AR-TTA: A Simple Method for Real-World Continual Test-Time Adaptation

Damian Sójka [1,2] Sebastian Cygert [1,3] Bartłomiej Twardowski [1,4,5] Tomasz Trzciński [1,6,7]

[1]IDEAS NCBR [2]Poznań University of Technology [3]Gdańsk University of Technology [4]Autonomous University of Barcelona
[5]Computer Vision Center [6]Warsaw University of Technology [7]Tooploox

ICCV23 PARIS

Introduction

- Current **machine learning models perform very well** only when **the test-time distribution is close to the training-time distribution**.
- **Continual Test-Time Adaptation (TTA) methods** allow the source model to **adapt** itself on-the-fly to **continual changes in data distribution** without any supervision.
- **Current techniques** are usually **evaluated on** benchmarks that are only **a simplification of real-world scenarios**.
- We observe that **current test-time adaptation methods struggle** to effectively handle **varying degrees of domain shift**, often resulting in **degraded performance that falls below that of the source model**.

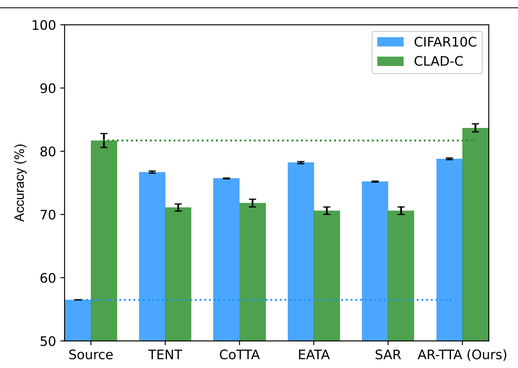


Figure 1. Continual test-time adaptation methods evaluated on synthetic (CIFAR-10C) and realistic (CLAD-C) domain shifts. Our method is the only one that consistently allows to improve over the naive strategy of using the (frozen) source model.

Contribution

Our main contributions can be summarized as follows:

- We **evaluate and analyze** current test-time adaptation methods **on realistic, continual domain shift image classification data from autonomous driving**.
- We **propose** a simple **continual TTA method**.
- Extensive evaluation shows that the **proposed method obtains state-of-the-art performance** on multiple benchmarks with both **artificial distortions** and **real-life domain shifts**.

Natural domain shifts

Figure 2. Example images from various domains within the CLAD-C benchmark.

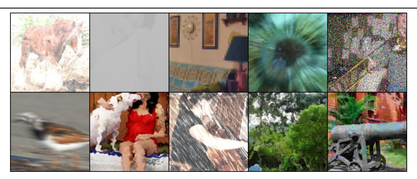
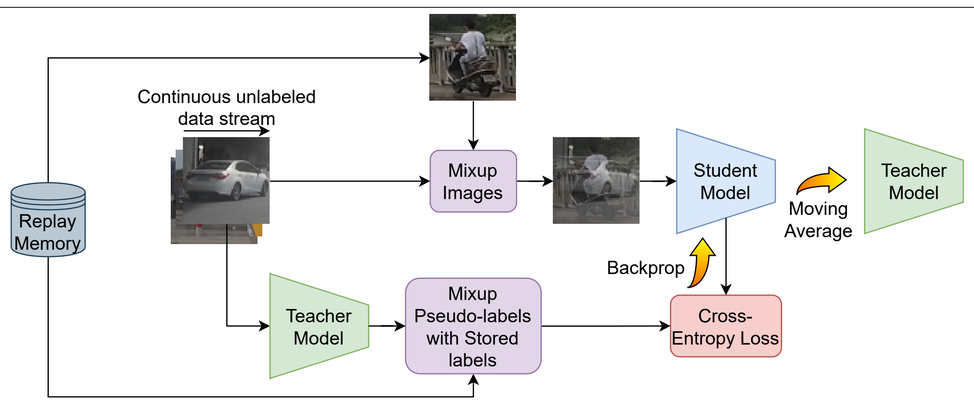


Figure 3. Example images with different corruptions from the ImageNet-C dataset.

- The most popular setting for test-time adaptation includes using different classes of **synthetic corruptions**.
- In **practical applications**, the **target distribution** can easily change in **different manner**, perpetually over time, e.g., due to changing weather, lighting conditions, or traffic intensity.
- Hence, we **propose to use two benchmarks** that consist of data with domain shifts that can occur in **real-world applications** - the CLAD-C benchmark [5] and the SHIFT dataset [4].

