Supplementary Materials for Robust Test-Time Adaptation in Dynamic Scenarios

A. Discussion

Societal impact. RoTTA enables adapting pre-trained models on continually changing distributions with correlatively sampled test streams without any more raw data or label requirements. Thus, our work may have a positive impact on communities to effectively deploy and adapt models in various real-world scenarios, which is economically and environmentally friendly. And since no training data is required, this protects data privacy and has potential commercial value. We carry out experiments on benchmark datasets and do not notice any societal issues. It does not involve sensitive attributes.

Future work. Our work suggests a few promising directions for future work. Firstly, the proposed RoTTA is a preliminary attempt to perform test-time adaptation for the more realistic test stream under the setup PTTA. One could experiment to improve the algorithm by replacing some parts of RoTTA. More importantly, we hope that with this work, we can open a path to the original goal of test-time adaptation, which is performing test-time adaptation in real-world scenarios. Thus, one could improve PTTA to make it more realistic.

Limitations. RoTTA achieves excellent performance on various tasks under the setup PTTA as demonstrated in Section 4 in the main paper, but we still find some limitations of it. Firstly, the adopted robust batch normalization (RBN) is a naive solution to the normalization of the correlatively sampled batch of data. This requires careful design of the value of α in RBN. Secondly, we observe that during the adaptation procedure of some methods like PL [3] and TENT [5], the model collapse finally. Although we design many strategies to stabilize the adaptation and model collapse never occurs in the experiments of RoTTA, we are still missing a way to recover the model from the collapse state as a remedy. Thirdly, category similarity is only one kind of correlation. Although we conduct experiments on different datasets with Dirichlet distribution to simulate correlatively sampled test streams, we still need to validate our approach in some real-world scenarios.

B. Sensitivity to different hyper-parameters

In this section, we conduct a detailed sensitivity analysis of the hyperparameters involved in RoTTA. All experiments are conducted on CIFAR100→CIFAR100-C, and the corruptions changes as *motion, snow, fog, shot, defocus, contrast, zoom, brightness, frost, elastic, glass, gaussian, pixelate, jpeg,* and *impulse*, and test streams are sampled correlatively with the Dirichlet parameter $\delta = 0.1$. When we investigate the sensitivity to a specific hyperparameter, other hyperparameters are fixed to the default values, i.e., $\lambda_t = 1.0$, $\lambda_u = 1.0$, $\alpha = 0.05$, and $\nu = 0.001$, for all experiments.

Table A. Classification error with different value of λ_t/λ_u.

λ_t/λ_u	0.0/2.0	0.5/1.5	1.0/1.0	1.5/ 0.5	2.0/ 0.0
CIFAR100-C	57.5	36.9	**35.0**	35.9	38.9

Trade-off between timeliness and uncertainty. When updating the memory bank, we take the timeliness and uncertainty of samples into account simultaneously, and λ_t and λ_u will make a trade-off between them. In Table A, we show the results of RoTTA with varying λ_t/λ_u, i.e., $\lambda_t/\lambda_u \in \{0.0/2.0, 0.5/1.5, 1.0/1.0, 1.5/0.5, 2.0/0.0\}$. When we consider both of them, the results are relatively stable (35.0-36.9%). When we only think about one side, the performance drops significantly. For example, when we set $\lambda_t/\lambda_u = 0.0/2.0$ which means only considering uncertainty, the performance drops 22.5%. That's because some confident samples get stuck in the memory bank, making it not work the way we design it.

Table B. Classification error with varying α

α	0.5	0.1	0.05	0.01	0.005	0.001
CIFAR100-C	39.0	36.0	**35.0**	36.0	38.1	41.5

Sensitivity to α. We show the results of RoTTA with varying α, i.e., $\alpha \in \{0.5, 0.1, 0.05, 0.01, 0.005, 0.001\}$ in Table B. A larger value of α means updating the global statistics faster and vice versa. We can see that RoTTA achieves competitive results $(35.0 - 36.0\%)$ at appropriate values of

α, i.e., $\alpha \in \{0.1, 0.05, 0.01\}$. Updating too aggressively or too gently can lead to unreliable estimates of statistics.

