturbed parameters and $\mathcal{L} = \mathcal{L}(\mathbf{x}; \tilde{\theta}, \alpha)$ is related to learnable parameters. The yellow blocks represent the learnable parameters, including α and $\tilde{\theta}$. The blue blocks represent the parameters perturbed by random vectors ν and ϵ. The gray blocks represent the frozen parameters that would not be updated.

Algorithm 1: The pipeline of our ZOA.

Input: Test samples $D_t = \{\mathbf{x}_i\}$, source model $f_{\theta^0}(\cdot)$, distribution $\mathcal{P}$, learning rate η_a and η_p.

Initialize $\mathcal{T} = \{\mathbf{0}\}$, $\tilde{\theta} = \mathbf{0}$ and $\alpha = \mathbf{0}$.
for $\mathbf{x} \in D_t$ **do**
  // *First forward pass.*
  Forward to obtain $\hat{\mathbf{y}}_i$ and compute $\mathcal{L}(\mathbf{x}; \tilde{\theta}, \alpha)$.
  Sample perturbations ν and ϵ from distribution $\mathcal{P}$.
  // *Second forward pass.*
  Forward to compute $\mathcal{L}(\mathbf{x}; \tilde{\theta} + c\epsilon, \alpha + c\nu)$.
  Estimate gradients $\hat{g}(\tilde{\theta})$ and $\hat{g}(\alpha)$ based on Eq. (5).
  Update $\tilde{\theta}$ using $\hat{g}(\tilde{\theta})$ and η_p.
  Update α using $\hat{g}(\alpha)$ and η_a.
  if *Domain changes* **then**
    Store current parameters Δ_t into $\mathcal{T}$ using Eq. (7).
    Add new learnable parameters $\tilde{\theta}'$ using Eq. (8).
  end
  if $|\mathcal{T}| > N$ **then**
    Remove redundant parameters via Eq. (9) and (10).
  end
end

Output: The predictions $\{\hat{\mathbf{y}}_i\}$.

TTA practicality. As shown in Fig. 2, ZOA comprises two key designs: 1) A continual domain knowledge learning scheme based on zeroth-order optimization, which learns and reuses domain knowledge without forgetting for more effective TTA using two forward passes (c.f. Sec. 4.1); and 2) A domain knowledge management strategy that preserves diverse, and informative domain knowledge to reduce memory overhead while maintaining TTA performance (c.f. Sec. 4.2). We summarize the pseudo-code of ZOA in Algorithm 1.

4.1 Continual Domain Knowledge Learning

Unlike conventional TTA methods that update the model via backpropagation, we aim to achieve both computationally and memory-efficient TTA—requiring only two forward passes per test sample. To this end, one direct solution is using the Covariance Matrix Adaptation Evolution Strategy [18] (CMA-ES), as in FOA [41]. However, this is inefficient and ineffective in our setting, due to: 1) CMA-ES relies solely on the rank of candidate solutions but ignores the detailed fitness (target loss) value, failing to provide sufficient learning signals when only two forward passes per sample are available, as in Fig. A of Appendix; 2) CMA-ES assumes a low-dimensional optimization space (*e.g.*, prompt learning on ViT [41]), restricting its efficacy on quantized models with high-dimensional parameters.

In this paper, we seek to exploit the exact loss values of candidate solutions to deliver richer learning signals for more efficient and effective forward-only adaptation, and thus introduce a zeroth-order adaptation method. Additionally, we devise a continual learning scheme that accumulates learned knowledge without forgetting, thereby further boosting the efficiency and effectiveness of subsequent forward-only adaptations under a very limited number (*i.e.*, 2) of forward passes. We depict them in the following.

Zeroth-Order Gradient Estimation. To efficiently update the QNN with two forward passes, we explore the zeroth-order optimization SPSA [46] in TTA, which leverages the exact loss values $\mathcal{L}(\cdot)$ of solutions for gradient estimation. Specifically, the one-sided averaged gradient estimator of SPSA [31, 51] can be formulated as:

$$\hat{g}(\theta) = \hat{\nabla}\mathcal{L}(\mathbf{x}; \theta) = \frac{1}{q}\sum_{i=1}^{q} \frac{\mathcal{L}(\mathbf{x}; \theta + c\epsilon_i) - \mathcal{L}(\mathbf{x}; \theta)}{c}\epsilon_i^{-1}, \quad (3)$$

where q denotes the step of perturbations, $c > 0$ is the perturbation scale, and $\boldsymbol{\epsilon}$ is a random perturbation vector sampled from mean-zero distributions, such as Rademacher and Segmented Uniform distribution [44]. For our ZOA, we set the number of steps q to be 1. Therefore, we only require two forward passes to estimate the gradients, including one forward pass for standard inference and computing the loss of current parameters $\mathcal{L}(\mathbf{x};\boldsymbol{\theta})$, and one forward pass for computing the loss of perturbed parameters $\mathcal{L}(\mathbf{x};\boldsymbol{\theta}+c\boldsymbol{\epsilon})$.

Continual Knowledge Reprogramming Learning. Although exploiting exact loss values in Eq. (3) improves learning efficiency, using only two forward passes provides only a coarse gradient estimation, which may still limit the learning efficacy for TTA. On the other hand, increasing the number of forward passes to deliver more informative gradients would compromise the efficiency we aim to preserve. Alternatively, we explore the potential of continual learning with zeroth-order optimization, which seeks to accumulate and leverage the historical adaptation knowledge to boost the efficacy of subsequent TTA without expensive computational cost.

