

# AdaShadow: Responsive Test-time Model Adaptation in Non-stationary Mobile Environments

Cheng Fang<sup>†</sup>, Sicong Liu<sup>†</sup>, Zimu Zhou<sup>§</sup>, Bin Guo<sup>†,\*</sup>, Jiaqi Tang<sup>‡</sup>, Ke Ma<sup>†</sup>, Zhiwen Yu<sup>†,+</sup>

<sup>†</sup>Northwestern Polytechnical University

<sup>§</sup>City University of Hong Kong

<sup>‡</sup>The Hong Kong University of Science and Technology

<sup>+</sup>Harbin Engineering University

## ABSTRACT

On-device adapting to continual, unpredictable domain shifts is essential for mobile applications like autonomous driving and augmented reality to deliver seamless user experiences in evolving environments. Test-time adaptation (TTA) emerges as a promising solution by tuning model parameters with *unlabeled* live data immediately before prediction. However, TTA's unique forward-backward-reforward pipeline notably increases the latency over standard inference, undermining the *responsiveness* in time-sensitive mobile applications. This paper presents AdaShadow, a responsive test-time adaptation framework for *non-stationary mobile data distribution* and *resource dynamics* via selective updates of adaptation-critical layers. Although the tactic is recognized in *generic on-device training*, TTA's *unsupervised* and *online* context presents unique challenges in estimating layer importance and latency, as well as scheduling the optimal layer update plan. AdaShadow addresses these challenges with a *backpropagation-free assessor* to rapidly identify critical layers, a unit-based *runtime predictor* to account for resource dynamics in latency estimation, and an *online scheduler* for prompt layer update planning. Also, AdaShadow incorporates a memory I/O-aware computation reuse scheme to further reduce latency in the reforward pass. Results show that AdaShadow achieves the best accuracy-latency balance under continual shifts. At low memory and energy costs, Adashadow provides a 2× to 3.5× speedup (*ms*-level) over state-of-the-art TTA methods with comparable accuracy and a 14.8% to 25.4% accuracy boost over efficient supervised methods with similar latency.

## CCS CONCEPTS

• **Human-centered computing** → Ubiquitous and mobile computing; • **Computing methodologies** → Artificial intelligence.

## KEYWORDS

Latency-efficient test-time adaptation, mobile environments

\*Corresponding email: scliu@nwpu.edu.cn, guob@nwpu.edu.cn.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SenSys '24, November 4–7, 2024, Hangzhou, China

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0697-4/24/11

<https://doi.org/10.1145/3666025.3699339>

## ACM Reference Format:

Cheng Fang<sup>†</sup>, Sicong Liu<sup>†</sup>, Zimu Zhou<sup>§</sup>, Bin Guo<sup>†,\*</sup>, Jiaqi Tang<sup>‡</sup>, Ke Ma<sup>†</sup>, Zhiwen Yu<sup>†,+</sup>. 2024. AdaShadow: Responsive Test-time Model Adaptation in Non-stationary Mobile Environments. In *ACM Conference on Embedded Networked Sensor Systems (SenSys '24)*, November 4–7, 2024, Hangzhou, China. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3666025.3699339>

## 1 INTRODUCTION

Deep neural networks (DNNs) pre-trained in the cloud are increasingly deployed onto mobile devices for autonomous intelligence at the edge [20, 42–44, 70, 76], especially in some cities where cloud streaming is restricted [56, 69]. Applications such as on-device motion tracking (e.g., Google ARcore [1]) and AR/VR (e.g., Apple ARKit [2], Meta Spark [68]). These DNNs must operate reliably in open-world mobile environments, necessitating *swift model adaptation to unforeseen domain shifts* which represent differences in the distribution of the pre-trained model's data and the data encountered during test. The shifts are mainly caused by *environmental changes* (e.g., weather, lighting) and *sensor degradation* (e.g., low resolution and Gaussian noise). For instance, an autonomous car may encounter temporary road signs or modified traffic signals due to construction work and must adjust its DNNs with a few live samples collected on-the-fly [41, 74]. Such adaptation is crucial for maintaining safety and operational efficiency. Similarly, AR headsets require seamless integration of virtual elements into the physical world. As conditions such as lighting, field of view, and backgrounds vary with user movement, these headsets need to *promptly* refine their object recognition/tracking models based on limited input video clips, keeping the virtual overlays consistent with the evolving physical surroundings [11, 30, 84]. For instance, a vehicle driving at 90km/h requires its 60fps traffic sign recognition model to adapt within 16.6ms to ensure safety.

Test-time adaptation (TTA) offers a compelling solution to combat *unpredictable domain shifts* in mobile environments [38]. It is an emerging domain adaptation paradigm that (i) utilizes unlabeled test data, and (ii) operates independently of source data and supervision from pre-training [73]. TTA typically works in three stages. A *forward* pass first generates initial inference results on the shifted test samples, i.e., the live data sensed in a new environment. If inference confidence is low (e.g., high entropy), then a *backward* pass follows to adjust the DNN parameters using unsupervised loss based on initial forward results. Afterward, a *reforward* pass applies the updated DNN to the same batch of input data to make the final predictions. This continual, *unsupervised, source-free* adaptation is particularly advantageous for mobile applications, which often lack access to pre-training data due to privacy or network

restrictions. Moreover, these applications typically preclude the possibility of pre-labeling target domains due to the requisite labour and resources.

Despite advances in TTA algorithms [24, 73, 77], their practical adoption in mobile applications is challenging. The *forward-backward-reforward* pipeline of TTA, essential for adaptability to domain shifts, adds considerable *latency* compared to standard inference (see § 2.2), which compromises the *responsiveness* in *time-sensitive* mobile applications e.g., AR and autonomous cars, where even minor delays can drastically impact user experience and operational safety. Despite pioneer studies on efficient TTA [24, 51, 67], they focus on decreasing computational or memory cost, which cannot easily translate into reduced *wall-clock time*. Backward-free TTA [4, 79, 80] experience unacceptable accuracy degradation during continual adaptation. This gap necessitates *latency-efficient* TTA for *fast* and *accurate* adaptation to non-stationary environments for mobile computing.

In this paper, we propose AdaShadow, a novel framework for *responsive* test-time adaptation, like how a shadow *swiftly* responds to the body's movements. The principle is *sparse updating*, i.e., selectively refining critical layers to reduce both computation and memory demands during backpropagation, and thus latency [28, 49, 86]. Although the tactic has been explored to improve the latency efficiency of generic DNN training on mobile devices [28], sparse updating in TTA encounters unique challenges due to its *unsupervised* and *online* nature. We elaborate on these challenges below.

• **Challenge #1: identifying critical layers at low latency.** The importance of each layer differs drastically across different data distributions [35]. In non-stationary environments for mobile computing, each data batch may exhibit unique domain shifts, demanding rapid layer importance assessment. Prior works [28, 40] assess layer importance by computing the gradients using *labeled* data, which are computation-intensive (i.e., higher latency) and error-prone in TTA as test data are usually unlabeled in highly dynamic mobile scenarios. For instance, ElasticTrainer [28] results in 2.98s delay and 14.8% gradient errors for ResNet50 on unlabeled data.

• **Challenge #2: predicting runtime latency for layer update.** To ensure the updated model architecture meets the latency requirement, we have to know exactly the run-time latency of each retained layer in this new model. Prior works [28, 88] measure layer latency offline and then use it to predict the run-time model inference latency. Yet mobiles are interactive devices, their hardware resource availability will change dynamically due to user operations, foreground-background APP switching, and computation and memory resource competitions, which makes offline estimation inaccurate for online model inference prediction.

• **Challenge #3: online scheduling layer update strategy efficiently.** Even with accurate information on each layer's importance and their runtime latency, finding the optimal sparse updating strategy to improve performance in the presence of rigid delay requirements is still challenging due to the large search space. Dedicated strategies are necessary in exploring and pruning the search space with high efficiency.

AdaShadow addresses these challenges with three functional module designs.

**First**, we notice that the forward pass executed with each inference in TTA offers an opportunity to evaluate layer importance, avoiding

**Table 1: Differences of our work from related schemes.**

Method	Support unlabeled data	Online adaptation	Low latency	High accuracy
Efficient training: Melon [78]	No	No	No	Yes
Efficient training: ElasticTrainer [28]	No	No	Yes	Yes
Generic TTA: CoTTA [77]	Yes	Yes	No	Yes
Efficient TTA: EcoTTA [67]	Yes	Yes	No	Yes
Efficient TTA: LAME [4]	Yes	No	Yes	No
AdaShadow	Yes	Yes	Yes	Yes

the higher latency and imprecise results of backward gradients. Based on this observation, we propose a backpropagation-free layer importance assessor that can timely assess the importance of each layer by measuring the *divergence* between the layers' *output feature maps* in different environments, without requiring labeled data.

**Second**, to deliver precise and timely adaptation latency feedback, we introduce an online, unit-based (i.e., layer) adaptation latency predictor. This predictor distinguishes between static and dynamic update delays and incorporates fine-grained, dynamic system metrics like cache hit rates, competing CPU/GPU processes, and frequency into each unit's latency for accurate runtime measurements and improved online prediction.

**Third**, Given predictions of unit importance and update latency, we further develop a lightweight dynamic programming (DP)-based online scheduler. This scheduler efficiently determines the optimal layer update strategy for three-stage TTA by clearly defining specific subproblems, utilizing recursion, and eliminating invalid subproblems.

In the implementation, AdaShadow optimizes the TTA loss for *small batches*, which is common in mobile applications, and harnesses *computation reuse* between forward and reforward passes to boost performance. We evaluate AdaShadow across three real-world mobile scenarios, addressing over 30 types of data shifts and 5 types of resource dynamics, using three mobile devices with varied hardware architectures. Results show that AdaShadow achieves the best accuracy-latency balance in scenarios with continual shifts. Adashadow achieves a 2× to 3.5× adaptation speedup compared to state-of-the-art TTA methods, while maintaining comparable accuracy. Additionally, it offers a 14.8% to 25.4% accuracy improvement over state-of-the-art efficient supervised adaptation methods with similar latency (§ 5). Our main contributions are summarized as follows.

- To our knowledge, this is the first work on near-/real-time on-device DNN adaptation in non-stationary environments for mobile computing without labels or source data access. It overcomes latency bottlenecks of TTA in mobile contexts without compromising accuracy.
- We propose AdaShadow, a holistic system design that includes a backpropagation-free layer importance assessor, a runtime latency predictor, an online update scheduler, and optimizations for efficient, resource-aware TTA. It seamlessly integrates with mainstream TTA pipelines, supports various DNN architectures.

**Table 2: Accuracy and latency of state-of-the-art TTA methods against inference w/o adaptation.**

Method	NICO++			CIFAR10-C		
	ACC.(%)	Latency (ms)	Th.(fps)	ACC.(%)	Latency (ms)	Th.(fps)
Source	60.7	10.2	98.2	52.2	9.8	102.2
Tent[73]	87.8	33.2	30.1	77.8	25.8	38.7
EATA[28]	88.4	36.5	27.4	79.2	27.5	36.4
EcoTTA[67]	89.2	42	23.8	80.1	31.1	32.2

- Experiments show that AdaShadow outperforms existing research on generic TTA [73, 77], efficient TTA [24, 51, 67], and efficient training (with labels) [28, 40, 78] in trading-off accuracy and latency at low memory and energy costs over diverse tasks, shifts, and devices.