Proposed method (AR-TTA)



Our proposed approach to AR-TTA can be divided into **three parts**:

- **Update procedure:** Inspired by [7], we keep two identical neural network models with different sets of weights: the **teacher model** and the **student model**. The **student model** is **updated** based on **cross-entropy loss** between its **predictions and pseudo-labels** generated by the teacher model. The **teacher model** is **adapted** based on an **exponential moving average of student's weights**. Predictions for each image are taken from the teacher model.
- **Experience replay:** We store a small **class-balanced buffer of random exemplars** from the labeled **source data** in the memory. We sample one exemplar from memory for each test image and **mixup image pairs**. Similarly, **pseudo-labels** from the teacher model are **mixupped** with the **labels of sampled exemplars**. The student model is trained on augmented test samples and augmented pseudo-labels.
- **Dynamic Barch Normalization Statistics:** To robustly **estimate** the correct **Batch Normalization (BN) statistics** we take the inspiration from [1] and propose to estimate BN statistics during test-time by **linearly interpolating between saved statistics of source data** and **statistics of the current batch**. The parameter that weights the influence of two sets of statistics is adjusted by the **exponential moving average**. Its value at the current batch is calculated using the **distance between the distribution of the current batch and the distribution used for the previous batch**. As a distance metric, we utilize **symmetric KL divergence**.

Results on artificial domain shifts

Table 1. Classification accuracy (%) for the standard ImageNet-to-ImageNetC and CIFAR10-to-CIFAR10C online continual test-time adaptation tasks.

Method	Mean	
	CIFAR10C	ImageNetC
Source	56.5	18.1
BN stats adapt	75.0	26.9
TENT-continual [6]	76.7	29.2
EATA [2]	78.2	31.5
COTTA [7]	75.7	15.5
SAR [3]	75.2	30.8
Ours (AR-TTA) w/o replay	$77.3_{\pm 0.07}$	$30.0_{\pm 0.45}$
Ours (AR-TTA)	$\mathbf{78.8}_{\pm 0.13}$	$\mathbf{32.0}_{\pm 0.07}$

- **Artificial domain shifts** pose a **great challenge** for source model.
- **Discarding BN statistics** calculated on the source training data and **estimating them for each batch separately**, already significantly **improves** the result on **corrupted images** (BN stats adapt method).
- Each of the compared **state-of-the-art TTA methods uses the BN stats adapt technique**, therefore they are able to **improve over it**, but the increase in accuracy value is not that significant.
- Our method **AR-TTA outperforms** all of the compared techniques.

Results on natural domain shifts

Table 2. Classification accuracy and average mean class accuracy (AMCA) (%) for the CLAD-C continual test-time adaptation task.

Method	T1 ($t \longrightarrow$)	T2	T3	T4	T5	Mean day	Mean night	Mean	AMCA
Source	75.6	85.9	73.3	87.5	66.2	86.6	71.2	81.3	57.6
BN stats adapt	73.2	69.9	75.0	75.5	59.7	72.2	69.1	71.1	48.3
TENT-continual [6]	73.4	69.8	76.5	76.1	59.7	72.4	69.8	71.5	47.6
EATA [2]	73.3	69.9	75.0	75.6	59.7	72.2	69.1	71.1	48.4
CoTTA [7]	75.2	69.3	80.2	77.0	62.7	72.4	72.9	72.6	44.8
SAR [3]	73.2	69.9	75.0	75.5	59.7	72.2	69.1	71.1	48.3
Ours (AR-TTA) w/o replay	76.9	**86.7**	**81.4**	87.9	**73.5**	87.2	**77.1**	$\mathbf{83.9}_{\pm 0.30}$	$59.6_{\pm 2.92}$
Ours (AR-TTA)	**77.2**	**86.7**	80.0	**89.6**	70.7	**87.8**	75.7	$83.7_{\pm 0.64}$	$\mathbf{63.1}_{\pm 3.32}$

Table 3. Classification accuracy and average mean class accuracy (AMCA) (%) for the SHIFT-C continual test-time adaptation task.

Method	daytime				dawn/dusk					night					Mean	AMCA
	cloudy	*overcast*	*rainy*	*foggy*	*clear*	*cloudy*	*overcast*	*rainy*	*foggy*	*clear*	*cloudy*	*overcast*	*rainy*	*foggy*		
Source	**97.9**	**98.2**	**97.5**	92.5	93.6	94.1	94.0	93.5	91.5	89.1	89.3	90.6	89.1	90.7	93.5	89.5
BN stats adapt	89.1	88.9	88.0	86.2	85.3	84.8	87.3	83.5	84.8	81.3	81.2	80.3	79.6	83.5	85.1	69.9
TENT-continual [6]	89.6	88.8	87.5	84.6	83.3	81.2	85.0	80.7	80.2	78.0	77.0	76.1	75.7	77.6	82.7	57.6
EATA [2]	89.1	88.9	88.0	86.2	85.3	84.8	87.4	83.6	84.9	81.4	81.4	80.3	79.7	83.7	85.1	70.5
CoTTA [7]	88.2	87.1	84.1	80.5	78.7	76.2	80.5	74.0	74.9	71.5	70.3	67.3	64.9	66.2	77.4	47.2
SAR [3]	89.1	88.9	88.0	86.2	85.3	84.8	87.3	83.5	84.8	81.3	81.2	80.3	79.6	83.6	85.1	69.9
Ours (AR-TTA) w/o replay	96.4	96.5	95.3	93.2	92.2	91.9	93.2	91.4	91.8	88.7	88.7	88.6	87.5	91.2	$92.4_{\pm 0.25}$	$83.5_{\pm 0.96}$
Ours (AR-TTA)	97.7	98.0	97.4	**94.3**	**94.2**	**95.5**	**94.8**	**95.2**	**93.1**	**92.3**	**92.7**	**93.0**	**91.4**	**92.6**	$\mathbf{94.8}_{\pm 0.03}$	$\mathbf{90.2}_{\pm 0.24}$