To achieve knowledge accumulation, we first decouple the knowledge of different domains into domain knowledge vectors [58], *i.e.*, $\Delta_j = \boldsymbol{\theta}_j - \boldsymbol{\theta}^0$, where $\boldsymbol{\theta}_j$ is adapted parameters on the j-th domain, and $\boldsymbol{\theta}^0$ is the source pre-trained parameters. We then save Δ_j in knowledge base $\mathcal{T} = \{\Delta_j\}_{j=0}^N$ upon distribution shifts (as discussed in Sec. 4.2). Given the knowledge base $\mathcal{T}$, our continual knowledge reprogramming learning process is formulated as:

$$\min_{\boldsymbol{\alpha},\tilde{\boldsymbol{\theta}}} \mathcal{L}(\mathbf{x};\tilde{\boldsymbol{\theta}}),\ \text{where}\ \boldsymbol{\theta} = \boldsymbol{\theta}^0 + \sum_{j=0}^{N} \alpha_j \Delta_j + \tilde{\boldsymbol{\theta}},\ \Delta_j \in \mathcal{T}. \tag{4}$$

Here, $\tilde{\boldsymbol{\theta}}$ denotes a set of new learnable parameters to enable learning new domain knowledge. $\boldsymbol{\alpha}$ is the learnable aggregation weights for different domain knowledge, normalized with the softmax function, *i.e.*, $\sum_{j=0}^N \alpha_j = 1$. The relevant domain knowledge is retrieved to enhance adaptation with a large coefficient α_j, while the irrelevant domain knowledge is suppressed with a small coefficient.

Within our zeroth-order optimization framework, we only update the parameters $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\alpha}$, while freezing other parameters. Notably, we perturb $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\alpha}$ simultaneously at each step to avoid introducing an additional forward pass for Eq. (3). The estimated gradients of domain knowledge parameters $\tilde{\boldsymbol{\theta}}$ and aggragation coefficients $\boldsymbol{\alpha}$ are computed by:

$$\begin{bmatrix} \hat{g}(\tilde{\boldsymbol{\theta}}) \\ \hat{g}(\boldsymbol{\alpha}) \end{bmatrix} = \frac{\mathcal{L}(\mathbf{x};\tilde{\boldsymbol{\theta}}+c\boldsymbol{\epsilon},\boldsymbol{\alpha}+c\boldsymbol{\nu}) - \mathcal{L}(\mathbf{x};\tilde{\boldsymbol{\theta}},\boldsymbol{\alpha})}{c} \begin{bmatrix} \boldsymbol{\epsilon}^{-1} \\ \boldsymbol{\nu}^{-1} \end{bmatrix}, \tag{5}$$

where $\boldsymbol{\epsilon}$ is the random perturbation vectors for $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\nu}$ is the random perturbation vectors for $\boldsymbol{\alpha}$. With the estimated gradients of learnable parameters, we can update the parameters using gradient descent methods, such as SGD or AdamW [35].

Objective Function for Quantized Model Adaptation. Our proposed continual zeroth-order optimization framework does not limit the objective function used to optimize $\boldsymbol{\alpha}$ and $\tilde{\boldsymbol{\theta}}$. In our implementation, we select FOA [41] as our test-time learning objective, which provides stable learning signals for forward optimization with a feature discrepancy loss and an entropy minimization loss. Specifically, we first calculate the mean and standard deviations of the output activations of L intermediate blocks $\{\mu_i^s, \sigma_i^s\}_{i=0}^L$ over a small set of unlabeled ID data. During testing, we calculate the statistics $\{\mu_i(\mathbf{x}_t), \sigma_i(\mathbf{x}_t)\}_{i=0}^L$ over the current batch of test samples $\mathbf{x}_t$. The objective function for B test samples $\mathbf{x}_t$ is then given by:

$$\begin{aligned} \mathcal{L}(\mathbf{x}_t;\boldsymbol{\theta}^t) &= \frac{1}{B\times C}\sum_{x\in\mathbf{x}_t}\sum_{c\in C} -y_c \log y_c \\ &+ \frac{\lambda}{L}\sum_{i=1}^{L} ||\mu_i(\mathbf{x}_t) - \mu_i^s||_2 + ||\sigma_i(\mathbf{x}_t) - \sigma_i^s||_2, \end{aligned} \tag{6}$$

where $\boldsymbol{\theta}^t$ denotes the adapted parameters at t-th step. y_c is the c-th element of the prediction $\hat{\mathbf{y}}_t = f_{\boldsymbol{\theta}^t}(\mathbf{x}_t)$, C is the dimension of class space C, and λ is a trade-off parameter. Note that a small number of unlabeled ID samples (*e.g.*, 32 samples) is sufficient to compute the feature statistics of ID data [41], making it practical for application.

4.2 Domain Knowledge Management

In this section, we introduce our domain knowledge management strategy that preserves valuable learned knowledge during the continual TTA while maintaining a manageable memory footprint.

