Supplementary Material for Continual Test-Time Domain Adaptation

Qin Wang[1] Olga Fink[1,3*] Luc Van Gool[1,4] Dengxin Dai[2]
[1]ETH Zurich, Switzerland [2]MPI for Informatics, Germany [3]EPFL, Switzerland [4]KU Lueven, Belgium
{qin.wang,vangool,dai}@vision.ee.ethz.ch olga.fink@epfl.ch

In this supplementary, we provide additional analysis and implementation details for the proposed test-time adaptation method. We also provide our code in the attachment.

1. Choice of Trainable Parameters

Unlike *TENT* based methods which only update the BN parameters, we update all trainable parameters in the network. This is achieved by reducing error accumulation using the improved pseudo-labels and our proposed stochastic restoration. As shown in Table A1, if we naively update all parameters using the entropy minimization techniques in *TENT*, the model quickly collapses because of the error accumulation. With the help of our proposed method which improves the target quality and preserve the information from the source model, the learning on all parameters can be successfully achieved. In the last two lines in Table A1, we show that learning all parameters instead of only Batch Normalization parameters can yield a 2.3% absolute improvement in terms of error rate for our CoTTA approach.

Table A1. Ablation study on the choice of trainable parameters. Classification error (%) is reported for the CIFAR10-to-CIFAR10C online continual test-time adaptation task.

Method	BN	All	Error
Source			43.5
TENT-continual(BN)	✓		20.7
TENT-continual(ALL)		✓	90.0
TENT-continual(ALL*) [4]		✓	19.8
CoTTA(BN)	✓		18.5
CoTTA(ALL)		✓	16.2

* indicates a frozen last layer and a smaller learning rate.

2. Effect of Learning Rate on Error Accumulation for Entropy Minimization

As shown in the previous section, TENT does not work when updating all trainable parameters by the same objective because it quickly reinforces the error predictions. One of the possible explanation could be that the default learning rate is too large for TENT and leads to fast error accumulation. Here, we evaluate this possibility and show that this is not the case. We show in Table A2 that tuning the learning rate cannot lead to the same performance level of the proposed CoTTA method. In contrast, the proposed method CoTTA works well on the default learning rate without the need to tune the learning rate.

3. Augmentation Confidence Threshold

As mentioned in the main paper, the proposed method uses augmentations to improve the pseudo-label quality. More specifically, we use a confidence threshold p_{th} in Equation 5 to determine whether to adopt this augmentation-averaged pseudo-label. This is necessary as we observe that naively augmenting all images can lead to performance decrease on some input images. We show this phenomenon in Table A3. Compared to the CoTTA model without any augmentation, naively using the additional augmentations does not provide improvement on some of the corruption types (e.g. performance on *Contrast* drops significantly). This negative effect is different across different corruption types, and seems to be more dominant when the domain difference is smaller. We also notice that the unimproved corruption types usually already have high confidence when predicted from the source model (f_{θ_0}). Therefore, we use a threshold p_{th} to filter the images, and do not apply augmentations on those with high confidence. More specifically, we design $p_{th} = \text{conf}^S - \delta$, where conf^S is the 5% quantile for the softmax predictions' confidence on the source images from the source model f_{θ_0}. $\delta = 0.05$ is a

Table A2. Ablation study on the choice of learning rate for TENT models when the optimization is on all parameters.

Method	All params	Learning Rate	Error
TENT-Continual (BN)		1e-3 (default)	20.7
TENT-Continual	x	1e-3	90.0
TENT-Continual	x	1e-4	66.9
TENT-Continual	x	1e-5	19.8
TENT-Continual	x	1e-6	19.8
CoTTA (Proposed)	x	1e-3	16.2

*The corresponding author

Table A3. Random Augmentations can bring negative effect when the target domain prediction is already confident. Results are collected from the CIFAR10-to-CIFAR10C online continual test-time adaptation task. All results (error rate in %) are evaluated on the ResNeXt-29 architecture with the largest corruption severity level.

Time	$t \longrightarrow$														
Target	*Gaussian*	*shot*	*impulse*	*defocus*	*glass*	*motion*	*zoom*	*snow*	*frost*	*fog*	*brightness*	*contrast*	*elastic_trans*	*pixelate*	*jpeg*
Average Confidence ($f_{\theta_0}(x^T)$)	0.91	0.92	0.89	0.96	0.89	0.95	0.95	0.94	0.94	0.95	0.97	0.94	0.92	0.93	0.92
CoTTA (w/o augmentation)	27.3	23.5	32.4	11.9	30.7	12.3	10.6	15.2	14.5	12.5	7.7	10.9	18.3	13.8	19.6
CoTTA (w/ augmentation w/o thresholding)	23.4	20.4	26.8	16.2	28.0	16.2	14.7	17.8	15.7	17.7	10.9	22.2	20.0	16.2	18.4
Improvement	3.9	3.1	5.6	-4.3	2.7	-3.9	-4.2	-2.6	-1.2	-5.1	-3.2	-11.4	-1.7	-2.3	1.2

small tolerance term. We use this definition of p_{th} for all our three experiments and find it effective. This design is also supported by a very recent work [2], where the aurhors also observe a positive correlation between the confidence and performance under domain shift. We would like to highlight that this design avoids using any test data to determine the threshold.