Table C. Classification error with varying ν

ν	0.05	0.01	0.005	0.001	0.0005	0.0001
CIFAR100-C	44.8	39.1	37.1	**35.0**	37.6	43.6

Sensitivity to ν. We show the results of RoTTA with varying ν, i.e., $\nu \in \{0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001\}$ in Table C. As we can see, the best performance is achieved at $\nu = 0.001$. Updating the teacher model too quickly or too slowly can cause performance degradation.

C. Additional experiment details and results

C.1 Compared methods

BN [4] utilizes statistics of the current batch of data to normalize their feature maps without tuning any parameters.

PL [3] is based on BN [4], and adopts pseudo labels to train the affine parameters in BN layers.

TENT [5] is the first to propose fully test-time adaptation. It adopts test-time batch normalization and utilizes entropy minimization to train the affine parameters of BN layers. We reimplement it following the released code https://github.com/DequanWang/tent.

LAME [1] adapts the output of the pre-trained model by optimizing a group of latent variables without tuning any inner parts of the model. We reimplement it following the released code https://github.com/fiveai/LAME.

CoTTA [6] considers performing test-time adaptation on continually changing distributions and propose augmentation-averaged pseudo-labels and stochastic restoration to address error accumulation and catastrophic forgetting. We reimplement it following the released code https://github.com/qinenergy/cotta.

NOTE [2] proposes instance-aware normalization and prediction-balanced reservoir sampling to stable the adaptation on temporally correlated test streams. We reimplement it following the released code https://github.com/TaesikGong/NOTE.

C.2 Simulate correlatively sampling

As we described in the scenarios of autonomous driving that the car will follow more vehicles on the highway or will encounter more pedestrians on the sidewalk, so we use the same category to simulate correlation. From a macro point of view, the test distribution $\mathcal{P}_{test}$ changes continually as $\mathcal{P}_0, \mathcal{P}_1, ..., \mathcal{P}_\infty$. During the period when $\mathcal{P}_{test} = \mathcal{P}_t$, we adopt Dirichlet distribution to simulate correlatively sampled test stream. More specifically, we consider dividing samples of $\mathcal{C}$ classes into T slots. Firstly, we utilize Dirichlet distribution with parameter γ to generate the partition criterion $q \in \mathbb{R}^{\mathcal{C}\times T}$. Then for each class c, we split samples into T parts according to q_c and assign each part to each slot respectively. Finally, we concatenate all slots to simulate the correlatively sampled test stream for $\mathcal{P}_{test} = \mathcal{P}_t$. And as $\mathcal{P}_{test}$ changes, we use the above method again to generate the test stream.

C.3 Detailed results of different orders

We report the average classification error of ten different distribution changing orders in Table 6 of the main paper. And then we present the specific results here, including Table D, E, F, G, H, I, J, K, L, and M for CIFAR10→CIFAR10-C and Table N, O, P, Q, R, S, T, U, V, and W for CIFAR100→CIFAR100-C. We can see consistently superior performance of RoTTA. One thing to mention is that on DomainNet we use alphabetical order to determine the order of domain changes.

References

[1] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *CVPR*, pages 8344–8353, 2022. 2, 3, 4, 5, 6

[2] Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. Robust continual test-time adaptation: Instance-aware BN and prediction-balanced memory. In *NeurIPS*, 2022. 2, 3, 4, 5, 6

[3] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 1, 2, 3, 4, 5, 6

[4] Zachary Nado, Shreyas Padhy, D. Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *CoRR*, abs/2006.10963, 2020. 2, 3, 4, 5, 6

[5] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 1, 2, 3, 4, 5, 6

[6] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *CVPR*, pages 7191–7201, 2022. 2, 3, 4, 5, 6