## 2 BACKGROUND

### 2.1 Primer on Test-time Adaptation (TTA)

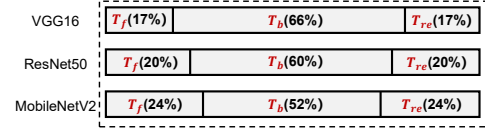
Test-time adaptation (TTA) is an emerging *unsupervised domain adaptation* paradigm to combat *domain shifts* [38]. It adapts the DNN *pre-trained* in the source domain to *unlabelled* target data during *testing*, enhancing the accuracy on target data [73]. Uniquely, TTA assumes no access to the *source data* and *supervision* from the pre-training stage. This setup is fit for mobile applications where (i) accessing source data is impractical due to privacy concerns or bandwidth limitations; and (ii) annotating the target domain is infeasible or labor-intensive. However, TTA often introduces significant latency (see details in § 2.2) or decline in accuracy if overfitting to the new environment [51, 77].

We explore *responsive* test-time adaptation of DNNs to *non-stationary* mobile domain shifts, with a focus on **latency efficiency**. This is because many mobile applications demand *real-time responsiveness*. Accordingly, the DNN should adapt swiftly, delivering accurate inference under continuous domain shifts without perceptible delays. Additionally, the non-stationary environment means the adaptation should (i) operate effectively with *small batches* of data; and (ii) account for *runtime resources dynamics*. Both requirements impose extra challenges when optimizing the latency of TTA (see § 1). Tab. 1 summarizes the differences of our work from other representative studies.

### 2.2 Latency of TTA

**Inefficiency of Prior Arts.** Despite emerging research on test-time adaptation for improved data efficiency [51] and memory efficiency [24, 67], latency remains a less-explored challenge. From Tab. 2, EcoTTA [67], a state-of-the-art, can be up to  $4.1\times$  slower than inference using the pre-trained model without adaptation (denoted as *source* in the table), even though they achieve higher accuracy on the drifted testing datasets. The unsatisfactory latency motivates us to zoom into the latency of TTA.

**Latency Bottleneck.** Latency is related to computing, memory access, and the availability of hardware resources [32, 88]. A typical TTA pipeline [38] includes three phases: *forward*, *backward*, and *re-forward*. In the first two phases (forward and backward), the model updates its parameters via standard gradient descent, based on the input batch data. In the last phase (re-forward), the model performs

**Figure 1: Latency breakdown of EcoTTA [67], a classic test-time adaptation scheme.**

inference on the *same* batch of data utilizing freshly updated parameters. This pipeline, known as the *sequential adaptation/inference mode*<sup>1</sup>, prioritizes inference accuracy by adapting the model before making predictions on it. However, it also implies that the inference for any batch of data must wait for the completion of adaptation, leading to noticeable latency. Formally, the overall TTA latency can be calculated as:

$$T = T_a + T_{re} = T_f + T_b + T_{re} \quad (1)$$

where  $T_a$  and  $T_{re}$  are the latency of model adaptation and inference (re-forward), and  $T_a$  is further decomposed into  $T_f$  and  $T_b$ , i.e., latency of forward and backward, respectively.

Fig. 1 show the latency breakdown of EcoTTA [67], a classic TTA method with three model architectures on NVIDIA Jetson NX, an off-the-shelf edge device. The *backward latency*  $T_b$  takes up over 50% of the overall delay, making it a *bottleneck*. This is because the forward pass only computes the activations, whereas the backward pass calculates the *gradients* for both the *activations* and *parameters*, which doubles the *computation* and demands *extra memory access* to the activations from the forward pass [37, 48].

### 2.3 Problem Statement

Our primary strategy to reduce adaptation latency is *sparse updating* [28, 34, 40], i.e., selectively updating a subset of layers that are crucial for TTA. It reduces the *computation cost of gradients* and the *memory accesses to retrieve intermediate activations*, thereby lowering latency in the backward pass. However, we observe that *updating fewer layers does not guarantee a proportional decrease in TTA latency on mobile devices*. As illustrated in Fig. 2, the latency for updating the first layer is comparable to updating the last four layers. This is because updating the first layer still necessitates *computing gradients* for all layers, leading to significant computation and memory access overhead.

Formally, we explore low-latency test-time adaptation to the unseen environment  $e$  via sparse updating by formulating the following constrained optimization problem.

$$\max \tilde{S}_e \cdot \vec{A}_e \quad s.t. \quad T_f + T'_b(\tilde{S}_e) + T'_{re}(\tilde{S}_e) \leq \sigma \cdot T \quad (2)$$

where  $\tilde{S}_e^*(N)$  is the optimal sparse updating strategy, which is a binary vector of length  $N$  (the number of layers), and  $\vec{A}_e$  is the layer importance vector of the same format as  $\tilde{S}_e^*(N)$ .  $T'_b$  and  $T'_{re}$  are the optimized backward and reforward latency under strategy  $\tilde{S}_e^*(N)$ , respectively.  $T$  is the overall latency without sparse updates, covering the time for a data batch to process from input to prediction

<sup>1</sup>A few studies [51, 73] prioritize latency over accuracy in process scheduling by only applying the newly adapted model to subsequent sample inference, thus eliminating the reforward latency. Our method achieves higher accuracy with lower latency than this execution mode (see § 5.4).

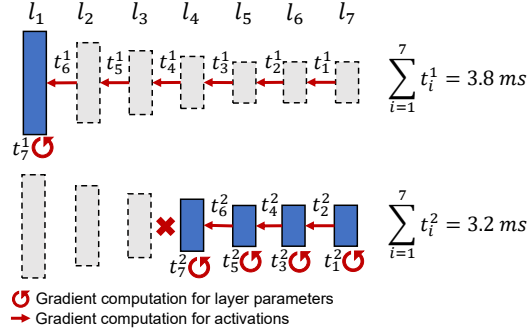


Figure 2: Example of sparse updating, where updating 1 or 4 layers results in almost the same latency.

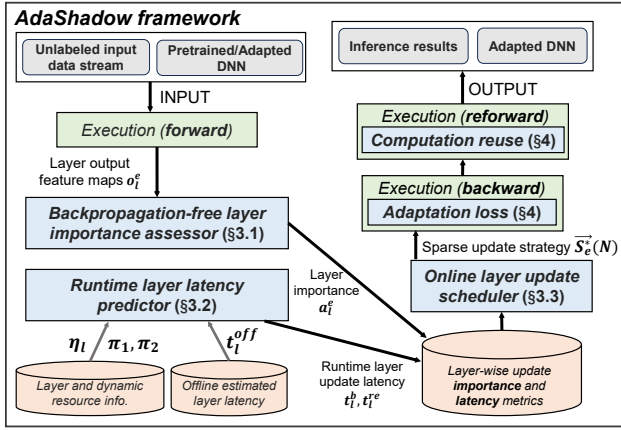


Figure 3: AdaShadow overview.

results, including forward, backward, and reforward stages.  $\sigma$  denotes the expected acceleration factor. Note that  $T_f$  is constant with diverse sparse update strategies. Sparse updating primarily reduces the backward latency, yet it also offers opportunities to decrease the reforward latency, as will be explained in § 4. We formulate latency as a constraint with a tunable acceleration factor for easy configuration of the real-time requirements in diverse mobile applications with streaming data, e.g., video processing [26, 46] (image classification, object detection, scene understanding) of vehicle cameras or AR headsets [33, 36, 55].

### 3 SYSTEM DESIGN

**System Overview.** Fig. 3 shows the system architecture of AdaShadow. It consists of three functional modules: (i) the *backpropagation-free layer importance assessor* resolves *Challenge #1* by evaluating layer importance via the *divergence* of output feature embeddings across environments, utilizing *unlabeled data* in the forward pass (§ 3.1). (ii) the *runtime layer latency predictor* overcomes *Challenge #2* by integrating mobile resource dynamics into the latency modeling for online calibration of offline estimates (§ 3.2). (iii) the *online layer update scheduler* addresses *Challenge #3* via a lightweight dynamic programming strategy to search for the

Table 3: Overhead of layer importance assessors evaluated on Raspberry Pi.

Methods	Latency (s)		Memory usage (MB)	
	ResNet50	MobileNetV2	ResNet50	MobileNetV2
TTE [40]	167.24	65.31	2108	695
ElasticTrainer [28]	2.98	1.29	2108	695
AdaShadow	<b>0.11</b>	<b>0.09</b>	<b>0.0086</b>	<b>0.0083</b>

optimal layer update strategy with high efficiency (§ 3.3). Additionally, we present the implementation and additional optimization of the AdaShadow system workflow in § 4.

#### 3.1 Backpropagation-free Layer Importance Assessor

A prerequisite for sparse updating is to estimate the layer importance for on-device model adaptation. This process should be *latency- and memory-efficient* to allow frequent revocation on low-resource mobile devices in non-stationary environments. This subsection presents a backpropagation-free approach to significantly enhance the efficiency of prior schemes [28, 40, 57] for layer importance assessment.

**Limitations of Prior Arts.** Existing studies [28, 40] rely on *gradients* to assess layer importance on adaptation accuracy, which faces two drawbacks. (i) Accurate *gradients* require *labels*, which are absent in mobile-end TTA. (ii) More critically, even with labels, the reliance on backpropagation to compute gradients for assessing layer importance results in substantial computational and memory overhead, as gradients must be propagated and stored for each layer. For instance, TTE [40] and ElasticTrainer [28] incur latencies of up to 167.2s and 2.98s, respectively, while consuming over 2,108 MB of memory for intermediate activations. Such overheads are unacceptable for *near-/real-time* DNN adaptation on mobile devices.

We propose to assess layer importance with *unlabeled data* merely in the *forward pass* for responsiveness by measuring the *divergence* between the layers' *output feature maps* in different environments. The idea is inspired by the distribution alignment techniques [47, 61], which quantifies drifts between the source and target data via the distributions of intermediate features in the forward stage. The *rationale* is that when a DNN encounters domain shifts from a new environment, it will be less confident about its predictions, as reflected by a deviation in the layer's output distribution from that observed in historical environments.

We cater this principle for TTA context as follows. Given an input data batch  $x_e$  from a new environment  $e$ , we assess the importance  $a_l^e$  of layer  $l$  in adapting to  $x_e$  by measuring the Kullback-Leibler (KL) divergence between the embedding  $E_l^e$  of layer  $l$ 's output feature  $o_l^e$  on  $x_e$ , i.e.,  $E_l^e = g(o_l^e)$ , and the embedding  $E_l^H$  on the average output features on historical environments  $H$ . Specifically, the layer importance  $a_l^e$  is estimated as:

$$a_l^e = D_{KL}(g(o_l^H) || g(o_l^e)) \quad (3)$$

where  $D_{KL}(\cdot)$  denotes the KL divergence. We justify our layer importance metric as follows.

- The *KL divergence* effectively quantifies the data distribution shift in mobile environment  $e$  from the history  $H$  perceived

at layer  $l$ . A larger divergence indicates a greater necessity to update layer  $l$ . Since we target at adaptation on data streams, *e.g.*, live videos, the domain shifts between adjacent batches tend to be mild. Hence KL divergence can effectively handle mild domain shifts, it is unnecessary to adopt computation-intensive metrics, *e.g.*, the Wasserstein distance, which are intended for distributions with low overlaps [63].

- Rather than the raw output features, we measure the divergence of their *embedding* via an embedding function  $g(\cdot)$ . It normalizes output features to the same dimensions for fair importance evaluation across layers. More critically, it allows robust estimation with small batches, which is essential for non-stationary environments, as will be discussed next.

Furthermore, we implement the layer importance metric with two optimizations for mobile environments.