- Calculating **BN statistics for each batch** (BN stats adapt method) does **not improve** the performance over the frozen source model on **natural domain shifts**.
- Similarly, the **state-of-the-art TTA methods** achieve significantly **lower mean accuracy**, compared to the **frozen source model**, rendering them not effective for natural domain shifts.
- **Our method**, which uses pre-calculated statistics and exemplars of source data during adaptation, **outperformed state-of-the-art methods** and achieves higher accuracy than the source model, which shows the **effectiveness and adapting capabilities**.
- **Average mean class accuracy (AMCA)** values show that the usage of **replay memory** might be **crucial** for high mean **per-class accuracy**.

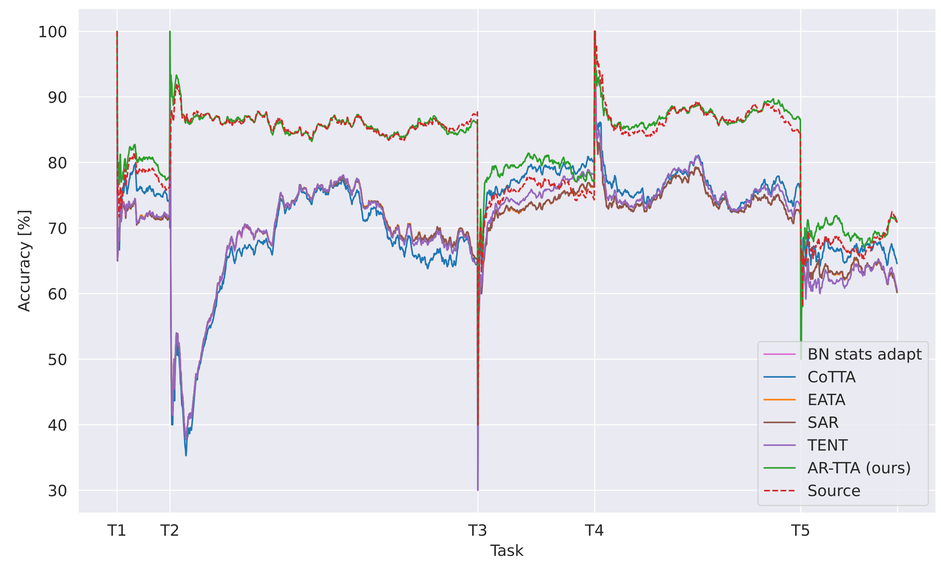


Figure 4. Batch-wise classification accuracy averaged in a window of 100 batches on CLAD-C benchmark for the chosen methods continually adapted to the sequences of data. The ticks on x-axis symbolize the beginning of next sequence and at the same time a different domain. Window for calculating average values is cleared in between sequences.

Conclusions

- **State-of-the-art methods** are **inadequate** in **real-life** settings, as they **fall short** in achieving accuracy comparable to the **frozen source model**.
- We **propose a novel and straightforward method** called AR-TTA. It achieves **state-of-the-art performance** on various benchmarks, **consistently outperforming** the source model, which serves as the ultimate baseline for **feasible TTA** methods.
- Our more **realistic evaluation of TTA** with a variety of different datasets provides a **better understanding** of their potential **benefits and shortcomings**.

References

[1] Junyuan Hong, Lingjuan Lyu, Jiayu Zhou, and Michael Spranger. Mecta: Memory-economic continual test-time adaptation. In *ICLR*, 2023.

[2] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 16888–16905. PMLR, 2022.

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

[4] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. SHIFT: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21371–21382, June 2022.

[5] Eli Verwimp, Kuo Yang, Sarah Parisot, Lanqing Hong, Steven McDonagh, Eduardo Pérez-Pellitero, Matthias De Lange, and Tinne Tuytelaars. Clad: A realistic continual learning benchmark for autonomous driving. *Neural Networks*, 161:659–669, 2023.

[6] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[7] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7191–7201. IEEE, 2022.

Read the paper at