4. Choice of Knowledge Preservation Method

Preserving knowledge learned from the past is an active research direction in continual learning, and is very related to the proposed stochastic restoration method. In this ablation study, we compare the performance of stochastic restoration with the popular learning without forgetting (LwF) [3] method. If we replace the restoration module from our final proposed model with LwF, the error rate increased from 16.2% to 17.3% for CIFAR-10C standard experiment. This indicates that the proposed stochastic restoration is more robust on the continual adaptation task. This is most likely because stochastic restoration is not linked to any gradient descent optimization process, while the regularization objective for the optimization in LwF can be unstable under the continually changing distribution shift because of the mis-calibrated and overconfident predictions under scenarios with large domain gaps.

5. Choice of the Restoration Factor

As shown in Equation 6, the restoration factor M is sampled from a Bernoulli distribution parameterized by probability p. We now present the ablation of p on CIFAR10C. A larger p restores more source knowledge and a p that is too large can prevent the model from adapting.

Choice of p	0	0.0001	0.0001	**0.01**	0.1
Error (%)	17.4	17.4	16.9	**16.2**	17.5

6. Ablation on the EMA Factor

The EMA model represents a temporal ensemble of models that were adapted to different test domains. In that way, a knowledge base of different domains is built, which will also improve the generalization capability of the EMA model to new unseen domains, which increases the pseudo-label quality. The adopted default EMA factor $\alpha = 0.999$ is well-studied in Section 3.4 in the mean teacher paper [53]. We find it suitable for all our exps. Here, we provide the ablation of α on CIFAR10-C.

Choice of α	0.99	0.995	**0.999**	0.9995	0.9999
Error (%)	19.0	16.7	**16.2**	16.9	18.3

7. Experiment Results on Cityscapes-to-ACDC

We present the complete experiment results on the continual test-time adaptation task for Cityscapes-to-ACDC in Table A4. Experiments show that while TENT model suffer largely from error accumulation over time, the proposed CoTTA model can largely maintain the strong performance in the long term.

8. Limitation

One limitation regarding our work is about the augmentation. While augmentation brings improvement on the test-time performance, it also requires extra computational power, which maybe unavailable during inference time for some real-time applications. One possible solution to this is to learn an efficient augmentation strategy for the test time data of the current time step, instead of applying a large number of random augmentations. While the idea of learning augmentation strategy were discussed for the training stage [1], it remains largely unexplored for test-time augmentation.

In addition, the proposed model is designed to be general and did not consider task-specific prior knowledge. Some prior knowledge maybe domain-invariant and might serve well as the test-time supervision. For example, for tasks with temporal information, it maybe useful to make use of the temporal consistency or other prior knowledge as the supervision to further improve the adaptation performance.

Finally, the evaluation tasks in our work try to mimic

Table A4. Semantic segmentation results (mIoU in %) on the Cityscapes-to-ACDC online continual test-time adaptation task. We evaluate the four test conditions continually for ten times to evaluate the long-term adaptation performance. All results are evaluated based on the Segformer-B5 architecture.

Time	$t \longrightarrow$																				
Condition	Fog	Night	rain	snow	Fog	Night	rain	snow	Fog	Night	rain	snow	Fog	Night	rain	snow	Fog	Night	rain	snow	cont.
Round	1				2				3				4				5				cont.
Source	69.1	40.3	59.7	57.8	69.1	40.3	59.7	57.8	69.1	40.3	59.7	57.8	69.1	40.3	59.7	57.8	69.1	40.3	59.7	57.8	cont.
BN Stats Adapt	62.3	38.0	54.6	53.0	62.3	38.0	54.6	53.0	62.3	38.0	54.6	53.0	62.3	38.0	54.6	53.0	62.3	38.0	54.6	53.0	cont.
Tent-continual	69.0	40.2	60.1	57.3	68.3	39.0	60.1	56.3	67.5	37.8	59.6	55.0	66.5	36.3	58.7	54.0	65.7	35.1	57.7	53.0	cont.
Proposed	70.9	41.2	62.4	59.7	70.9	41.1	62.6	59.7	70.9	41.0	62.7	59.7	70.9	41.0	62.7	59.7	70.9	41.0	62.8	59.7	cont.
Round	6				7				8				9				10				Mean
Source	69.1	40.3	59.7	57.8	69.1	40.3	59.7	57.8	69.1	40.3	59.7	57.8	69.1	40.3	59.7	57.8	69.1	40.3	59.7	57.8	56.7
BN Stats Adapt	62.3	38.0	54.6	53.0	62.3	38.0	54.6	53.0	62.3	38.0	54.6	53.0	62.3	38.0	54.6	53.0	62.3	38.0	54.6	53.0	52.0
Tent-continual	64.9	34.0	56.5	52.0	64.2	32.8	55.3	50.9	63.3	31.6	54.0	49.8	62.5	30.6	52.9	48.8	61.8	29.8	51.9	47.8	52.3
Proposed	70.9	41.0	62.8	59.7	70.9	41.0	62.8	59.7	70.9	41.0	62.8	59.7	70.8	41.0	62.8	59.7	70.8	41.0	62.8	59.7	58.6

the real-world adaptation scenarios by making use of corruption and weather changes. While this imitation is reasonable, it does not consider some real-world restrictions. For example, in the real world, the data distribution is often long-tailed and is changing continuously. Extending test-time adaptation to more real-world scenarios in the wild can be an interesting future work.

References

[1] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, pages 113–123, 2019. 2

[2] Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *ICCV*, pages 1134–1144, 2021. 2

[3] Zhizhong Li and Derek Hoiem. Learning without forgetting. *T-PAMI*, 40(12):2935–2947, 2017. 2

[4] Dequan Wang, Shaoteng Liu, Sayna Ebrahimi, Evan Shelhamer, and Trevor Darrell. On-target adaptation. *arXiv preprint arXiv:2109.01087*, 2021. 1