- For efficient distribution estimation with *small-batch* mobile data, we devise  $g(\cdot)$  with only first and second-order moments. Compared to existing methods [5, 87] that minimize differences in higher-order sample moments, it can effectively extract sufficient distribution information from small batch data. Concretely, we compute the channel-wise means  $\mu(o_{l,c}^e) = \frac{1}{N} \sum_{c \in \mathbb{C}} o_{l,c}^e$ , and the channel-wise variances  $\sigma^2(o_{l,c}^e) = \frac{1}{N} \sum_{c \in \mathbb{C}} (o_{l,c}^e - \mu(o_{l,c}^e))^2$  from layer  $l$ 's output features  $o_{l,c}^e$ . The embedding  $E_l^e$  is then represented as  $[\mu(o_{l,1}^e), \sigma^2(o_{l,1}^e), \dots, \mu(o_{l,N_c}^e), \sigma^2(o_{l,N_c}^e)]$ .
- To track the *non-stationary* mobile environments, we gradually update the historical embedding  $E_l^H = g(o_l^H)$  by integrating the new environment  $e$  into  $H$  via a moving average strategy:  $g(o_l^H) = \alpha \cdot g(o_l^e) + (1 - \alpha) \cdot g(o_l^H)$ , where  $\alpha \in [0, 1]$  is a hyperparameter controlling the incorporating rate of new environment  $e$ .

The layer importance assessor reuses existing embeddings, eliminating the need for additional forward propagation. This limits the computational cost to statistic extraction and KL divergence calculation. For example, in ResNet50, our importance profiler incurs only 0.2 GFLOPs overhead for these computations, significantly lower than the 4.1 GFLOPs saved by halving backpropagation when  $\sigma = 0.5$ . Consequently, the theoretical latency can be optimized to as low as 4.8% of the original (0.2 GFLOPs) or up to 104.8% of the original (4.3 GFLOPs) when  $\sigma = 1$ . As shown in Tab. 3, our layer importance assessor reduces evaluation latency by 95% compared to gradient-based methods like TTE [40] and ElasticTrainer [28]. It also has low memory requirements, using less than 8.6 KB to store historical embeddings for ResNet50. Importantly, this efficiency does not compromise accuracy; our method accurately identifies the importance of individual layers for TTA, as illustrated by the blue bars in Fig. 4, which represent the layers selected for updating.

### 3.2 Runtime Layer Latency Predictor

Precise layer update latency estimation is another key input to determine the optimal sparse updating strategy. As we target at TTA on mobile devices with limited and dynamic resources, *offline* predicting schemes [28, 88] incur large estimation errors. This subsection introduces an *accurate* and *lightweight* approach to model

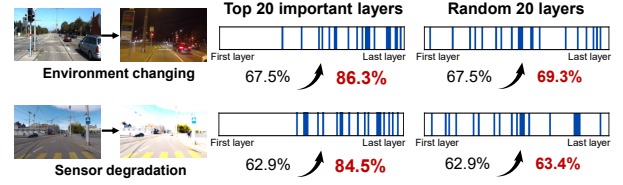


Figure 4: Accuracy gains when updating top 20 important layers or 20 random layers (blue bars).

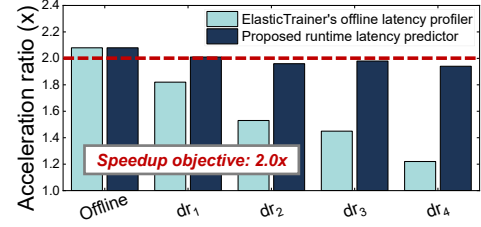


Figure 5: Comparison of offline/runtime latency predicting with dynamic resource availability.

and calibrate the backward<sup>2</sup> latency of individual layers to *runtime* resource dynamics.

We empirically demonstrate the necessity of estimating the training latency at runtime. Specifically, we first apply ElasticTrainer [28], the state-of-the-art sparse training scheme that predicts layer latency offline, to speed up the adaptation of a ResNet50 and NVIDIA Jetson NX by an acceleration ratio of about 2× under *abundant resources* (denoted as *offline* in Fig. 5). Then we simulate four dynamic resource conditions [62]:  $dr_1$ : increase the mobile CPU/GPU temperature to 60 °C;  $dr_2$ : add three competing compute-intensive processes;  $dr_3$ : occupy the cache to tune the cache hit rate to be 30%;  $dr_4$ : combine all factors from  $dr_1 \sim dr_3$ . When running ElasticTrainer [28] in these four mobile system conditions using configurations obtained in the offline case, we observe mild to drastic decrease in acceleration ratios (see Fig. 5). The experiment shows that the *offline layer latency* estimates can be significantly *inaccurate* due to dynamic resource availability. [62] incorporates dynamic resources into latency predictions using a Graph Neural Network (GNN), however, incurring substantial overhead. Furthermore, latency profiling is ineffective without actual execution, as noted in [10]. In contrast, our proposed method delivers higher acceleration ratios by adapting to these runtime conditions.

**Key Idea.** We observe that the execution latency of a DL unit mainly hinges on on-chip kernel computation and off-chip memory access. Each unit's kernel execution strategy and memory allocation are static, whereas kernel utilization, temperature, and cache-hit-rate are dynamic and measurable at runtime. Moreover, this unit-based approach is versatile across different DL models. For example, basic units in ResNet include Conv2d, BatchNorm, and Linear, while

<sup>2</sup>Although we optimize both the latency during backward and reforward (see Equ.(2)), we mainly model and calibrate the estimation on backward latency for two reasons. (i) The backward pass is the latency bottleneck (see § 2.2). (ii) In practices, the reforward latency can be easily hidden via pipelining adaptation and inference (see § 5.4).



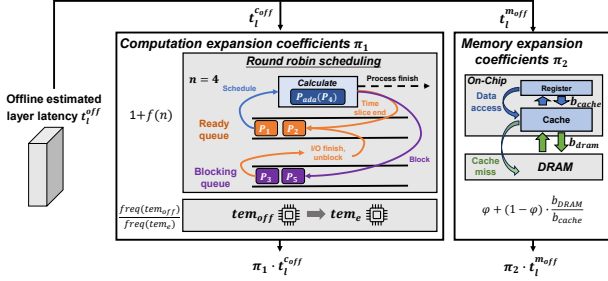


Figure 6: Layer update latency in offline and online.

in Transformer, they consist of projectors Q, K, V, LayerNorm, and the feed-forward network (FFN).

**Methods.** In offline predicting with *abundant, static* resources, the latency  $t_l^{off}$  of layer  $l$  can be estimated as  $t_l^{off} = t_l^{c_{off}} + t_l^{m_{off}}$ , where  $t_l^{c_{off}}$  and  $t_l^{m_{off}}$  are corresponding latency for computation and memory access profiled offline, respectively. We account for *limited, dynamic* runtime resources on device by modeling the layer latency  $t_l = \pi_1 \cdot t_l^{c_{off}} + \pi_2 \cdot t_l^{m_{off}}$ , where  $\pi_1 > 1$  and  $\pi_2 > 1$  are the computation and memory expansion coefficients. As measuring  $t_l^{c_{off}}$  and  $t_l^{m_{off}}$  separately is challenging, we correlate  $t_l$  to  $t_l^{off}$  as follows.

$$t_l = (\pi_1 \cdot \frac{\eta_l}{\eta_l + 1} + \pi_2 \cdot \frac{1}{\eta_l + 1}) \cdot t_l^{off} \quad (4)$$

Note that  $t_l^{off}$  is measurable offline. To simplify Eq. 3.2, we define  $\eta_l = \frac{t_l^{c_{off}}}{t_l^{m_{off}}}$ , a layer-dependent metric that remains constant regardless of runtime resource conditions. This metric captures the effects of the computation expansion coefficient  $\pi_1$  and the memory expansion coefficient  $\pi_2$  on overall unit latency. By using distinct coefficients for computation and memory, we can effectively profile their impacts on runtime performance. The two coefficients  $\pi_1$  and  $\pi_2$  reflect resource limitations on computation and memory latency, applicable across layers and measurable during runtime. We'll detail how to assess these three hyperparameters next.

• **Profiling  $\pi_1$ .** The computation expansion coefficient  $\pi_1$  depicts two runtime factors that affect computation latency: (i) number  $n$  of competing processes, and (ii) core temperature  $tem^{on}$  of mobile CPU/GPU.

- The OS kernel of the mobile CPU [3, 65] and GPU [54, 72] often employs round robin scheduling (see Fig. 6). Hence an increase in process number leads to a linear growth in process waiting time.
- Dynamic Voltage Frequency Scaling (DVFS) [22, 31, 39] is often activated in the mobile CPU/GPU to avoid overheating by reducing the clock frequency. This also increases the computation time.

Accordingly, we set  $\pi_1 = \frac{freq(tem^{off})}{freq(tem^{on})} \cdot [1 + f(n)]$ , where  $f(\cdot)$  is a linear function determined offline depicting the process switching overhead. The dynamic frequency  $freq(\cdot)$  accounts for the DVFS.

• **Profiling  $\pi_2$ .** We relate the memory expansion coefficient  $\pi_2$  to the cache-hit-rate  $\varphi$ , which is directly measurable at runtime. This is because sharing the limited cache among processes increases the memory access latency. Also, in case of cache miss, there is an extra latency for data movement, which is determined by the ratio between the bus bandwidth  $b_{DRAM}$  and the cache bandwidth  $b_{cache}$  (see Fig. 6). Hence, we set  $\pi_2 = \varphi + (1 - \varphi) \cdot \frac{b_{DRAM}}{b_{cache}}$ .

• **Profiling  $\eta_l$ .** Although it is challenging to measure  $t_l^{c_{off}}$  and  $t_l^{m_{off}}$  separately, their ratio  $\eta_l = \frac{t_l^{c_{off}}}{t_l^{m_{off}}}$  can be profiled offline. Specifically, we transform  $\eta_l = \frac{c^l \cdot \delta_c}{m^l \cdot \delta_m}$ , where  $c^l$ ,  $m^l$ ,  $\delta_c$ , and  $\delta_m$  denote

amount of MAC, amount of memory accesses, unit MAC time, and unit memory access time of layer  $l$ , respectively.  $c^l$  and  $m^l$  can be directly derived according to the type and hyperparameters of layer  $l$ .  $\delta_c$  is obtained from the processor's FLOPS  $F$ :  $\delta_c = \frac{1}{F}$ , while  $\delta_m$  is obtained from the cache bandwidth  $b_{cache}$ :  $\delta_m = \frac{1}{b_{cache}}$ . We also include non-parameterized layers, e.g., activation layers, in  $t_l^{c_{off}}$  as they also involve in gradient computation. The cost of latency prediction only involves sensing dynamic resource states and calculating execution latency using Eq. , resulting in negligible computational and memory overhead.

### 3.3 Online Layer Update Scheduler

This subsection presents a dynamic programming (DP) based lightweight online scheduler that efficiently decides the best layer update strategy  $\tilde{S}_e^*(N)$  for Equ.(2), given the layer-wise metrics of a DNN. Picking the optimal solution to Equ.(2), an NP-hard nonlinear integer programming problem, can be time-consuming[40]; We formulate specific subproblems, recursion, and space for three-stage TTA.

Consider a DNN with  $N$  layers. Let the importance and update latency of layer  $l$  (counting from the last layer) be  $a_l^e$  and  $t_l^e$  in environment  $e$ . A brute-force search for the optimal update strategy  $\tilde{S}_e^*(N)$  has a complexity of  $O(2^N)$ , which is impractical for online scheduling. To address this, we propose a dynamic programming (DP) formulation to solve Equ.(2). Specifically, we aim to find the optimal layer selection  $\tilde{S}_e^*(N)$  that maximizes cumulative importance within the latency budget  $T_{bgt} = \sigma \cdot T - T_f$ . Let  $P[l][t]$  represent the maximum cumulative layer importance from the last layer to layer  $l$  for a given latency budget  $t$ . The binary vector  $\tilde{S}_e[l][t]$  corresponds to one solution for  $P[l][t]$ , while  $\tilde{S}_e^*[l]$  denotes the optimal solution. We formulate the DP at the granularity of  $t$  and  $l$ .