Domain Knowledge Preservation. In the long-term TTA, we use domain shift detection [5, 22] to identify whether the test distribution changes. We compute the distribution distance of current test samples and historical samples, the domain change occurs when the distance is larger than a predefined threshold (refer to Appendix B for details). Once changes occur, we accumulate the learned domain vectors in $\mathcal{T}$ and initialize a new set of parameters $\tilde{\boldsymbol{\theta}}$ in Eq. (4) for further domain adaptation. Specifically, the knowledge preservation process is formulated as:

$$\mathcal{T} = \mathcal{T} \cup \{\Delta_t\},\ \text{where}\ \Delta_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}^0. \tag{7}$$

Here, $\boldsymbol{\theta}_t$ is the ensemble parameters at the t-th domain from Eq. (4). Note that the sum of all coefficients equals to 1 due to the softmax function, *i.e.*, $\sum_{j=0}^n \alpha_j + \alpha_t = 1$, thus adding new domain vectors Δ_t in $\mathcal{T}$ can change the value of the ensemble parameter in Eq. (4).

To keep TTA stability, we seek to keep the ensemble parameters $\boldsymbol{\theta}_t$ unchanged after storing new domain vectors Δ_t in $\mathcal{T}$. To this end, let $\boldsymbol{\theta}'$ be the updated ensemble parameters after adding new parameters in $\mathcal{T}$, we re-initialize the $\tilde{\boldsymbol{\theta}}$ for new domain as:

$$\tilde{\boldsymbol{\theta}}' = \tilde{\boldsymbol{\theta}} - (\boldsymbol{\theta}' - \boldsymbol{\theta}_t), \tag{8}$$

where $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}'$ denote the learnable parameters before and after vector preservation, respectively. We also carefully initialize the coefficient α_t of the domain parameters Δ_t to keep the parameters $\tilde{\boldsymbol{\theta}}'$ initialized as small as possible, thereby reducing cross-domain interference. We put more details of $\boldsymbol{\alpha}$ initialization in Appendix C.

Redundant Domain Knowledge Removal. Over time, the number of stored domain vectors in $\mathcal{T}$ may grow excessively large, resulting in an unbearable memory footprint that hinders TTA practicality. To prevent this, we aim to preserve only the most diverse and informative ones to reduce the memory overhead.

To keep the memory footprint manageable, we restrict $\mathcal{T}$ to a maximum size N and remove redundant domain vectors (*i.e.*, those that are outdated and show a large similarity with other vectors) when the number of vectors in $\mathcal{T}$ exceeds N. Formally, the similarity

Table 1: Effectiveness of our ZOA on Quantized ViT-B models. We report the Accuracy (%) on ImageNet-C (severity level 5) during the 10th round of continual adaptation. "WNAN" indicates that the ViT-B model is quantized to N-bit precision.

		Noise			Blur				Weather				Digital				Average
Model	Method	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pix.	JPEG	Acc.
ViT-B W8A8	Source	56.1	56.0	56.7	46.7	34.9	52.5	42.6	61.1	61.7	65.9	77.1	24.3	44.5	65.6	66.8	54.2
	T3A [23]	52.9	53.1	53.5	45.0	33.7	50.2	40.5	58.2	60.5	62.6	74.6	42.3	39.2	63.8	64.8	53.0
	FOA [41] (K=2)	57.1	57.5	58.0	49.0	37.8	54.1	45.6	63.8	64.7	70.5	77.5	54.0	48.6	66.4	67.8	58.2
	ZOA (ours)	60.2	62.0	61.8	54.6	50.2	59.1	55.2	67.0	65.5	68.0	79.1	58.0	59.6	70.7	70.5	**62.8**
ViT-B W6A6	Source	44.2	42.0	44.8	39.8	28.9	43.4	34.7	53.2	59.8	59.0	75.1	27.4	39.0	59.1	65.3	47.7
	T3A [23]	37.1	36.6	37.9	28.0	25.1	36.6	27.7	45.2	54.6	52.8	69.8	19.7	31.5	56.4	61.5	41.4
	FOA [41] (K=2)	46.4	45.4	47.0	42.5	33.1	46.4	39.0	57.3	63.0	66.1	75.7	35.9	44.2	60.3	66.2	51.3
	ZOA (ours)	52.6	54.2	55.0	48.4	42.8	53.7	46.7	60.5	63.8	64.8	76.5	42.0	50.0	65.8	68.2	**56.3**
ViT-B W4A4	Source	46.0	45.5	46.2	41.4	28.8	44.3	37.3	50.7	57.5	62.0	74.5	26.0	39.6	56.9	63.8	48.0
	T3A [23]	41.3	41.6	42.1	39.1	27.9	41.5	34.6	47.1	55.8	60.1	71.7	30.0	36.1	53.8	61.5	45.6
	FOA [41] (K=2)	47.5	46.8	47.6	43.4	31.2	46.8	40.4	54.8	59.6	64.7	74.8	39.5	43.4	57.7	64.8	50.9
	ZOA (ours)	49.8	50.7	52.6	48.6	38.2	52.8	47.6	55.5	56.1	59.7	75.3	38.1	47.5	62.6	65.4	**53.4**

Table 2: Effectiveness of our ZOA on Quantized ViT-B models in long-term continual adaptation. We report the average Accuracy (%) on ImageNet-C (severity level 5) at each round of adaptation. "WNAN" indicates that the ViT-B model is quantized to N-bit precision. The bold number indicates the best result. "#FP" is the number of forward passes to obtain output and update models.