Table D. Average classification error of the task CIFAR10 → CIFAR10-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	t →															
Method	brightness	pixelate	gaussian	motion	zoom	glass	impulse	jpeg	defocus	elastic	shot	frost	snow	fog	contrast	Avg.
Source	9.3	58.5	72.3	34.8	42.0	54.3	72.9	30.3	46.9	26.6	65.7	41.3	25.1	26.0	46.7	43.5
BN [4]	71.1	75.2	76.8	74.2	73.7	80.1	79.3	77.5	73.8	77.7	77.2	73.3	73.8	72.7	71.7	75.2
PL [3]	71.7	75.9	80.2	78.4	80.2	85.2	85.3	85.4	85.1	86.7	87.9	87.9	88.1	88.3	87.9	83.6
TENT [5]	71.6	75.9	81.3	80.5	82.3	85.6	87.1	87.0	87.1	88.1	88.2	87.8	87.9	88.3	88.2	84.4
LAME [1]	**5.4**	56.8	73.1	29.1	37.0	50.5	71.4	**22.3**	42.8	**18.6**	65.5	37.3	**18.8**	20.4	43.6	39.5
CoTTA [6]	75.0	79.8	83.1	83.4	83.2	84.0	84.5	83.2	83.5	83.3	83.6	83.0	83.0	83.4	83.7	82.6
NOTE [2]	10.1	29.9	47.1	**23.4**	28.4	48.4	46.1	41.8	26.9	36.1	37.5	25.0	25.0	23.2	**14.2**	30.9
RoTTA	10.4	**26.6**	**37.5**	23.9	**17.0**	**40.9**	**39.7**	30.1	**18.0**	29.9	**30.1**	**23.6**	21.7	**17.6**	19.0	**25.7**(+5.2)

Table E. Average classification error of the task CIFAR10 → CIFAR10-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	t →															
Method	jpeg	shot	zoom	frost	contrast	fog	defocus	elastic	gaussian	brightness	glass	impulse	pixelate	snow	motion	Avg.
Source	30.3	65.7	42.0	41.3	46.7	26.0	46.9	26.6	72.3	9.3	54.3	72.9	58.5	25.1	34.8	43.5
BN [4]	77.6	75.8	73.4	74.1	73.1	72.5	72.9	77.1	77.2	72.2	79.9	79.9	75.5	74.6	72.9	75.2
PL [3]	77.6	77.1	76.6	78.3	77.5	79.8	82.0	84.8	86.1	83.5	87.8	87.1	86.5	85.6	85.7	82.4
TENT [5]	78.5	78.2	79.2	81.8	84.8	84.8	86.4	87.3	87.9	86.7	87.3	87.8	87.2	87.5	87.1	84.8
LAME [1]	**22.5**	65.2	37.0	37.1	44.0	**20.3**	41.7	**18.7**	72.8	**5.2**	51.2	71.5	57.0	**19.0**	29.4	39.5
CoTTA [6]	78.5	81.0	82.8	84.1	84.9	83.4	83.5	83.5	84.5	83.3	84.7	84.6	83.0	84.4	83.4	83.3
NOTE [2]	35.4	36.1	22.1	**21.3**	**11.6**	24.8	24.5	36.0	37.7	18.4	49.0	47.4	43.9	30.4	29.2	31.2
RoTTA	33.2	**33.3**	**19.8**	24.1	24.9	20.5	**16.2**	31.7	**28.4**	11.8	**43.1**	**36.9**	**32.5**	20.7	**20.6**	**26.5**(+4.7)

Table F. Average classification error of the task CIFAR10 → CIFAR10-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	t →															
Method	contrast	defocus	gaussian	shot	snow	frost	glass	zoom	elastic	jpeg	pixelate	brightness	impulse	motion	fog	Avg.
Source	46.7	46.9	72.3	65.7	25.1	41.3	54.3	42.0	26.6	30.3	58.5	9.3	72.9	34.8	26.0	43.5
BN [4]	72.3	72.6	76.9	77.1	74.8	73.5	80.0	73.2	77.4	78.6	76.4	71.0	79.1	73.9	71.5	75.2
PL [3]	72.4	75.3	80.7	82.6	83.3	83.5	86.6	85.7	86.6	88.4	87.5	86.6	88.3	88.2	86.8	84.1
TENT [5]	73.5	77.9	85.5	86.9	87.6	87.8	88.3	87.7	88.6	89.2	88.5	88.5	89.3	88.6	88.6	86.4
LAME [1]	43.5	42.3	73.1	65.3	**19.2**	37.3	51.1	36.8	**18.5**	**22.5**	56.9	**5.5**	71.1	29.1	20.5	39.5
CoTTA [6]	79.4	80.3	83.8	83.9	83.9	83.4	85.0	83.2	85.1	84.3	83.9	83.3	84.7	83.9	82.5	83.4
NOTE [2]	**9.6**	21.8	40.1	**31.0**	25.5	22.6	44.8	22.8	33.2	39.4	33.2	18.1	50.0	28.3	29.8	30.0
RoTTA	18.4	**17.9**	**38.4**	31.9	23.3	**19.8**	**40.7**	**17.4**	31.4	29.8	**27.8**	11.3	**43.8**	**19.7**	**18.8**	**26.0**(+4.0)