- **Discrete time:** We discretize time into  $N_T$  units, i.e., at a resolution of  $T_{bgt}/N_T$ . We empirically set  $N_T = 500$  which effectively balances search cost and accuracy.
- **Layer-wise:** It is natural to decompose sub-problems layer-wise to support diverse DNN architectures. It also aligns with the layer-wise DNN execution scheme on mobile devices with limited resources [88]. Finer-grained decomposition e.g., tensor-wise as [28] can be inefficient ( $214 \cdot N_T$  sub-problems in tensor-wise vs.  $107 \cdot N_T$  sub-problems in layer-wise for ResNet50).

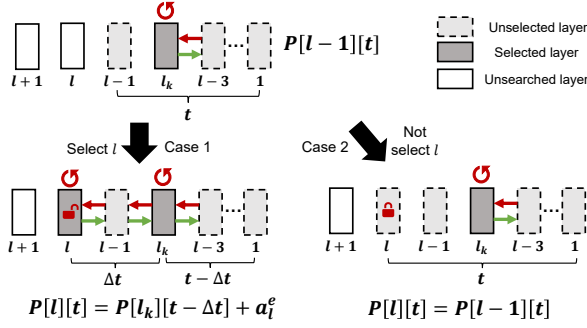


Figure 7: Illustration of recursion in two different cases.

Accordingly,  $0 \leq l \leq N$  and  $0 \leq t \leq N_T$ , and the DP formulation would reduce the time complexity to  $O(N^2 \cdot N_T)$  considering the time complexity  $O(N)$  of solving each sub-problem (details below).

The base cases are trivial. We set  $P[0][t]$  for all  $0 \leq t \leq N_T$  as 0. This is because the maximum cumulative layer importance is 0 if no layer is selected for updating. The recursion for  $P[l][t]$  when  $l > 0$  is more subtle.

i) **Case 1:** If layer  $l$  is selected, then the maximum cumulative layer importance becomes  $P[l_k][t - \Delta t] + a_l^e$ , where layer  $l_k$  is the last layer selected in  $P[l-1][t]$ , i.e., the closest selection to layer  $l$ . This is because updating one more layer decreases the latency budget when backpropagating till layer  $l-1$  by  $\Delta t$  (details below) and increases importance by  $a_l^e$ .

ii) **Case 2:** If layer  $l$  is not selected, the maximum cumulative layer importance remains  $P[l-1][t]$ .

As shown in Fig. 7, we clarify the two cases assuming  $l_k = l-2$ . Since we search for the maximum, the final recursion becomes:  $P[l, t] = \max\{P[l-1, t], P[l_k, t - \Delta t] + a_l^e\}$ . Note that the extra latency  $\Delta t$  induced by selecting layer  $l$  in Case 1 is not simply its layer update latency  $t_l^e$ . From Fig. 7,  $\Delta t = t_l^{\delta w} + \sum_{m=l_k}^l t_m^{\delta x} + \sum_{m=l_k+1}^l t_m^{re}$ , where  $t_l^{\delta w}$  is the gradient calculation time of weight  $w_l$ ,  $\sum_{m=l_k}^{l-1} t_m^{\delta x}$  is the gradient calculation time of activations between layer  $l-1$  and layer  $l_k$ , and  $\sum_{m=l_k+1}^l t_m^{re}$  is the reforward time between layer  $l$  and layer  $l_k+1$ . Noting that  $t_l^b = t_l^{\delta x} + t_l^{\delta w}$ , we utilize the amount of MAC of  $t_l^{\delta x}$  and  $t_l^{\delta w}$  to infer their proportion of latency in  $t_l^b$ . As  $l_k$  is unknown in advance, we gradually decrease  $l_k$  from  $l-1$  to 0 to explore all possible  $l_k$  to find the  $l_k$  that maximizes  $P[l_k][t - \Delta t]$  with a time complexity of  $O(N)$ .

Given the above formulations, we can explore all  $P[l][t]$  recursively till  $P[N][T_{bgt}]$ , and then we can search the optimal layer update strategy  $\tilde{S}_e^*[N]$ . To further improve the efficiency, we refine the search space by removing two types of invalid sub-problems.

- For  $T_{bgt}$ , we iteratively exclude sub-problem  $P[l, t]$  if  $t$  exceeds the latency budget  $T_{bgt}$ . We can also discard all sub-problems generated from an invalid  $P[l, t]$ .
- For  $t$  in sub-problem  $P[l, t]$ , if the gradient computation time  $\sum_{m=1}^l t_m^{\delta x}$  exceeds  $t$ , the total latency of  $P[l, t]$  will also exceed  $t$  even if no layers are selected for updating. Hence, we can discard such  $P[l, t]$  and all sub-problems it generates.

To eliminate redundant calculations, we identify overlapping solutions by formulating the latency of both backward and reforward into the same recursive sub-problem. This is because, as noted in § 2.3, the position of the first update layer impacts the latency of both backward and reforward simultaneously.

## 4 IMPLEMENTATION

This subsection presents how the layer importance assessor (§ 3.1), layer latency predictor (§ 3.2), and layer update scheduler (§ 3.3) cooperates in the forward-backward-reforward pipeline (see Fig. 3) as well as our additional optimizations to improve the accuracy and efficiency of TTA on mobiles.

At initialization, DNNs are pre-trained in the cloud using source datasets and then deployed to mobile devices. Upon receiving a batch of test samples  $x_e$  from a new environment  $e$ , AdaShadow adapts the DNN as follows.

- **Forward phase.** AdaShadow employs the layer importance assessor and the latency predictor to derive the layer importance  $a_l^e$  and the backward and reforward latency  $t_l^b$  and  $t_l^{re}$  for each layer of the DNN. Meanwhile, it prepares each layer's output feature maps  $o_l^e$  and computation graph for the backward phase.
- **Backward phase.** AdaShadow calls the layer update scheduler to search for the optimal strategy  $\tilde{S}_e^*(N)$ , based on  $a_l^e$ ,  $t_l^b$ , and  $t_l^{re}$  obtained in the forward stage. AdaShadow prunes computation graph nodes of layers not in  $\tilde{S}_e^*(N)$  to skip gradient calculation at the compiler level, to realize the selective propagation process. The training loss will be discussed later.
- **Reforward phase.** AdaShadow infers on  $x_e$  using the updated model from the backward stage. AdaShadow partially reuses computation results from the forward stage to accelerate the inference, as described next.

We propose two more optimizations during backward and reforward to boost the performance.

- **Adaptation loss for small batches.** Existing TTA methods [52, 73] mainly use the entropy of the final layer output as the unsupervised adaptation loss, often resulting in low accuracy with small batches [52]. In AdaShadow, we refine the loss  $L$  by summing the KL-divergence between the new environment and the history from all layers, i.e.,  $L = \sum_{l=1}^N D_{KL}(E_l^H \| E_l^e)$ . Evaluations show that the refined loss is more robust to small batch sizes.
- **Memory I/O-aware computation reuse between forward and reforward.** Since both forward and reforward computations are performed on the same batch, we can partially reuse computations from the forward stage to reduce overall latency, as illustrated in Fig. 8. We track memory addresses for input tensors during forward. After defining the layer update strategy, we detach the input activation  $x_n$  of the first updated layer  $n$  in  $\tilde{S}_e^*(N)$  from the gradient computation graph and retain it in memory, while releasing input tensors  $x_i$  for other layers. Before reforward begins, we load  $x_n$  from memory and reuse computation results for layers not updated during the backward pass, skipping layers before the first updated layer  $n$  during reforward.

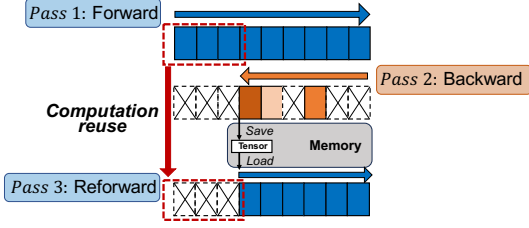


Figure 8: Illustration of memory I/O-aware computation reuse implementation.

## 5 EVALUATION

### 5.1 Experiment Setup

**Implementation.** We prototype AdaShadow on three mobile devices.  $D_1$ : NVIDIA Jetson NX (Cortex-A57 CPU and 256 CUDA cores GPU);  $D_2$ : NVIDIA Jetson Nano (Cortex-A57 CPU and 128 CUDA core GPU);  $D_3$ : Raspberry Pi 4B (Cortex-A72 CPU). We installed Jetpack 4.6.4 OS based on Ubuntu 18.04 on  $D_1$  and  $D_2$ , and the 64-bit Raspbian 11 OS on  $D_3$ . Before DNN deployment, AdaShadow measures the multi-process scheduling latency function  $f(n)$  and the offline layer execution latency via one pass of backpropagation using PyTorch profiler plugin, and detects frequency-temperature function  $freq(tem)$  through the device DVFS strategy. During test-time adaptation, AdaShadow uses the SGD optimizer with a learning rate  $lr$  of  $5e-3$ .

**Tasks, Datasets, and Models.** We test three mobile sensing tasks with diverse continual domain shifts.

- **NICO++ [89].** It covers 6 types of *environmental* shifts, including outdoor, autumn, dim, grass, rock, and water. The model is pre-trained on 'outdoor' for adaptation to other five environments. NICO++ are commonly used datasets for evaluating domain adaptation [18, 90].
- **CIFAR10-C and CIFAR100-C [21].** CIFAR10-C and CIFAR100-C has been widely used for evaluation adaptation performance in mobile systems [12, 13, 16]. They simulate 15 types of *corruption* shifts, induced by lens shakes, low-light conditions, and sensor degradation such as Gaussian noise, motion blur, pixelation, *etc.* We experiment with two dataset combinations: [CIFAR10, CIFAR10-C] and [CIFAR100, CIFAR100-C], where the first is for pre-training and the second for adaptation.
- **Self-collected mobile videos.** It contains 10 natural, continual, potentially *mixed* shifts collected in open mobile environments. Specifically, we built an Ackermann vehicle platform equipped with Jetson NX development board, STM32 control board, visual sensors, *etc.* to move on campus and collect 10 hours of videos. We manually sample video clips that contain transitions of shifts, *e.g.*, indoor-wild, sunny-rainy-cloudy, daytime-night, and different view angles, *etc.*
- **Self-collected layer adaptation latency records.** We build a dataset of 15,000 records capturing various real-world system factors and layer adaptation latency.

We adopt three typical model architectures, *i.e.*, vision transformer (ViT-B/16) [14], VGG16 (without BN layer) [64], MobileNetV2 (with BN) [60], and ResNet50 (with BN) [19].