Models	Methods	#FP	1	2	3	4	5	6	7	8	9	10	Average
ViT-B W8A8	Source	1	54.2	54.2	54.2	54.2	54.2	54.2	54.2	54.2	54.2	54.2	54.2
	T3A [23]	1	55.6	55.6	55.1	54.9	54.6	54.3	53.9	53.7	53.4	53.0	54.4
	FOA [41] (K=2)	2	58.0	57.9	58.1	58.1	58.2	58.1	58.2	58.2	58.1	58.2	58.1
	ZOA (ours)	2	59.7	61.0	61.4	61.5	62.0	61.9	62.1	62.1	62.1	62.8	**61.7**
ViT-B W6A6	Source	1	47.7	47.7	47.7	47.7	47.7	47.7	47.7	47.7	47.7	47.7	47.7
	T3A [23]	1	43.3	44.7	44.3	43.9	43.5	43.0	42.6	42.2	41.9	41.4	43.1
	FOA [41] (K=2)	2	51.1	51.6	51.5	51.3	51.3	51.1	51.0	51.1	51.1	51.3	51.3
	ZOA (ours)	2	54.3	55.2	55.8	55.7	55.8	55.7	56.2	56.4	56.5	56.3	**55.8**
ViT-B W4A4	Source	1	48.0	48.0	48.0	48.0	48.0	48.0	48.0	48.0	48.0	48.0	48.0
	T3A [23]	1	47.3	47.6	47.3	47.0	46.6	46.4	46.2	46.0	45.9	45.6	46.6
	FOA [41] (K=2)	2	50.7	50.8	50.9	50.9	50.9	50.8	50.8	50.9	50.9	50.9	50.8
	ZOA (ours)	2	52.1	52.8	53.2	53.5	53.3	53.4	53.9	53.3	53.6	53.4	**53.2**

between domain knowledge vectors is computed as:

$$k, p = \operatorname*{arg\,max}_{\Delta_i, \Delta_j \in \mathcal{L}} \cos(\Delta_i, \Delta_j),$$
$$\cos(\Delta_i, \Delta_j) = \frac{1}{L} \sum_{l=0}^{L} \frac{\langle \Delta_{i,l}, \Delta_{j,l} \rangle}{\|\Delta_{i,l}\| \|\Delta_{j,l}\|}, \tag{9}$$

where k and p are the indices of the selected domain parameters, $\Delta_{j,l}$ is the parameters in the l-th layer of Δ_j, and $\langle \cdot, \cdot \rangle$ represents the inner product. In implementation, we pre-compute the similarities between all $\Delta_j \in \mathcal{T}$ to efficiently obtain the index k and p. Based on k and p, we remove the most outdated domain vectors from $\mathcal{T}$:

$$\mathcal{T} = \mathcal{T} \setminus \{\Delta_o\}, \tag{10}$$

where $\setminus$ denotes the remove operation and $o = \min(k, p)$.

5 Experiment

Datasets and Models. We conduct experiments on the benchmark of test-time adaptation, such as ImageNet-C [21]. ImageNet-C contains corrupted images in 15 types and each type has 5 severity levels. In the experiments, we mainly report the results on ImageNet-C at severity level 5. For simplicity, we abbreviate the 15 types of ImageNet-C as Gauss., Shot, Impul., Defoc., Glass, Motion, Zoom, Snow, Frost, Fog, Brit., Contr., Elas., Pix., and JPEG, respectively. We evaluate the different methods using ViT-B [12], ViM-S [64], and ResNet-50 [20]. These models are trained on the ImageNet-1K [9] training set. We obtain the model weights from the *timm* repository [55]. To obtain the quantized models, we adopt PTQ4ViT [62] to obtain the W8A8 and W6A6 ViT-B models. And we use AdaLog [56] to obtain the W4A4 and W3A3 ViT-B models. For the ViM-S model, we adopt PTQ4VM [7] to quantize the model into W8A8 precision. Moreover, we use MQBench [28] to obtain the quantized W8A8 and W2A4 ResNet-50 models. In the experiment, we use 'Source' to represent the deployed models without adaptation. We put more details in Appendix D.

Compared Methods. In the experiment, we compare our ZOA with there BP-free methods, including LAME [2], T3A [23], and FOA [41]. Since the edge devices are resource-limited, we mainly compare

Table 3: Effectiveness of our ZOA on W8A 8 ResNet-50 and ViM-S models. We report the Accuracy (%) on ImageNet-C (severity level 5) at the 10th round of adaptation. The bold number indicates the best result.

		Noise			Blur				Weather				Digital				Average
Model	Method	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pix.	JPEG	Acc.
ResNet-50 W8A8	Source	3.5	4.2	3.2	17.6	9.5	15.0	22.7	16.5	22.8	23.8	59.0	5.5	16.4	21.3	32.6	18.3
	BN Adapt	16.1	17.3	16.9	14.4	15.1	25.8	38.4	34.2	33.2	47.6	65.3	15.9	44.4	49.1	40.1	31.6
	T3A [23]	13.5	14.4	14.4	12.8	12.8	22.1	32.5	30.0	28.4	40.7	56.6	13.9	37.1	40.9	34.0	26.9
	FOA [41] (K=2)	15.9	17.4	15.3	12.8	13.3	20.2	30.2	25.7	27.4	37.0	56.9	12.0	36.2	43.2	37.8	26.7
	ZOA (ours)	22.9	23.8	23.4	18.8	21.1	29.5	40.5	35.9	35.3	47.2	64.4	20.7	48.1	50.2	43.8	**35.0**
ViM-S W8A8	Source	42.2	42.2	42.7	29.7	17.1	38.1	31.1	48.9	55.1	61.9	71.2	46.7	34.6	40.5	53.8	43.7
	T3A [23]	39.3	39.3	39.2	29.4	16.7	36.9	30.0	48.2	53.9	59.0	70.3	49.2	34.1	39.4	53.3	42.6
	FOA [41] (K=2)	42.8	42.0	44.1	27.6	15.2	37.1	32.2	49.3	55.7	60.5	70.4	51.6	35.6	40.1	53.7	43.9
	ZOA (ours)	46.5	47.9	48.7	31.4	23.8	43.5	36.2	50.7	55.0	63.5	71.6	49.6	37.0	43.4	55.6	**47.0**

Table 4: Effectiveness of our ZOA on Quantized ResNet-50 and ViM-S models in long-term continual adaptation. We report the average Accuracy (%) on ImageNet-C (severity level 5) at each round of adaptation. Both models are quantized into 8-bit precision. The bold number indicates the best result. "#FP" is the number of forward passes to obtain output and update models.