Table G. Average classification error of the task CIFAR10 → CIFAR10-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	t →															
Method	shot	fog	glass	pixelate	snow	elastic	brightness	impulse	defocus	frost	contrast	gaussian	motion	jpeg	zoom	Avg.
Source	65.7	26.0	54.3	58.5	25.1	26.6	9.3	72.9	46.9	41.3	46.7	72.3	34.8	30.3	42.0	43.5
BN [4]	76.4	72.0	80.4	76.2	74.8	77.0	71.1	79.6	73.8	74.4	73.0	77.0	72.5	78.3	72.5	75.3
PL [3]	77.0	73.3	82.4	79.8	81.0	82.3	79.5	84.4	82.7	83.5	83.5	85.5	84.8	87.0	84.5	82.1
TENT [5]	76.9	74.6	82.3	81.7	82.0	84.9	84.8	87.3	86.6	87.3	87.6	89.2	88.3	88.9	87.3	84.6
LAME [1]	65.3	**20.6**	50.9	56.7	**19.2**	**18.8**	**5.4**	71.8	42.8	37.2	43.3	73.2	29.4	**22.6**	36.9	39.6
CoTTA [6]	77.4	77.6	83.8	81.9	82.2	82.6	80.4	83.3	82.3	81.5	82.7	82.6	81.1	82.9	81.0	81.6
NOTE [2]	**34.0**	20.9	**43.1**	36.6	24.0	36.4	12.1	48.0	25.9	23.9	**13.4**	38.1	**25.0**	43.2	24.2	29.9
RoTTA	35.0	21.1	43.9	**29.2**	22.1	29.7	10.8	**44.6**	**25.3**	**22.7**	24.6	**29.4**	26.9	34.4	**16.1**	**27.7**(+2.2)

Table H. Average classification error of the task CIFAR10 → CIFAR10-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	t →															
Method	pixelate	glass	zoom	snow	fog	impulse	brightness	motion	frost	jpeg	gaussian	shot	contrast	defocus	elastic	Avg.
Source	58.5	54.3	42.0	25.1	26.0	72.9	9.3	34.8	41.3	30.3	72.3	65.7	46.7	46.9	26.6	43.5
BN [4]	76.0	79.6	73.3	75.2	72.9	79.8	71.1	73.5	74.1	78.6	77.4	76.1	72.0	73.8	76.4	75.3
PL [3]	76.7	81.3	77.4	80.3	81.2	86.3	83.3	85.9	86.2	87.7	88.1	88.4	87.4	87.6	87.7	84.4
TENT [5]	76.4	80.2	77.8	81.2	83.0	87.1	85.6	87.2	87.6	88.7	88.6	88.9	88.5	88.6	88.2	85.2
LAME [1]	56.9	50.7	37.0	**19.0**	20.3	71.5	**5.4**	29.2	37.2	**22.5**	73.0	65.3	43.8	42.4	**18.7**	39.5
CoTTA [6]	77.1	83.6	84.1	84.8	84.4	85.2	84.0	84.3	84.9	84.9	85.0	84.7	85.3	84.4	84.3	84.1
NOTE [2]	27.8	52.2	24.5	22.3	21.6	44.5	14.5	**21.3**	25.9	42.5	38.8	36.0	**16.7**	28.1	40.6	30.5
RoTTA	**25.9**	**43.3**	**17.7**	22.1	**20.2**	**41.5**	12.2	22.9	**22.5**	31.2	**33.8**	**26.0**	31.4	**17.7**	27.6	**26.4**(+4.1)