**Baselines.** Generic and supervised TTA set strict accuracy lines, while efficient TTA set latency benchmarks.

- **Generic TTA (no target label).** They aim at accurate TTA without explicit efficiency optimization.
  - **Tent [73]:** the first TTA that updates the *BN* layers via entropy minimization.
  - **CoTTA [77]:** another representative TTA strategy via *knowledge distillation* for anti-forgetting.
- **Efficient TTA (no target label).** Besides accuracy, this category optimizes efficiency metrics of TTA.
  - **EATA [51]:** *data-efficient* TTA by updating BN layers on selective samples.
  - **EcoTTA [67]:** typical *memory-efficient* TTA that reduces the memory footprint to update the BN layers with lightweight adapters.
  - **MECTA [24]:** *memory-efficient* TTA that selectively updates the proposed *adaptive BN* layers.
  - **BFTT3D [80]:** *backpropagation-free* TTA via comparing source and target features.
  - **LAME [4]:** a *latency-efficient* TTA that adapts the model's output rather than parameters using manifold-regularized likelihood loss.
  - **OFTTA [79]:** a *latency-efficient* TTA that selectively adjust the proposed *exponential decay BN* layers.
- **Efficient Training (supervised by target labels).** This category focuses on efficient training on low-resource mobile and embedded devices. They provide unsupervised loss (*i.e.*, entropy) at test time.
  - **Melon [78]:** a *memory-efficient* on-device training scheme that uses recomputation and micro-batch to fit into memory-scarce mobile devices.
  - **TTE [40]:** a *memory-efficient* on-device training scheme that selectively updates tensors via *offline* profiling.
  - **ElasticTrainer [28]:** a *latency-elastic* on-device training method that uses *runtime* profiling to identify important tensors and *offline* latency profiling, with the importance metric based on *backpropagation*.

**Process Schedule Modes.** EATA(P) and EcoTTA(P) refer to the EATA and EcoTTA baselines implemented in parallel adaptation/inference mode. In contrast, other baselines operate in sequential adaptation/inference mode, applying the newly adapted model to subsequent sample inference for accuracy testing, as discussed in § 2.2.

### 5.2 Main Results

**5.2.1 Accuracy vs. Latency.** This experiment evaluates performance during adaptation to two types of *continual shifts*. We pre-train a ResNet50 on the "outdoor" sub-dataset in NICO++ and sequentially adapt it to five other sub-datasets with *environmental* shifts (autumn, dim, grass, rock, and water) on  $D_1$  (see Tab. 4). AdaShadow achieves a better balance between adaptation accuracy and latency, achieving  $2\times \sim 3.5\times$  lower latency than Tent, EATA,



**Table 4: Test accuracy (%), adaptation throughput (fps), and latency (ms) with continual environmental shifts.**

Time		Continual data shift over time →										Mean ACC.(%)	Th. (fps)	Latency (ms)
Dataset	Method	Autumn	Dim	Grass	Rock	Water	Autumn	Dim	Grass	Rock	Water			
NICO++	TENT [73]	84.2	81.8	86.6	86.2	86.5	89.9	89.1	91.3	90.9	91.7	87.8	30.1	33.2
	CoTTA [77]	85.0	83.3	87.7	86.3	85.8	90.3	89.2	92.4	90.5	91.2	88.2	4.2	238.1
	EATA(C) [51]	81.8	79.5	84.2	83.7	82.3	86.0	86.1	87.9	88.6	86.9	84.7	35.3	28.3
	EATA(O) [51]	85.2	83.0	87.5	86.6	86.1	90.1	89.8	92.2	91.8	91.7	88.4	27.4	36.5
	EcoTTA(C) [67]	84.6	80.9	85.5	84.7	84.3	88.0	86.9	91.0	89.8	89.5	86.5	31.2	32.1
	EcoTTA(O) [67]	85.4	<b>84.9</b>	88.3	87.4	86.5	91.7	90.7	92.9	91.9	92.3	89.2	23.8	42.0
	MECTA [24]	86.6	83.6	88.4	88.2	87.0	<b>92.1</b>	<b>91.0</b>	93.6	91.7	92.7	89.5	31.2	32.1
	Melon [78]	72.0	51.4	47.0	40.3	26.1	23.4	13.9	9.4	9.1	5.4	29.8	22.0	45.5
	TTE [40]	68.7	65.8	71.5	70	68.9	73.8	72.2	74.7	74	75.5	71.5	29.5	33.9
	ElasticTrainer [28]	72.3	69.9	73.7	74.2	72.3	76.3	77	78.3	78.5	78.5	75.1	57.6	17.4
	AdaShadow	<b>86.9</b>	84.8	<b>89.0</b>	<b>88.5</b>	<b>87.7</b>	91.6	<b>91.0</b>	<b>93.7</b>	<b>92.9</b>	<b>92.9</b>	<b>89.9</b>	<b>76.4</b>	<b>13.1</b>

**Table 5: Test accuracy (%), adaptation throughput (fps), and latency (ms) with continual corruption shifts.**

Time		t →															Mean ACC.(%)	Th. (fps)	Latency (ms)
Dataset	Method	Gaus.	Shot	Impu.	Defo.	Glas.	Moti.	Zoom	Snow	Fros.	Fog	Brig.	Cont.	Elas.	Pixe.	Jpeg			
CIFAR-10C	TENT [73]	73.1	75.2	67.8	79.6	68.6	81.2	85.6	78.0	87.1	86.5	81.8	<b>90.3</b>	69.5	77.4	<b>77.3</b>	78.6	29.2	34.3
	EATA [51]	69.4	71.5	71.6	83.5	70.6	81.8	85.9	79.5	83.1	86.1	87.1	86.8	75.1	83.1	73.1	79.2	27.5	36.4
	EcoTTA [67]	65.7	68.7	70.6	78.1	68.1	79.7	<b>87.9</b>	77.3	85.8	89.3	90.2	82.3	72.9	77.4	72.9	77.8	23.8	42.1
	MECTA [24]	<b>74.7</b>	70.1	70.4	83.2	68.5	<b>84.9</b>	83.3	79.2	85.2	<b>90.9</b>	91.9	85.2	76.3	85.7	72.2	80.1	31.0	32.3
	BFTT3D[80]	44.5	45.3	51.2	61.9	41.7	54.9	66.0	56.4	60.0	65.7	65.3	57.9	51.9	58.9	47.9	55.3	163.9	6.1
	LAME[4]	47.8	43.4	50.6	59.0	44.9	55.4	67.4	56.6	62.1	58.5	62.6	68.0	55.7	62.6	45.3	56.0	175.4	5.7
	OFTTA[79]	48.6	46.1	46.9	57.9	47.0	64.6	63.3	54.5	57.6	67.6	66.9	66.4	53.9	61.6	50.6	56.9	<b>192.3</b>	<b>5.2</b>
	ElasticTrainer [28]	53.6	59.9	59.8	72.8	57.9	70.4	74.3	68.7	71.5	69.2	73.7	75.8	61.4	67.2	60.5	66.4	58.6	17.1
	AdaShadow	69.2	<b>75.4</b>	<b>72.6</b>	<b>83.6</b>	<b>71.7</b>	79.6	84.7	<b>85.2</b>	<b>87.4</b>	85.8	<b>92.6</b>	84.8	<b>76.6</b>	<b>85.9</b>	74.8	<b>80.7</b>	79.1	12.6

**Table 6: Performance comparison on CIFAR100-C.**

Time		Mean ACC.(%)	Th.(fps)	Latency(ms)
Dataset	Method			
CIFAR100-C	TENT [73]	60.1	26.1	38.3
	EATA [51]	58.8	24.3	41.2
	EcoTTA [67]	59.3	21.5	46.6
	MECTA [24]	61.8	27.2	36.8
	BFTT3D[80]	44.1	158.7	6.3
	LAME[4]	41.8	163.9	6.1
	OFTTA[79]	42.5	<b>175.4</b>	<b>5.7</b>
	ElasticTrainer [28]	49.1	56.9	17.6
	AdaShadow	<b>62.2</b>	76.2	13.1

and EcoTTA, and up to 18.2× lower than CoTTA. While ElasticTrainer achieves a 2× speedup over other baselines, AdaShadow’s latency is only 75.3% of ElasticTrainer’s. Additionally, AdaShadow outperforms existing TTA methods in parallel inference/adaptation, achieving the highest adaptation accuracy with up to a 5.2% improvement over generic and efficient TTA baselines. MECTA offers comparable accuracy due to its adaptive BN layers. In label-free environments, AdaShadow surpasses Melon by 60.1%, effectively reducing overfitting through sparse updating. It also shows accuracy gains of 18.4% over TTE and 14.8% over ElasticTrainer, thanks to its backpropagation-free design and timely layer importance assessment. In contrast, TTE and ElasticTrainer’s reliance on backpropagation for layer importance evaluation results in significant performance drops when labels are unavailable.

We further compare AdaShadow with baselines under *corruption* shifts, another common in mobile applications due to sensor degradation. AdaShadow maintains the best accuracy-latency balance

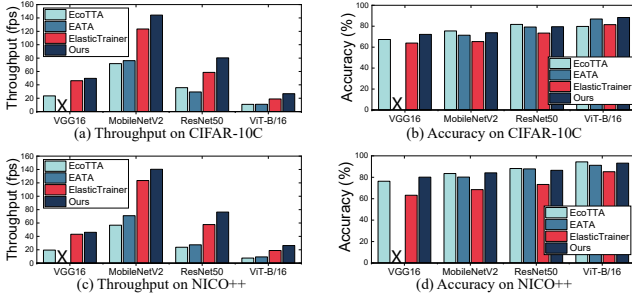
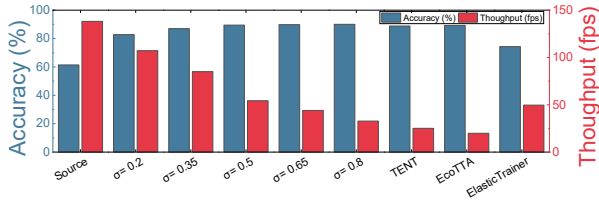
while adapting to corrupted CIFAR10 (see Tab. 5) and CIFAR100 (see Tab. 6) datasets on device D1. In CIFAR10-C shifts, AdaShadow outperforms EcoTTA, MECTA, and ElasticTrainer, achieving accuracy improvements of 2.9%, 0.6%, and 14.3%, along with accelerations of 3.3× 2.6× and 1.2×, respectively. For CIFAR100-C, AdaShadow surpasses these baselines with similar accuracy gains and accelerations. We also compare AdaShadow with backward-free TTA methods (BFTT3D, LAME, and OFTTA) on CIFAR10-C and CIFAR100-C (see Tab. 5). Although AdaShadow is on average 2.17 times slower, backward-free methods experience a 18.1%-25.4% accuracy drop after more than two continual shifts, which is unacceptable for mobile tasks. These methods perform adequately in single shift scenarios but lack adaptability during continual shifts and significant image changes because prohibiting backward and parameter updating limit the model’s optimization capacity[85]. OFTTA and BFTT3D address prior distribution shifts but provide minimal improvement for appearance shifts, while LAME performs worse than a non-adaptive model in scenarios with sensor degradation, such as *Shot*, *Gaus.*, and *Pixe.*

In summary, AdaShadow outperforms other baselines in adaptation latency and accuracy across diverse non-stationary shifts due to its efficient sparse updating with accurate online predictors for unit update importance and latency.

**5.2.2 Performance across Diverse Devices.** This experiment verifies AdaShadow’s effectiveness under different hardware and DVFS strategies by comparing to EATA, EcoTTA, and ElasticTrainer on three devices (*i.e.*,  $D_1 \sim D_3$ ) using NICO++ and CIFAR10-C shifted data. As shown in Tab. 7, AdaShadow achieves the lowest adaptation latency, being at most 32.7%, 29.3%, and 77.1% that of EATA,

**Table 7: Adaptation latency on diverse devices.**

Diverse devices	NICO++			CIFAR10-C		
	$D_1$	$D_2$	$D_3$	$D_1$	$D_2$	$D_3$
EATA [51]	39.5	318.5	1960.8	36.4	344.8	2325.6
EcoTTA [67]	44.1	322.6	2234.7	43.1	405.2	3031.8
ElasticTrainer [28]	16.4	144.9	831.6	16.2	136.6	1013.2
Ours	<b>12.9</b>	<b>113.6</b>	<b>769.2</b>	<b>12.5</b>	<b>121.9</b>	<b>909.1</b>

**Figure 9: Adaptation performance with diverse DNNs.****Figure 10: Impact of acceleration factor.**

EcoTTA, and ElasticTrainer, respectively. The latency difference between AdaShadow and ElasticTrainer is smaller on device  $D_3$  due to increased GPU core parallelization, increasing differences between runtime and offline latency, highlighting AdaShadow's advantage.