Models	Methods	#FP	1	2	3	4	5	6	7	8	9	10	Average
ResNet50 W8A8	Source	1	18.3	18.3	18.3	18.3	18.3	18.3	18.3	18.3	18.3	18.3	18.3
	BN Adapt	1	31.6	31.6	31.6	31.6	31.6	31.6	31.6	31.6	31.6	31.6	31.6
	T3A [23]	1	30.3	29.8	29.1	28.7	28.3	28.0	27.6	27.4	27.2	26.9	28.3
	FOA [41] (K=2)	2	24.9	26.2	26.2	26.4	26.5	26.7	26.7	26.8	26.8	26.7	26.4
	ZOA (ours)	2	32.8	33.7	34.1	34.1	34.4	34.6	34.3	34.4	34.6	35.0	**34.2**
ViM-S W8A8	Source	1	43.7	43.7	43.7	43.7	43.7	43.7	43.7	43.7	43.7	43.7	43.7
	T3A [23]	1	43.7	43.9	43.8	43.6	43.5	43.4	43.2	43.1	42.9	42.6	43.4
	FOA [41] (K=2)	2	43.9	43.9	43.9	43.9	43.9	43.9	43.9	43.9	43.9	43.9	43.9
	ZOA (ours)	2	46.1	46.4	46.5	46.8	46.8	47.0	46.9	46.9	46.8	47.0	**46.7**

FOA (K=2) in our experiments, which only uses two forward passes. We also compare our methods with BP-based methods on the full-precision ViT-B model to show the effectiveness of our method on resource-abundant devices. The compared BP-based methods include TENT [52], MEMO [63], SAR [43], CoTTA [53], EATA [42], and DeYO [26]. We report the adaptation results on the test dataset in terms of Top-1 Accuracy (abbreviated as Acc.).

Implementation Details. For fair comparison, we set the batch size of test samples to 64 as TENT [52] and SAR [43]. Following FOA [41], we use the validation set of ImageNet-1K to compute the statistics of ID data. During adaptation, we use the SGD optimizer to update the learnable domain parameters $\tilde{\theta}$, and use the AdamW [35] to update the aggregation coefficients $\boldsymbol{\alpha}$. For all quantized models, we sample the perturbation vectors $\boldsymbol{\epsilon}$ and $\boldsymbol{\nu}$ from a Rademacher and Segmented Uniform distribution [44]. For the W8A8 ViT-B model, we add the perturbation vectors $\boldsymbol{\epsilon}$ and $\boldsymbol{\nu}$ with the step size of 0.02 and 0.05, respectively. We set the learning rate of $\tilde{\theta}$ and $\boldsymbol{\alpha}$ to be 0.0005 and 0.01, respectively. We set the maximum number of domain knowledge parameters as $N = 32$ to avoid large memory consumption. We put more details in Appendix D.

5.1 Results on Quantized Neural Networks

Results on Quantized ViT-B models. We evaluate our ZOA on the quantized ViT-B models. Since the edge devices are resource-limited, we configure FOA to employ two forward passes as our ZOA. As shown in Tab. 2, our ZOA achieves the best performance in the first round of adaptation on ImageNet-C dataset. Moreover, during the long-term adaptation process, our ZOA is able to accumulate the learned domain knowledge and further boost the adaptation performance. We also show the results at the 10th round of adaptation in Tab. 1. As shown in Tab. 1, our methods achieve a 5.0% improvement over FOA on average on the quantized W6A6 model. These results on quantized models with different bit precisions demonstrate the superiority of our ZOAover existing methods.

Results on Quantized ResNet-50. We evaluate our ZOA on the classical convolutional neural network, *i.e.*, the quantized W8A8 ResNet-50 model in Tab. 3 and Tab. 4. From Tab. 3, FOA achieves limited results compared with the BN adapt baseline with only two forward passes. This limitation arises because FOA relies on low-dimensional prompt learning to adapt the model, which is less flexible on CNN-based architectures. Instead, our ZOA consistently adapts the quantized ResNet-50 to the target domains. As shown in Tab. 4, our ZOA benefits from our continual knowledge accumulation to achieve better performance in long-term adaptation.

Results on Quantized ViM-S. We also evaluate the different BP-free adaptation methods on the quantized Vision Mamba model, *i.e.*, the quantized W8A8 ViM-S model from PTQ4VM [7]. Since PTQ4VM [7] uses the per-token quantization scheme, it requires appropriate quantization parameters for activations related to the inserted prompts by FOA. However, obtaining suitable parameters

Table 5: Comparisons w.r.t. computational complexity. FP/BP is short for forward/backward propagation. Acc. (%) is the average accuracy on ImageNet-C (level 5) at the 10th round of continual adaptation. Memory (MB) is measured for processing 64 images on a single GPU. "*" denotes resetting the model parameters after adaptation on each corruption.