Table I. Average classification error of the task CIFAR10 → CIFAR10-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	$t \longrightarrow$															
Method	motion	snow	fog	shot	defocus	contrast	zoom	brightness	frost	elastic	glass	gaussian	pixelate	jpeg	impulse	Avg.
Source	34.8	25.1	26.0	65.7	46.9	46.7	42.0	9.3	41.3	26.6	54.3	72.3	58.5	30.3	72.9	43.5
BN [4]	73.2	73.4	72.7	77.2	73.7	72.5	72.9	71.0	74.1	77.7	80.0	76.9	75.5	78.3	79.0	75.2
PL [3]	73.9	75.0	75.6	81.0	79.9	80.6	82.0	83.2	85.3	87.3	88.3	87.5	87.5	87.5	88.2	82.9
TENT [5]	74.3	77.4	80.1	86.2	86.7	87.3	87.9	87.4	88.2	89.0	89.2	89.0	88.3	89.7	89.2	86.0
LAME [1]	29.5	**19.0**	20.3	65.3	42.4	43.4	36.8	**5.4**	37.2	**18.6**	51.2	73.2	57.0	**22.6**	71.3	39.5
CoTTA [6]	77.1	80.6	83.1	84.4	83.9	84.2	83.1	82.6	84.4	84.2	84.5	84.6	82.7	83.8	84.9	83.2
NOTE [2]	**18.0**	22.1	20.6	35.6	26.9	**13.6**	26.5	17.3	27.2	37.0	48.3	38.8	42.6	41.9	49.7	31.1
RoTTA	18.1	21.3	**18.8**	**33.6**	**23.6**	16.5	**15.1**	11.2	**21.9**	30.7	**39.6**	**26.8**	**33.7**	27.8	**39.5**	**25.2**(+5.9)

Table J. Average classification error of the task CIFAR10 → CIFAR10-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	$t \longrightarrow$															
Method	frost	impulse	jpeg	contrast	zoom	glass	pixelate	snow	defocus	motion	brightness	elastic	shot	fog	gaussian	Avg.
Source	41.3	72.9	30.3	46.7	42.0	54.3	58.5	25.1	46.9	34.8	9.3	26.6	65.7	26.0	72.3	43.5
BN [4]	73.8	79.1	77.9	73.0	73.7	80.1	75.7	74.4	73.7	74.0	71.7	77.0	75.9	72.8	76.2	75.3
PL [3]	74.2	80.9	80.4	79.5	81.8	85.9	83.9	85.1	84.7	85.9	85.9	86.7	87.2	87.0	87.8	83.8
TENT [5]	73.9	80.3	81.8	81.6	83.6	86.3	85.6	85.7	86.4	87.7	87.4	88.8	88.8	88.5	88.4	85.0
LAME [1]	37.4	71.8	**22.4**	43.5	37.0	50.5	57.0	**19.0**	42.8	29.1	**5.4**	**18.7**	65.2	20.4	72.9	39.5
CoTTA [6]	76.5	82.2	82.8	85.0	82.9	85.0	83.0	82.9	83.5	83.4	82.6	83.7	83.2	83.3	83.6	82.9
NOTE [2]	**21.1**	**41.4**	36.3	**10.2**	21.7	46.7	37.5	26.4	26.1	21.4	14.3	37.9	38.5	24.4	40.7	29.6
RoTTA	22.2	44.9	35.2	18.8	**19.7**	**41.5**	**28.5**	23.2	**21.2**	**18.6**	12.4	30.0	**27.4**	**20.0**	**31.2**	**26.3**(+3.3)