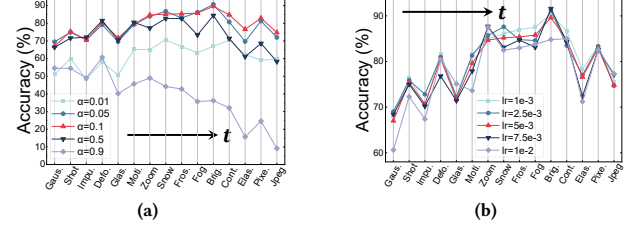
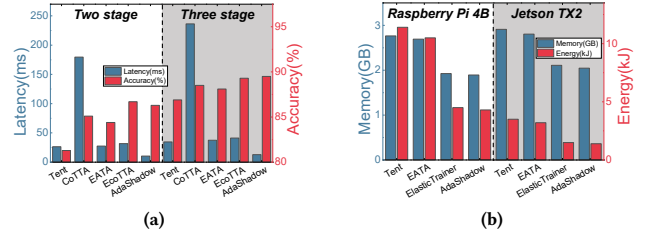
**5.2.3 Performance using Diverse Model Architectures.** We compare AdaShadow's adaptation accuracy and throughput with EATA, EcoTTA, and ElasticTrainer on ViT-B/16, VGG16, MobileNetV2, and ResNet50 using the NICO++ and CIFAR10-C datasets on device  $D_1$ . As shown in Fig. 9, AdaShadow achieves  $2\times \sim 3\times$  acceleration while maintaining high accuracy across all models. ElasticTrainer struggles with low accuracy in unsupervised conditions relying on backpropagation. MobileNetV2's simplified structure results in inferior learnability and adaptability, requiring more epochs to converge. EcoTTA and EATA fall short for lightweight DNNs due to fixed sparse updating. In contrast, AdaShadow flexibly selects layers to reduce accuracy gaps and employs a unit-based latency predictor to ensure an accuracy-latency balance, even on transformer architectures like ViT-B/16.

### 5.3 Micro-benchmarks

**5.3.1 Impact of Expected Acceleration Factor  $\sigma$ .** This experiment tests the impact of the expected acceleration factor  $\sigma$  in Equ.(2) on AdaShadow. From Fig. 10, AdaShadow achieves relatively balanced latency and accuracy when  $\sigma$  varies from 0.2 to 0.5. When  $\sigma$  is

**Table 8: Impact of batch sizes (BS).**

	BS=1	BS=2	BS=4	BS=16
Accuracy (%)	79.5	84.5	88.5	89.7
Throughput (fps)	19.2	43.2	79.6	340.2
Peak memory (MB)	2108	2210	2462	3794

**Figure 11: Impact of (a)  $\sigma$  and (b)  $lr$ .****Figure 12: Evaluation of (a) different adaptation modes and (b) system overhead.**

larger than 0.5, the accuracy saturates. When  $\sigma$  is lower than 0.2 AdaShadow shows notable accuracy loss. We empirically set  $\sigma = 0.33$  in AdaShadow.

**5.3.2 Impact of Adaptation Rate.** We test the AdaShadow's sensitivity to adaptation rate using CIFAR10-C and ResNet50, including integration rate  $\alpha$  and learning rate  $lr$ . AdaShadow shows robustness to  $lr$  variations (see Fig. 11b) but is sensitive to  $\alpha$ . A small  $\alpha$  leads to the model hardly learning new scenes, while a large  $\alpha$  causes the model to easily overfit to new scenes (see Fig. 11a). Therefore, we empirically set  $\alpha = 0.1$  in AdaShadow and  $lr=5e-3$  by default.

**5.3.3 Impact of Batch Sizes.** For stay responsive to non-stationary environments, effective adaptation on small batch sizes is expected. Tab. 8 shows the performance of AdaShadow with batch sizes of 1, 2, 4, and 16 with NICO++ on device  $D_1$ . Even with a batch size of 1, Adashadow keeps a high accuracy with 79.5%, supporting dynamic batch size adjustments to maintain responsiveness. As the batch size increases, throughput and accuracy improve at the cost of memory usage. With a batch size of 16, it surpasses the device's memory budget. Thus, we set *batch size* = 4 with 79.2fps throughput by default in AdaShadow.

**Table 9: Error of factors  $\pi_1$  and  $\pi_2$ .**

	Offline	$dr_1$	$dr_2$	$dr_3$	$dr_4$
$\pi_1$	1.0	1.6	4.3	1.0	7.0
$\pi_2$	1.0	1.0	1.0	2.4	2.4
Offline latency	45.8	45.8	45.8	45.8	45.8
Actual latency	45.3	73.5	187.2	50.5	309.9
Predicted latency	45.8	71.7	181.7	52.2	299.8
Error rate (%)	1.2	2.4	2.9	3.3	3.2

**Table 10: Performance comparison in sequential adaptation/inference mode.**

	Accuracy (%)	Throughput (fps)	Turnaround time
EATA [51]	70.8	35.3	2.11
EcoTTA [67]	71.4	22.8	2.42
ElasticTrainer [28]	60.7	70.6	1.47
AdaShadow	76.5	80.8	1.37

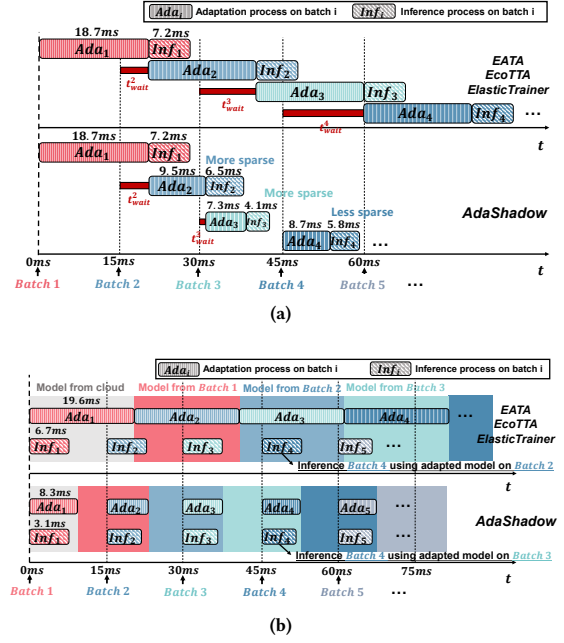
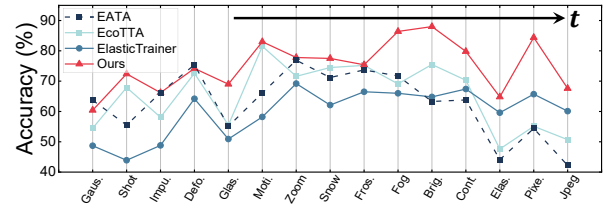
**5.3.4 Generalizing to Two-stage Adaptation Mode.** This experiment tests the AdaShadow’s efficiency of different adaptation modes using NICO++ and ResNet50. We compare the effectiveness of AdaShadow in three-stage (forward-backward-reforward) and two-stage (forward-backward) online TTA settings ( Tab. 12a). Results show that AdaShadow also aligns well with the traditional two-stage (forward-backward) mode, achieving significant acceleration by optimizing the sparse updating in the backward pass.

**5.3.5 Performance of Latency Predictor.** We test AdaShadow’s latency predictor under offline and  $dr_1 \sim dr_4$  conditions (the same settings mentioned in ) using 15,000 self-collected records from different systems, devices, and models. We run and measure the network’s *offline latency* in a stable environment (25°C, with no competing processes for computation or memory). In the dynamic online environment, after predicting execution latency, the network is executed to collect *actual latency* immediately. Considering dynamic kernel and memory resources, the average prediction error between AdaShadow’s latency predictor and ground truth under  $dr_1 \sim dr_4$  are  $\leq 3.3\%$  (see Tab. 9), showing high accuracy and stability.

**5.3.6 System Overhead.** AdaShadow can efficiently tune memory usage by selective updates of crucial layers in a dynamic adaptive manner. In Table 12b, AdaShadow achieves a peak memory usage of  $\leq 2045MB$ , lower than Tent, EATA, and ElasticTrainer. Moreover, AdaShadow significantly decreases the frequency and duration of backpropagation, reducing 56% and 60% energy cost than Tent and EATA.

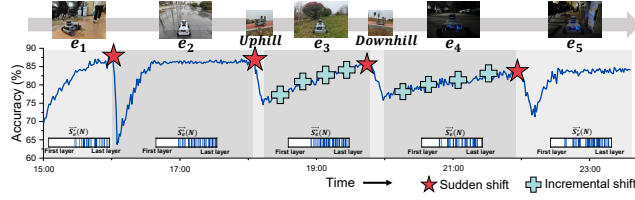
## 5.4 Case Studies

**5.4.1 Sequential Adaptation/Inference Mode.** This study showcases the usage of AdaShadow, where it first adapts the model upon a new batch of data before inference on the same batch using the updated model (see Fig. 13a). That is, the inference on *batch1* is after the adaptation on *batch1*, and the adaption on *batch2* also starts after the adaptation on *batch1*. This execution mode slightly prioritizes accuracy over latency and is widely used in applications such as AR and VR. We set the time window to 15ms, the batch size

**Figure 13: Illustration of (a) sequential and (b) parallel adaptation/inference mode.****Figure 14: Accuracy comparison in parallel adaptation/inference mode.**

to 4, and test with continuously played CIFAR10-C on device  $D_1$ . As shown in Fig. 13a, this execution mode introduces waiting time, making latency a potential bottleneck. We quantify such waiting time via the *turnaround time* for batch  $i$ , i.e.,  $r^i = \frac{T_{wait}^i + T_a^i + T_{re}^i}{T_a^i + T_{re}^i}$ , a common metric in OS process scheduling, where  $T_{wait}^i$  is the wait time of batch  $i$ . We then report the average  $r$  across all arriving data batches as the total waiting time. Since AdaShadow dynamically adjusts the acceleration factor  $\sigma$ , it can gradually decrease  $r$ , where AdaShadow scales  $\sigma$  to tune  $t_{wait}^2$ ,  $t_{wait}^3$ , and  $t_{wait}^4$  of subsequent arriving batch. From Tab. 10, AdaShadow achieves higher inference accuracy with a lower  $r$  than the baselines.

**5.4.2 Parallel Adaptation/Inference Mode.** This study demonstrates the usage of AdaShadow, where when a new batch of data arrives, it directly utilizes the current model for inference without waiting for the newly adapted one (see Fig. 13b). This execution mode is preferred in real-time applications like drone object detection and autonomous driving. We set the time window to 15ms, the



**Figure 15: Case study illustration.**  $\tilde{S}_e^*(N)$  is the corresponding optimal sparse updating strategy.

batch size to 4, and test with CIFAR10-C on device  $D_1$ . From Fig. 14, AdaShadow maintains the highest accuracy for its rapid and accurate model adaptation. In contrast, other schemes fail to update model in time, leading to inaccurate inference using outdated models.