Model	Method	BP	#FP	#BP	Acc.	Memory (MB)
ViT-B	Source	✗	1	0	55.5	819
	MEMO [63]	✔	65	64	57.2	11,058
	TENT* [52]	✔	1	1	59.6	5,165
	TENT [52]	✔	1	1	0.1	5,165
	CoTTA* [53]	✔	3 or 35	1	61.7	16,836
	CoTTA [53]	✔	3 or 35	1	34.0	16,836
	LAME [2]	✗	1	0	54.1	819
ViT-B W8A8	Source	✗	1	0	54.2	205
	T3A [23]	✗	1	0	53.0	239
	FOA [41] (K=2)	✗	2	0	58.2	208
	ZOA(ours)	✗	2	0	**62.8**	**207**
ViT-B W4A4	Source	✗	1	0	47.7	102
	T3A [23]	✗	1	0	45.6	120
	FOA [41] (K=2)	✗	2	0	50.9	104
	ZOA(ours)	✗	2	0	**53.4**	**103**
ResNet-50 W8A8	Source	✗	1	0	18.3	195
	T3A [23]	✗	1	0	26.9	249
	FOA [41] (K=2)	✗	2	0	26.7	195
	ZOA(ours)	✗	2	0	**35.0**	**197**

Table 6: Ablation studies of each component on ImageNet-C (level 5). ZO denotes we use the zeroth-order optimizer for adaptation, and DRL denotes our continual domain reprogramming learning. We report the average accuracy at the first (R1) and the last round (R10) of continual adaptation.

Method	#FP	ZO	DRL	R1	R10	Memory (MB)
Source	1			54.2	54.2	205
FOA	2			58.0	58.2	208
V1	2	✔		59.7	60.2	206
ZOA (Ours)	2	✔	✔	**59.7**	**62.8**	207

Table 7: Effect of the number of domain parameters at the 10th round adaptation on ImageNet-C

N	0	2	4	8	16	32	64	128
Accuracy (%)	33.6	33.5	34.1	34.6	34.9	35.0	35.4	35.3
Memory (MB)	0.00	0.02	0.03	0.07	0.13	0.27	0.53	1.06

for these new prompts is challenging during testing. This mismatch in quantization parameters can degrade the performance of the pre-quantized model, further affecting the adaptation capability of FOA. As shown in Tab. 3, FOA achieves a limited performance gain compared with the source model. Instead, without requiring new prompt tokens, our ZOA does not suffer from the quantization scheme. Thus, our method is still able to adapt the quantized models and achieve the best results. From Tab. 3, our method achieves a 3.1% improvement over FOA on the quantized W8A8 ViM-S model. The results on three quantized models demonstrate the effectiveness of our method on different architectures.

Computational Complexity Analyses. In this part, we analyze the computational cost of different adaptation methods in Tab. 5. Following [34, 41], the memory consumption for quantized W8A8 and W4A4 models is an ideal estimation by 0.25× and 0.125× memory of 32-bit models, respectively. For the adaptation of the W8A8 ViT-B model, our ZOA achieves the best results with merely two forward passes, whose computational cost is only approximately twice that of the standard forward of the quantized model. The memory consumption of our ZOA is almost the same as that of the quantized model inference. On the quantized W8A8 ViT-B model, our ZOA even achieves a better performance than the BP-based methods on full-precision models, such as MEMO [63], TENT [52], and CoTTA [53]. Moreover, for the adaptation of the quantized ViT-B model with the W4A4 precision, our ZOA is able to adapt the model with merely 103 MB memory consumption. In conclusion, the cost of our ZOA is affordable for many resource-limited devices, such as smartphones, FPGAs, and laptops.

5.2 Further Experiments

In this part, we evaluate the component of our ZOA and provide more results and discussions in the following. Due to the page limit, we put more results and discussions in Appendix E.

Ablation Studies. In this part, we analyze the effectiveness of each component of our ZOA. We construct an additional baseline, namely V1, which does not use our continual domain knowledge learning scheme. To show the effectiveness of our method, we compare these methods with the state-of-the-art FOA in terms of accuracy at the 10th round of adaptation. As shown in Tab. 6, our baseline V1 using zero-order adaptation already achieves better results than FOA on ImageNet-C. Moreover, the results of the last row show that our continual domain knowledge learning greatly boost the long-term adaptation for quantized models with merely 2 MB extra memory consumption compared with the source model.

Effect of the number of domain parameters N. To investigate the hyperparameters N, we evaluate our ZOA on the W8A8 ResNet50 model. As shown in Tab. 7, when $N \geq 4$, our ZOA achieves promising improvements compared to the baseline without our continual knowledge learning scheme (*i.e.*, $N = 0$). In our experiments, we set N=32 for all quantized models by default.

6 Conclusion

In this paper, we theoretically and empirically illustrate the high sensitivity of quantized models to out-of-distribution data. To this end, we propose a continual zeroth-order adaptation framework (ZOA) for efficient quantized model adaptation with only two forward passes per test data. To further enhance the efficacy of forward-only TTA, we propose to accumulate and leverage the historical TTA knowledge with a continual knowledge reprogramming learning scheme, and devise a domain knowledge management strategy to reduce memory overhead. Experiments on quantized models across different architectures demonstrate the effectiveness of our framework for efficient quantized model adaptation.