Table K. Average classification error of the task CIFAR10 → CIFAR10-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	$t \longrightarrow$															
Method	defocus	motion	zoom	shot	gaussian	glass	jpeg	fog	contrast	pixelate	frost	snow	brightness	elastic	impulse	Avg.
Source	46.9	34.8	42.0	65.7	72.3	54.3	30.3	26.0	46.7	58.5	41.3	25.1	9.3	26.6	72.9	43.5
BN [4]	72.8	72.7	73.3	77.2	77.3	80.0	77.6	72.6	73.3	76.6	73.8	74.1	70.3	77.5	79.0	75.2
PL [3]	73.2	74.6	76.5	81.7	82.8	84.6	85.1	84.6	86.2	86.4	86.1	87.1	86.8	88.4	88.1	83.5
TENT [5]	73.7	74.3	77.1	82.5	84.3	86.9	87.4	86.6	88.0	88.5	88.1	88.5	88.4	89.4	88.9	84.8
LAME [1]	42.5	29.3	37.0	65.3	73.2	50.5	**22.5**	20.5	43.5	56.9	37.1	**18.9**	**5.4**	**18.5**	71.3	39.5
CoTTA [6]	76.3	79.8	82.4	83.3	83.8	84.5	83.1	82.7	84.7	82.9	83.0	83.3	81.4	83.8	83.8	82.6
NOTE [2]	18.5	18.8	23.6	36.5	33.7	47.8	38.6	22.8	**13.0**	40.0	29.2	26.3	17.5	44.0	52.9	30.9
RoTTA	**17.0**	**17.5**	**16.5**	**33.8**	**33.3**	**42.7**	29.4	**18.0**	19.6	**29.5**	**20.7**	22.1	11.5	29.5	**38.1**	**25.3**(+5.6)

Table L. Average classification error of the task CIFAR10 → CIFAR10-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	$t \longrightarrow$															
Method	glass	zoom	impulse	fog	snow	jpeg	gaussian	frost	shot	brightness	contrast	motion	pixelate	defocus	elastic	Avg.
Source	54.3	42.0	72.9	26.0	25.1	30.3	72.3	41.3	65.7	9.3	46.7	34.8	58.5	46.9	26.6	43.5
BN [4]	79.7	72.3	79.8	73.2	74.7	77.7	76.6	73.2	77.1	72.2	73.0	73.3	75.5	73.8	76.4	75.2
PL [3]	79.6	73.2	81.3	77.3	79.1	83.0	83.2	83.0	85.5	84.3	87.0	86.9	86.4	86.5	87.6	82.9
TENT [5]	79.5	74.1	84.2	82.2	84.5	86.5	86.7	85.9	87.2	86.6	86.8	87.3	86.9	87.4	87.3	84.9
LAME [1]	50.8	36.9	71.3	**20.6**	**19.2**	**22.4**	72.5	37.2	65.4	**5.2**	43.3	29.1	57.0	42.4	**18.7**	39.5
CoTTA [6]	81.5	79.4	85.2	84.1	84.5	84.2	84.8	84.0	84.8	83.2	85.2	83.8	83.2	84.6	83.6	83.7
NOTE [2]	45.0	21.2	**42.3**	21.0	21.6	38.4	36.4	21.4	33.1	16.7	**14.6**	25.4	43.5	29.1	38.5	29.9
RoTTA	**42.6**	**17.6**	48.1	23.9	21.9	32.6	**32.1**	**20.7**	**30.2**	12.0	21.9	**20.0**	**33.7**	**16.4**	28.1	**26.8**(+3.1)

Table M. Average classification error of the task CIFAR10 → CIFAR10-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	$t \longrightarrow$															
Method	contrast	gaussian	defocus	zoom	frost	glass	jpeg	fog	pixelate	elastic	shot	impulse	snow	motion	brightness	Avg.
Source	46.7	72.3	46.9	42.0	41.3	54.3	30.3	26.0	58.5	26.6	65.7	72.9	25.1	34.8	9.3	43.5
BN [4]	72.4	76.2	73.2	73.7	73.6	80.0	77.6	72.6	76.4	77.7	77.2	79.9	73.8	73.9	70.0	75.2
PL [3]	73.0	78.2	76.7	79.7	81.6	85.6	86.0	85.3	87.2	88.2	88.3	88.9	88.5	89.2	88.2	84.3
TENT [5]	73.6	80.9	83.1	85.6	87.1	88.5	88.8	88.4	89.2	89.3	89.0	89.0	89.3	89.9	89.1	86.7
LAME [1]	43.5	73.2	42.3	37.0	37.2	50.5	**22.5**	20.5	57.0	**18.6**	65.5	71.5	**18.8**	29.1	**5.6**	39.5
CoTTA [6]	79.5	81.4	83.4	83.6	83.9	85.0	84.0	82.8	84.8	84.8	84.5	84.7	84.1	84.4	82.8	83.6
NOTE [2]	**9.6**	43.6	26.5	24.8	23.9	46.9	38.0	23.4	**34.0**	41.2	41.5	45.0	27.6	25.8	19.0	31.4
RoTTA	18.4	**36.0**	**21.1**	**15.6**	**23.0**	**41.7**	30.8	**19.1**	34.1	31.1	**31.3**	**39.9**	26.0	**18.8**	12.8	**26.6**(+4.8)