**5.4.3 Use Case.** This study evaluates AdaShadow in real-world mobile environments. We built a vehicle platform to continuously collect video streams in a university campus in varying weather, illumination, and locations. We performed test-time DNN adaptation and inference during vehicle driving. We summarized the environments encountered over time as: bright indoor ( $e_1$ )  $\rightarrow$  rainy road ( $e_2$ )  $\rightarrow$  uphill  $\rightarrow$  rainy grass ( $e_3$ )  $\rightarrow$  downhill  $\rightarrow$  night road ( $e_4$ )  $\rightarrow$  dark indoor ( $e_5$ ). From Fig. 15, AdaShadow can promptly adapt to all these natural shifts with high accuracy and low latency by updating diverse layers. The blue bars below represent the layers that are selected for updating.

## 6 RELATED WORK

**Test-time Adaptation.** Test-time adaptation (TTA) is an emerging *domain adaptation* setting with *unlabeled* target data and assuming no access to any source data or supervision [73]. Generic domain adaptation techniques [7, 15, 23] are inherently supervised. More recent label-free domain adaptation schemes [71, 82] explored unsupervised adaption, yet still require assistance from the source data to assess and align data distributions. We refer readers to [38] for a comprehensive review on domain adaptation.

Earlier TTA efforts focus on improving the adaptation *accuracy* in this unsupervised, source-free setup. Classical methods [50, 61, 73] adapt the *batch normalization* (BN) layers to the domain shifts by *minimizing the entropy of the last layer’s outputs on the training and testing data*. Recent studies [6, 67, 77] show that updating the BN layers alone might be insufficient for adaption to more drastic shifts and when the batch size is small and propose to also update parameters in the convolutional layers [6, 77] or adapters [67]. Since we aim at TTA with small batches, we also update all layers rather than restrict to BN layers. Furthermore, we refine the adaption loss by incorporating outputs from all layers to enhance the robustness to small batches.

A few pioneer studies explore improving the *efficiency* of TTA. EATA [51] focuses on data efficiency by only performing adaptation on important test samples. EcoTTA [67] adopts lightweight meta-networks attached to the original model for memory-efficient adaptation. MECTA [24] improves memory efficiency via an adaptive BN layer. Yet, decreased computation or memory does not easily translate into reduced latency. Additionally, *retraining-free*

TTA, including prototype-based approaches [79, 80] and consistency regularization [4], enhance efficiency by eliminating the need for backpropagation. However, these approaches do not update model parameters, limiting their ability to refine representations. This limitation leads to dramatic accuracy losses in mobile environments with continuous changes or drastic shifts [85] (evaluated in § 5.2.1), thereby motivating our work.

**On-device DNN Training.** Resource-efficient DNN training on mobile devices has attracted increasing research interest. Common techniques include recomputation [9, 17, 53, 78], micro-batch [29, 45, 66], quantization [58, 83, 91], sparse updating [28, 49, 86], and memory swapping [8, 27, 59, 75]. While memory swapping, recomputation, and micro-batch techniques can reduce memory usage, they cannot decrease or even increase training latency because of extra computations [78] or data transfers [59]. In contrast, sparse updating [28, 40, 86] often results in decreased latency as it reduces computations during backpropagation.

Sparse updating can be *static* [25, 49, 86], which updates fixed crucial DNN modules, or *dynamic* [24, 28, 81], where the DNN modules to be updated are adjusted at runtime. A static strategy is unfit for TTA since the modules critical for one domain may be less important for another. Our work is most relevant to ElasticTrainer [28], which also adopts a dynamic sparse updating strategy to accelerate DNN training. However, ElasticTrainer [28] assumes labelled data, making it not directly applicable to TTA. More importantly, it is primarily designed for offline training, where the training (adaptation) and inference are tightly coupled in TTA. This coupling imposes stringent latency budget for layer importance assessment and necessitates awareness to resource dynamics for layer latency profiling, making ElasticTrainer [28] sub-optimal. In contrast, AdaShadow achieves accurate and low-latency adaptation in mobile environments with fast-changing unlabelled data and highly dynamic resources.

## 7 CONCLUSIONS

This paper presents AdaShadow, a responsive test-time adaptation framework tailored for non-stationary mobile environments. It is the first endeavor toward near-real-time on-device DNN adaptation without source data or supervision. It accurately estimates layer importance in the forward pass without backpropagation, calibrates the offline layer latency measurements to runtime resource dynamics, and efficiently schedules optimal layer update strategy, overcoming TTA’s latency bottleneck in mobile context without compromising accuracy. Evaluations across diverse real-world scenarios and mobile devices show that AdaShadow achieves significant adaptation latency reductions and accuracy improvements. We plan to extend AdaShadow to mobile applications with stricter latency requirements and more complex DNNs. Integrating AdaShadow with federated learning, such as federated TTA, could enhance model robustness and user privacy.

## ACKNOWLEDGMENTS

This work was partially supported by the National Science Fund for Distinguished Young Scholars (62025205), the National Natural Science Foundation of China (No. 62032020, 6247074224, 62102317), and CityU APRC grant (No. 9610633).



## REFERENCES

- [1] Google ARCore. 2019. <https://developers.google.com/ar>.
- [2] Apple ARKit. 2019. <https://developer.apple.com/augmented-reality/arkit/>.
- [3] Taghreed Balharith and Fahd Alhaidari. 2019. Round robin scheduling algorithm in CPU and cloud computing: a review. In *Proceedings of the IEEE ICCAIS*. 1–7.
- [4] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. 2022. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF CVPR*. 8344–8353.
- [5] Chao Chen, Zhihang Fu, Zhihong Chen, Sheng Jin, Zhaowei Cheng, Xinyu Jin, and Xian-Sheng Hua. 2020. Homm: Higher-order moment matching for unsupervised domain adaptation. In *Proceedings of the AAAI*. 3422–3429.
- [6] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. 2022. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF CVPR*. 295–305.
- [7] Minghao Chen, Hongyang Xue, and Deng Cai. 2019. Domain adaptation for semantic segmentation with maximum squares loss. In *Proceedings of the IEEE/CVF ICCV*. 2090–2099.
- [8] Ping Chen, Shuibing He, Xuechen Zhang, Shuaibin Chen, Peiyi Hong, Yanlong Yin, Xian-He Sun, and Gang Chen. 2021. CSWAP: A self-tuning compression framework for accelerating tensor swapping in GPUs. In *Proceedings of the IEEE CLUSTER*. 271–282.
- [9] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174* (2016).
- [10] Haolin Chu, Xiaolong Zheng, Liang Liu, and Huadong Ma. 2023. nnPerf: Demystifying DNN Runtime Inference Latency on Mobile Platforms. In *Proceedings of the ACM SenSys*. 125–137.
- [11] Matthew Corbett, Brendan David-John, Jiacheng Shang, Y Charlie Hu, and Bo Ji. 2023. Bystander: Protecting bystander visual data in augmented reality systems. In *Proceedings of the ACM MobiSys*. 370–382.
- [12] Kaikai Deng, Dong Zhao, Qiaoyue Han, Shuyue Wang, Zihan Zhang, Anfu Zhou, and Huadong Ma. 2022. Geryon: Edge assisted real-time and robust object detection on drones via mmWave radar and camera fusion. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–27.
- [13] Chuntao Ding, Zhichao Lu, Felix Juefei-Xu, Vishnu Naresh Boddeti, Yidong Li, and Jiannong Cao. 2022. Towards transmission-friendly and robust cnn models over cloud and device. *IEEE Transactions on Mobile Computing* 22, 10 (2022), 6176–6189.
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the ICLR*.
- [15] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the ACM ICML*. 1180–1189.
- [16] Soumendu Kumar Ghosh, Arnab Raha, and Vijay Raghunathan. 2023. Energy-efficient approximate edge inference systems. *ACM Transactions on Embedded Computing Systems* 22, 4 (2023), 1–50.
- [17] In Gim and JeongGil Ko. 2022. Memory-efficient DNN training on mobile devices. In *Proceedings of the ACM MobiSys*. 464–476.
- [18] Shurui Gui, Meng Liu, Xiner Li, Youzhi Luo, and Shuiwang Ji. 2024. Joint learning of label and environment causal independence for graph out-of-distribution generalization. *Advances in Neural Information Processing Systems* 36 (2024).
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF CVPR*. 770–778.
- [20] Yuze He, Chen Bian, Jingfei Xia, Shuyao Shi, Zhenyu Yan, Qun Song, and Guoliang Xing. 2023. VI-Map: Infrastructure-Assisted Real-Time HD Mapping for Autonomous Driving. In *Proceedings of the ACM MobiCom*. 1–15.
- [21] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the ICLR*.
- [22] Sebastian Herbert and Diana Marculescu. 2007. Analysis of dynamic voltage/frequency scaling in chip-multiprocessors. In *Proceedings of the ISLPED*. 38–43.
- [23] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *Proceedings of the ACM ICML*. 1989–1998.
- [24] Junyuan Hong, Lingjuan Lyu, Jiayu Zhou, and Michael Spranger. 2023. Mecta: Memory-economic continual test-time model adaptation. In *Proceedings of the ICLR*.
- [25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the ICLR*.
- [26] Jiagao Hu, Zhengxing Sun, Bo Li, Kewei Yang, and Dongyang Li. 2017. Online user modeling for interactive streaming image classification. In *Proceedings of the MMM*. Springer, 293–305.
- [27] Chien-Chin Huang, Gu Jin, and Jinyang Li. 2020. Swapadvisor: Pushing deep learning beyond the gpu memory limit via smart swapping. In *Proceedings of ASPLOS*. 1341–1355.
- [28] Kai Huang, Boyuan Yang, and Wei Gao. 2023. ElasticTrainer: Speeding Up On-Device Training with Runtime Elastic Tensor Selection. In *Proceedings of the ACM MobiSys*. 56–69.
- [29] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, et al. 2019. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In *Proceedings of the NeurIPS*. Vol. 32.
- [30] Puneet Jain, Justin Manweiler, and Romit Roy Choudhury. 2015. Overlay: Practical mobile augmented reality. In *Proceedings of the ACM MobiSys*. 331–344.
- [31] Venkatesh Kodukula, Mason Manetta, and Robert LiKamWa. 2023. Squint: A Framework for Dynamic Voltage Scaling of Image Sensors Towards Low Power IoT Vision. In *Proceedings of the ACM MobiCom*. 1–15.
- [32] Rui Kong, Yuanchun Li, Yizhen Yuan, and Linghe Kong. 2023. Convrelu++: Reference-based lossless acceleration of conv-relu operations on mobile cpu. In *Proceedings of the ACM MobiSys*. 503–515.
- [33] Z Jonny Kong, Qiang Xu, Jiayi Meng, and Y Charlie Hu. 2023. AccuMO: Accuracy-Centric Multitask Offloading in Edge-Assisted Mobile Augmented Reality. In *Proceedings of the ACM MobiCom*. 1–16.
- [34] Jaeeun Lee, Raphael Tang, and Jimmy Lin. 2019. What would elsa do? freezing layers during transformer fine-tuning. *arXiv preprint arXiv:1911.03090* (2019).
- [35] Yoonho Lee, Annie S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. 2023. Surgical fine-tuning improves adaptation to distribution shifts. In *Proceedings of the ICLR*.
- [36] Jingyu Li, Fengling Jiang, Jing Yang, Bin Kong, Mandar Gogate, Kia Dashtipour, and Amir Hussain. 2021. Lane-deeplab: Lane semantic segmentation in automatic driving scenarios for high-definition maps. *Neurocomputing* 465 (2021), 15–25.
- [37] Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, et al. 2020. Pytorch distributed: Experiences on accelerating data parallel training. In *Proceedings of the VLDB*. 3005–3018.
- [38] Jian Liang, Ran He, and Tieniu Tan. 2023. A comprehensive survey on test-time adaptation under distribution shifts. *arXiv preprint arXiv:2303.15361* (2023).
- [39] Chengdong Lin, Kun Wang, Zhenjiang Li, and Yu Pu. 2023. A Workload-Aware DVFS Robust to Concurrent Tasks for Mobile Devices. In *Proceedings of the ACM MobiCom*. 1–16.
- [40] Ji Lin, Ligeng Zhu, Wei-Ming Chen, Wei-Chen Wang, Chuang Gan, and Song Han. 2022. On-device training under 256kb memory. In *Proceedings of the NeurIPS*.
- [41] Liangkai Liu, Sidi Lu, Ren Zhong, Baofu Wu, Yongtao Yao, Qingyang Zhang, and Weisong Shi. 2020. Computing systems for autonomous driving: State of the art and challenges. *IEEE Internet of Things Journal* 8, 8 (2020), 6469–6486.
- [42] Sicong Liu, Bin Guo, Cheng Fang, Ziqi Wang, Shiyao Luo, Zimu Zhou, and Zhiwen Yu. 2023. Enabling Resource-Efficient AIoT System With Cross-Level Optimization: A Survey. *IEEE Communications Surveys & Tutorials* (2023).
- [43] Sicong Liu, Bin Guo, Ke Ma, Zhiwen Yu, and Junzhao Du. 2021. AdaSpring: Context-adaptive and runtime-evolutionary deep model compression for mobile applications. *Proceedings of the ACM on IMWUT* 5, 1 (2021), 1–22.
- [44] Sicong Liu, Xiaochen Li, Zimu Zhou, Bin Guo, Meng Zhang, Haocheng Shen, and Zhiwen Yu. 2023. AdaEnlight: Energy-aware low-light video stream enhancement on mobile devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–26.
- [45] Weijie Liu, Shiquan Lai, Shengwei Li, Yabo Duan, Keshi Ge, and Dongsheng Li. 2022. AutoPipe: A fast pipeline parallelism approach with balanced partitioning and micro-batch slicing. In *Proceedings of the IEEE CLUSTER*. IEEE, 301–312.
- [46] Shengyu Lu, Beizhan Wang, Hongji Wang, Lihao Chen, Ma Linjian, and Xiaoyan Zhang. 2019. A real-time object detection algorithm for video. *Computers & Electrical Engineering* 77 (2019), 398–408.
- [47] M Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. 2022. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *Proceedings of the IEEE/CVF CVPR*. 14765–14775.
- [48] Ji Joong Moon, Hyun Suk Lee, Jiho Chu, Donghak Park, Seungbaek Hong, Hyungjun Seo, Donghyeon Jeong, Sungsik Kong, and MyungJoo Ham. 2024. A New Frontier of AI: On-Device AI Training and Personalization. In *Proceedings of the ICSE*.
- [49] Pramod Kaushik Mudrakarta, Mark Sandler, Andrey Zhmoginov, and Andrew Howard. 2019. K for the price of 1: Parameter-efficient multi-task and transfer learning. In *Proceedings of the ICLR*.
- [50] 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. In *Proceedings of the ACM ICML Workshops*.
- [51] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yafo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. 2022. Efficient test-time model adaptation without forgetting. In *Proceedings of the ACM ICML*. 16888–16905.
- [52] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiqian Wen, Yafo Chen, Peilin Zhao, and Mingkui Tan. 2023. Towards stable test-time adaptation in dynamic wild world. *Proceedings of the ICLR*.
- [53] Shishir G Patil, Paras Jain, Prabal Dutta, Ion Stoica, and Joseph Gonzalez. 2022. Poet: Training neural networks on tiny devices with integrated rematerialization