Acknowledgments

This work was partially supported by the Science and Technology Innovation 2030--Major Projects (No. 2022ZD0208900), the National Natural Science Foundation of China (Grant No.U24A20327, U23B2013 and 62276176), Key-Area Research and Development Program of Guangdong Province (2018B010107001), TCL Science and Technology Innovation Fund, and the Young Scholar Project of Pazhou Lab (No.PZL2021KF0021).

References

[1] Alexander Bartler, Andre Bühler, Felix Wiewel, Mario Döbler, and Bin Yang. 2022. Mt3: Meta test-time training for self-supervised test-time adaption. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 3080–3090.

[2] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. 2022. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8344–8353.

[3] Aochuan Chen, Yimeng Zhang, Jinghan Jia, James Diffenderfer, Konstantinos Parasyris, Jiancheng Liu, Yihua Zhang, Zheng Zhang, Bhavya Kailkhura, and Sijia Liu. 2024. DeepZero: Scaling Up Zeroth-Order Optimization for Deep Model Training. In *The Twelfth International Conference on Learning Representations*.

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

[5] Guohao Chen, Shuaicheng Niu, Deyu Chen, Shuhai Zhang, Changsheng Li, Yuanqing Li, and Mingkui Tan. 2024. Cross-Device Collaborative Test-Time Adaptation. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 122917–122951.

[6] Shuyu Cheng, Guoqiang Wu, and Jun Zhu. 2021. On the Convergence of Prior-Guided Zeroth-Order Optimization Algorithms. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 14620–14631.

[7] Younghyun Cho, Changhun Lee, Seonggon Kim, and Eunhyeok Park. 2025. PTQ4VM: Post-Training Quantization for Visual Mamba. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1176–1185.

[8] Wonjeong Choi, Do-Yeon Kim, Jungwuk Park, Jungmoon Lee, Younghyun Park, Dong-Jun Han, and Jaekyun Moon. 2024. Adaptive Energy Alignment for Accelerating Test-Time Adaptation. In *The Thirteenth International Conference on Learning Representations*.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[10] Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing Discrete Text Prompts with Reinforcement Learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 3369–3391.

[11] Zeshuai Deng, Zhuokun Chen, Shuaicheng Niu, Thomas Li, Bohan Zhuang, and Mingkui Tan. 2023. Efficient test-time adaptation for super-resolution with second-order degradation and reconstruction. *Advances in Neural Information Processing Systems* 36 (2023), 74671–74701.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[13] Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. 2019. Learned Step Size Quantization. In *International Conference on Learning Representations*. 1–12.

[14] François Fleuret et al. 2021. Test time adaptation through perturbation robustness. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*.

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

[16] Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tianxiang Li, Peng Hu, Jiazhen Lin, Fengwei Yu, and Junjie Yan. 2019. Differentiable Soft Quantization: Bridging Full-Precision and Low-Bit Neural Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4852–4861.

[17] Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. 2022. Note: Robust continual test-time adaptation against temporal correlation. *Advances in Neural Information Processing Systems* 35 (2022), 27253–27266.

[18] Nikolaus Hansen. 2016. The CMA evolution strategy: A tutorial. *arXiv preprint arXiv:1604.00772* (2016).

[19] Nikolaus Hansen, Sibylle D Müller, and Petros Koumoutsakos. 2003. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evolutionary computation* 11, 1 (2003), 1–18.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[21] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*.

[22] Junyuan Hong, Lingjuan Lyu, Jiayu Zhou, and Michael Spranger. 2023. MECTA: Memory-Economic Continual Test-Time Model Adaptation. In *2023 International Conference on Learning Representations*.

[23] Yusuke Iwasawa and Yutaka Matsuo. 2021. Test-Time Classifier Adjustment Module for Model-Agnostic Domain Generalization. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 2427–2440.

[24] Shuoran Jiang, Qingcai Chen, Youcheng Pan, Yang Xiang, Yukang Lin, Xiangping Wu, Chuanyi Liu, and Xiaobao Song. 2024. ZO-AdaMU Optimizer: Adapting Perturbation by the Momentum and Uncertainty in Zeroth-Order Optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18363–18371.

[25] Sangil Jung, Changyong Son, Seohyung Lee, Jinwoo Son, Jae-Joon Han, Youngjun Kwak, Sung Ju Hwang, and Changkyu Choi. 2019. Learning to Quantize Deep Networks by Optimizing Quantization Intervals with Task Loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4350–4359.

[26] Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. 2024. Entropy is not Enough for Test-Time Adaptation: From the Perspective of Disentangled Factors. In *The Twelfth International Conference on Learning Representations*.

[27] Yuhang Li, Xin Dong, and Wei Wang. 2020. Additive Powers-of-Two Quantization: An Efficient Non-uniform Discretization for Neural Networks. In *International Conference on Learning Representations*. 1–15.

[28] Yuhang Li, Mingzhu Shen, Jian Ma, Yan Ren, Mingxin Zhao, Qi Zhang, Ruihao Gong, Fengwei Yu, and Junjie Yan. 2021. MQBench: Towards Reproducible and Deployable Model Quantization Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

[29] Ji Lin, Chuang Gan, and Song Han. 2019. Defensive Quantization: When Efficiency Meets Robustness. In *International Conference on Learning Representations*.

[30] Sijia Liu, Pin-Yu Chen, Bhavya Kailkhura, Gaoyuan Zhang, Alfred O Hero III, and Pramod K Varshney. 2020. A Primer on Zeroth-Order Optimization in Signal Processing and Machine Learning: Principals, Recent Advances, and Applications. *IEEE Signal Processing Magazine* 37, 5 (2020), 43–54.