Table N. Average classification error of the task CIFAR100 → CIFAR100-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	$t \longrightarrow$															
Method	brightness	pixelate	gaussian	motion	zoom	glass	impulse	jpeg	defocus	elastic	shot	frost	snow	fog	contrast	Avg.
Source	29.5	74.7	73.0	30.8	28.8	54.1	39.4	41.2	29.3	37.2	68.0	45.8	39.5	50.3	55.1	46.4
BN [4]	46.5	52.0	58.6	47.4	47.4	57.6	58.2	56.9	47.0	53.4	56.0	52.5	53.1	57.7	49.1	52.9
PL [3]	48.5	60.7	77.1	85.9	91.5	95.5	95.8	96.6	96.8	96.9	97.3	97.5	97.6	97.7	97.9	88.9
TENT [5]	49.8	69.4	92.2	96.0	96.7	97.3	97.5	97.9	97.5	97.9	98.0	98.2	98.2	98.2	98.2	92.2
LAME [1]	**21.7**	75.1	72.7	**22.9**	**20.6**	49.0	**32.1**	**33.3**	**21.2**	**28.0**	66.8	40.0	**30.6**	43.9	51.3	40.6
CoTTA [6]	46.8	48.4	54.7	48.7	48.6	53.5	55.4	52.8	49.8	51.8	53.5	52.9	54.1	56.7	53.6	52.1
NOTE [2]	42.6	53.0	69.9	52.1	53.3	70.4	73.1	76.7	80.8	96.0	97.7	97.1	96.6	97.2	95.8	76.8
RoTTA	28.4	**37.3**	**44.6**	31.9	28.3	**41.8**	43.6	39.9	28.0	35.2	**38.2**	**33.7**	33.0	**39.5**	**31.0**	**35.6**(+5.0)

Table O. Average classification error of the task CIFAR100 → CIFAR100-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	$t \longrightarrow$															
Method	jpeg	shot	zoom	frost	contrast	fog	defocus	elastic	gaussian	brightness	glass	impulse	pixelate	snow	motion	Avg.
Source	41.2	68.0	28.8	45.8	55.1	50.3	29.3	37.2	73.0	29.5	54.1	39.4	74.7	39.5	30.8	46.4
BN [4]	58.3	56.8	47.8	51.8	48.9	57.3	46.8	53.5	57.8	45.5	57.1	58.5	51.7	53.3	48.8	52.9
PL [3]	59.4	66.3	74.9	87.5	94.2	95.5	96.2	97.1	97.4	97.2	97.5	97.7	98.0	98.2	98.2	90.4
TENT [5]	62.0	79.3	91.7	95.8	96.9	97.0	97.4	97.7	97.6	97.7	97.9	97.9	98.0	97.9	97.9	93.5
LAME [1]	**33.6**	66.7	**21.1**	39.9	50.6	43.9	**21.0**	**28.6**	72.5	**21.6**	48.6	**32.5**	74.5	**30.6**	**22.5**	40.6
CoTTA [6]	54.6	54.1	49.6	52.1	52.7	58.0	50.3	53.3	55.0	49.1	55.4	55.7	51.0	54.6	52.1	53.2
NOTE [2]	60.4	63.0	49.9	55.7	47.0	65.2	59.4	76.6	90.9	87.2	96.8	97.0	97.3	96.7	96.8	76.0
RoTTA	43.9	**45.3**	31.0	**37.3**	**35.7**	**41.2**	27.7	34.8	**39.7**	26.6	**39.5**	41.9	**32.0**	33.0	30.5	**36.0**(+4.6)