- and paging. In *Proceedings of the ACM ICML*. 17573–17583.
- [54] Ashutosh Pattnaik, Xulong Tang, Adwait Jog, Onur Kayiran, Asit K Mishra, Mahmut T Kandemir, Onur Mutlu, and Chita R Das. 2016. Scheduling techniques for GPU architectures with processing-in-memory capabilities. In *Proceedings of the PACT*. 31–44.
- [55] Stefano Petrangeli, Gwendal Simon, Haoliang Wang, and Vishy Swaminathan. 2019. Dynamic adaptive streaming for augmented reality applications. In *Proceedings of the ISM*. IEEE, 56–567.
- [56] Protecting data privacy using Microsoft Azure. 2021.
- [57] Zhongnan Qu, Zimu Zhou, Yongxin Tong, and Lothar Thiele. 2022. p-meta: Towards on-device deep model adaptation. In *Proceedings of the ACM KDD*. 1441–1451.
- [58] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. 2016. Xnor-net: Imagenet classification using binary convolutional neural networks. In *Proceedings of the ECCV*. 525–542.
- [59] Minsoo Rhu, Natalia Gimelshein, Jason Clemons, Arslan Zulfiqar, and Stephen W Keckler. 2016. vDNN: Virtualized deep neural networks for scalable, memory-efficient neural network design. In *Proceedings of the MICRO*. 1–13.
- [60] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF CVPR*. 4510–4520.
- [61] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. 2020. Improving robustness against common corruptions by covariate shift adaptation. *Proceedings of the NeurIPS* 33, 11539–11551.
- [62] Haihong She, Yigui Luo, Zhaohong Xiang, Weiming Liang, and Yin Xie. 2023. Accurate Latency Prediction of Deep Learning Model Inference Under Dynamic Runtime Resource. In *Proceedings of the ICONIP*. Springer, 495–510.
- [63] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2018. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI*, Vol. 32.
- [64] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the ICLR*.
- [65] Ajit Singh, Priyanka Goyal, and Sahil Batra. 2010. An optimized round robin scheduling algorithm for CPU scheduling. *International Journal on Computer Science and Engineering* 2, 07 (2010), 2383–2385.
- [66] Nimit S Sohoni, Christopher R Aberger, Megan Leszczynski, Jian Zhang, and Christopher Ré. 2019. Low-memory neural network training: A technical report. *arXiv preprint arXiv:1904.10631* (2019).
- [67] Junha Song, Jungsoo Lee, In So Kweon, and Sungha Choi. 2023. EcoTTA: Memory-Efficient Continual Test-time Adaptation via Self-distilled Regularization. In *Proceedings of the IEEE/CVF CVPR*. 11920–11929.
- [68] Meta Spark. 2024. <https://spark.meta.com>.
- [69] Sweden Data Collection & Processing. 2024.
- [70] Xiaohu Tang, Yang Wang, Ting Cao, Li Lyna Zhang, Qi Chen, Deng Cai, Yunxin Liu, and Mao Yang. 2023. Lut-nn: Empower efficient neural network inference with centroid learning and table lookup. In *Proceedings of the ACM MobiCom*. 1–15.
- [71] Abu Md Niamul Taufique, Chowdhury Sadman Jahan, and Andreas Savakis. 2022. Unsupervised continual learning for gradually varying domains. In *Proceedings of the IEEE/CVF CVPR*. 3740–3750.
- [72] Uri Verner, Assaf Schuster, Mark Silberstein, and Avi Mendelson. 2012. Scheduling processing of real-time data streams on heterogeneous multi-GPU systems. In *Proceedings of the ACM SYSTOR*. 1–12.
- [73] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. Tent: Fully test-time adaptation by entropy minimization. In *Proceedings of the ICLR*.
- [74] Jiadai Wang, Jiajia Liu, and Nei Kato. 2018. Networking and communications in autonomous driving: A survey. *IEEE Communications Surveys & Tutorials* 21, 2 (2018), 1243–1274.
- [75] Linnan Wang, Jinmian Ye, Yiyang Zhao, Wei Wu, Ang Li, Shuaiwen Leon Song, Zenglin Xu, and Tim Kraska. 2018. Superneurons: Dynamic GPU memory management for training deep neural networks. In *Proceedings of the PPOPP*. 41–53.
- [76] Lehao Wang, Zhiwen Yu, Haoyi Yu, Sicong Liu, Yaxiong Xie, Bin Guo, and Yunxin Liu. 2023. AdaEvo: Edge-Assisted Continuous and Timely DNN Model Evolution for Mobile Devices. *IEEE Transactions on Mobile Computing* (2023).
- [77] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. 2022. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF CVPR*. 7201–7211.
- [78] Qipeng Wang, Mengwei Xu, Chao Jin, Xinran Dong, Jinliang Yuan, Xin Jin, Gang Huang, Yunxin Liu, and Xuanzhe Liu. 2022. Melon: Breaking the memory wall for resource-efficient on-device machine learning. In *Proceedings of the ACM MobiSys*. 450–463.
- [79] Shuoyuan Wang, Jindong Wang, Huajun Xi, Bob Zhang, Lei Zhang, and Hongxin Wei. 2024. Optimization-Free Test-Time Adaptation for Cross-Person Activity Recognition. *Proceedings of the ACM on IMWUT* 7, 4 (2024), 1–27.
- [80] Yanshuo Wang, Ali Cheraghian, Zeeshan Hayder, Jie Hong, Sameera Ramasinghe, Shafin Rahman, David Ahméd-Aristizabal, Xuesong Li, Lars Petersson, and Mehrtash Harandi. 2024. Backpropagation-free Network for 3D Test-time Adaptation. In *Proceedings of the IEEE/CVF CVPR*. 23231–23241.
- [81] Yue Wang, Ziyu Jiang, Xiaohan Chen, Pengfei Xu, Yang Zhao, Yingyan Lin, and Zhangyang Wang. 2019. E2-train: Training state-of-the-art cnns with over 80% energy savings. In *Proceedings of the NeurIPS*, Vol. 32.
- [82] Markus Wulfmeier, Alex Bewley, and Ingmar Posner. 2018. Incremental adversarial domain adaptation for continually changing environments. In *Proceedings of the ICRA*. 4489–4495.
- [83] Daliang Xu, Mengwei Xu, Qipeng Wang, Shangguang Wang, Yun Ma, Kang Huang, Gang Huang, Xin Jin, and Xuanzhe Liu. 2022. Mandheling: Mixed-precision on-device dnn training with dsp offloading. In *Proceedings of the ACM MobiCom*. 214–227.
- [84] Fang Xu, Tianyu Zhou, Hengxu You, and Jing Du. 2024. Improving indoor wayfinding with AR-enabled egocentric cues: A comparative study. *Advanced Engineering Informatics* 59 (2024), 102265.
- [85] Yongcan Yu, Lijun Sheng, Ran He, and Jian Liang. 2023. Benchmarking test-time adaptation against distribution shifts in image classification. *arXiv preprint arXiv:2307.03133* (2023).
- [86] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2022. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *Proceedings of the ACL* (2022).
- [87] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschlager, and Susanne Saminger-Platz. 2017. Central moment discrepancy (cmd) for domain-invariant representation learning. In *Proceedings of the ICLR*.
- [88] Li Lyna Zhang, Shihao Han, Jianyu Wei, Ningxin Zheng, Ting Cao, Yuqing Yang, and Yunxin Liu. 2021. Nn-meter: Towards accurate latency prediction of deep-learning model inference on diverse edge devices. In *Proceedings of the ACM MobiSys*. 81–93.
- [89] Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyang Shen, and Peng Cui. 2023. Nico++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF CVPR*. 16036–16047.
- [90] Xingxuan Zhang, Renzhe Xu, Han Yu, Yancheng Dong, Pengfei Tian, and Peng Cui. 2023. Flatness-aware minimization for domain generalization. In *Proceedings of the IEEE/CVF ICCV*. 5189–5202.
- [91] Qihua Zhou, Song Guo, Zhihao Qu, Jingcai Guo, Zhenda Xu, Jiewei Zhang, Tao Guo, Boyuan Luo, and Jingren Zhou. 2021. Octo: INT8 Training with Loss-aware Compensation and Backward Quantization for Tiny On-device Learning. In *Proceedings of the USENIX ATC*. 177–191.