[31] Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini. 2018. Zeroth-order stochastic variance reduction for nonconvex optimization. In *Advances in Neural Information Processing Systems*, Vol. 31.

[32] Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. 2021. TTT++: When Does Self-Supervised Test-Time Training Fail or Thrive? *Advances in Neural Information Processing Systems* 34 (2021), 21808–21820.

[33] Yong Liu, Zirui Zhu, Chaoyu Gong, Minhao Cheng, Cho-Jui Hsieh, and Yang You. 2024. Sparse mezo: Less parameters for better performance in zeroth-order llm fine-tuning. *arXiv preprint arXiv:2402.15751* (2024).

[34] Zhenhua Liu, Yunhe Wang, Kai Han, Wei Zhang, Siwei Ma, and Wen Gao. 2021. Post-Training Quantization for Vision Transformer. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 28092–28103.

[35] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

[36] Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. 2023. Fine-tuning language models with just forward passes. In *Advances in Neural Information Processing Systems*, Vol. 36. 53038–53075.

[37] Massimiliano Mancini, Hakan Karaoguz, Elisa Ricci, Patric Jensfelt, and Barbara Caputo. 2018. Kitting in the wild through online domain adaptation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1103–1109.

[38] Muhammad Jehanzeb Mirza, Pol Jané Soneira, Wei Lin, Mateusz Kozinski, Horst Possegger, and Horst Bischof. 2023. Actmad: Activation matching to align distributions for test-time-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24152–24161.

[39] Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. 2020. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963* (2020).

[40] Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. 2020. Up or down? adaptive rounding for post-training quantization. In *International conference on machine learning*. PMLR, 7197–7206.

[41] Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. 2024. Test-Time Model Adaptation with Only Forward Passes. In *International*

Conference on Machine Learning.

[42] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. 2022. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning*. 16888–16905.

[43] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. 2023. Towards Stable Test-time Adaptation in Dynamic Wild World. In *The International Conference on Learning Representations*.

[44] Changdae Oh, Hyeji Hwang, Hee-young Lee, YongTaek Lim, Geunyoung Jung, Jiyoung Jung, Hosik Choi, and Kyungwoo Song. 2023. Blackvip: Black-box visual prompting for robust transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24224–24235.

[45] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. 2020. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems* 33 (2020), 11539–11551.

[46] James C Spall. 1992. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control* 37, 3 (1992), 332–341.

[47] Tianxiang Sun, Zhengfu He, Hong Qian, Yunhua Zhou, Xuan-Jing Huang, and Xipeng Qiu. 2022. BBTv2: Towards a Gradient-Free Future with Large Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 3916–3930.

[48] Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022. Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning*. PMLR, 20841–20855.

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

[50] Mingkui Tan, Guohao Chen, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Peilin Zhao, and Shuaicheng Niu. 2025. Uncertainty-calibrated test-time model adaptation without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

[51] Yun-Yun Tsai, Pin-Yu Chen, and Tsung-Yi Ho. 2020. Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources. In *International Conference on Machine Learning*. PMLR, 9614–9624.

[52] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*.

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

[54] Xiuying Wei, Ruihao Gong, Yuhang Li, Xianglong Liu, and Fengwei Yu. 2022. QDrop: Randomly Dropping Quantization for Extremely Low-bit Post-Training Quantization. In *International Conference on Learning Representations*.

[55] Ross Wightman. 2019. PyTorch Image Models. https://github.com/rwightman/pytorch-image-models. doi:10.5281/zenodo.4414861

[56] Zhuguanyu Wu, Jiaxin Chen, Hanwen Zhong, Di Huang, and Yunhong Wang. 2024. AdaLog: Post-Training Quantization for Vision Transformers with Adaptive Logarithm Quantizer. In *European Conference on Computer Vision*.

[57] Yisong Xiao, Aishan Liu, Tianyuan Zhang, Haotong Qin, Jinyang Guo, and Xianglong Liu. 2023. RobustMQ: benchmarking robustness of quantized models. *Visual Intelligence* 1, 1 (2023), 30.

[58] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, Vol. 36.

[59] Jiwei Yang, Xu Shen, Jun Xing, Xinmei Tian, Houqiang Li, Bing Deng, Jianqiang Huang, and Xian-sheng Hua. 2019. Quantization networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7308–7316.

[60] Yifan Yang, Kai Zhen, Ershad Banijamali, Athanasios Mouchtaris, and Zheng Zhang. 2024. AdaZeta: Adaptive Zeroth-Order Tensor-Train Adaption for Memory-Efficient Large Language Models Fine-Tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 977–995.

[61] Yige Yuan, Bingbing Xu, Liang Hou, Fei Sun, Huawei Shen, and Xueqi Cheng. 2024. TEA: Test-time energy adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23901–23911.

[62] Zhihang Yuan, Chenhao Xue, Yiqi Chen, Qiang Wu, and Guangyu Sun. 2022. PTQ4ViT: Post-training Quantization for Vision Transformers with Twin Uniform Quantization. In *European conference on computer vision*. Springer, 191–207.

[63] Marvin Zhang, Sergey Levine, and Chelsea Finn. 2022. MEMO: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*, Vol. 35. 38629–38642.

[64] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. In *Proceedings of the International Conference on Machine Learning* (Vienna, Austria). JMLR.org, Article 2584, 14 pages.