Table P. Average classification error of the task CIFAR100 → CIFAR100-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	$t \longrightarrow$															
Method	contrast	defocus	gaussian	shot	snow	frost	glass	zoom	elastic	jpeg	pixelate	brightness	impulse	motion	fog	Avg.
Source	55.1	29.3	73.0	68.0	39.5	45.8	54.1	28.8	37.2	41.2	74.7	29.5	39.4	30.8	50.3	46.4
BN [4]	49.4	47.2	58.6	56.2	52.7	52.0	57.9	46.1	54.4	57.7	50.5	46.2	58.2	47.6	58.5	52.9
PL [3]	54.8	64.2	83.3	92.4	95.5	96.5	96.9	96.4	97.2	97.4	97.8	97.8	97.9	97.7	98.0	90.9
TENT [5]	60.2	83.1	95.2	96.5	96.9	97.3	97.0	97.3	97.8	97.8	97.6	97.9	97.8	97.9	98.1	93.9
LAME [1]	51.3	**21.3**	72.7	66.3	**30.2**	40.0	48.6	**20.9**	**27.7**	**33.3**	75.0	**21.5**	**32.2**	**22.5**	43.8	40.5
CoTTA [6]	52.1	48.6	55.1	52.7	53.4	51.9	55.9	49.2	53.2	52.8	49.2	49.7	56.2	50.7	58.1	52.6
NOTE [2]	**39.5**	45.9	68.8	61.8	57.4	58.5	71.4	66.5	80.8	90.9	94.2	94.9	97.0	95.5	96.6	74.6
RoTTA	41.7	30.5	**44.9**	**40.5**	35.4	**34.1**	**40.5**	28.2	34.5	39.5	**31.1**	26.7	43.3	31.4	**38.8**	**36.1**(+4.4)

Table Q. Average classification error of the task CIFAR100 → CIFAR100-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	$t \longrightarrow$															
Method	shot	fog	glass	pixelate	snow	elastic	brightness	impulse	defocus	frost	contrast	gaussian	motion	jpeg	zoom	Avg.
Source	68.0	50.3	54.1	74.7	39.5	37.2	29.5	39.4	29.3	45.8	55.1	73.0	30.8	41.2	28.8	46.4
BN [4]	57.5	58.6	58.5	50.5	52.7	53.1	45.9	57.9	47.0	51.5	47.8	58.2	48.2	57.1	47.7	52.8
PL [3]	59.5	72.9	85.1	89.6	94.5	96.8	97.1	97.9	97.8	98.0	98.3	98.2	98.0	98.0	98.2	92.0
TENT [5]	60.3	81.4	95.0	96.6	97.0	97.3	97.3	97.7	97.7	97.7	97.8	97.7	97.6	97.6	97.9	93.8
LAME [1]	66.4	**43.2**	49.0	75.2	**30.2**	**28.5**	**21.6**	**32.5**	**21.2**	39.5	52.0	72.8	**22.3**	**33.1**	**20.5**	40.5
CoTTA [6]	54.5	58.4	55.6	50.0	53.9	53.4	50.3	56.7	51.3	53.2	53.7	56.1	52.0	54.5	51.5	53.7
NOTE [2]	61.8	60.2	63.4	55.6	59.8	65.9	58.6	75.1	77.8	93.8	94.2	97.0	95.0	95.5	94.4	76.5
RoTTA	**45.5**	44.5	**43.5**	**35.6**	35.1	35.7	26.2	44.0	29.7	**34.2**	**32.0**	**40.7**	31.4	39.4	27.7	**36.3**(+4.2)

Table R. Average classification error of the task CIFAR100 → CIFAR100-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	$t \longrightarrow$															
Method	pixelate	glass	zoom	snow	fog	impulse	brightness	motion	frost	jpeg	gaussian	shot	contrast	defocus	elastic	Avg.
Source	74.7	54.1	28.8	39.5	50.3	39.4	29.5	30.8	45.8	41.2	73.0	68.0	55.1	29.3	37.2	46.4
BN [4]	51.7	58.6	47.8	52.9	57.1	58.2	45.9	47.6	52.9	57.8	57.5	56.7	49.5	46.1	54.0	52.9
PL [3]	52.4	68.0	73.4	87.9	93.7	96.1	95.7	96.0	96.5	96.7	97.5	97.7	97.7	97.3	97.7	89.6
TENT [5]	53.5	77.8	91.1	96.0	97.0	97.6	97.4	97.6	97.9	98.1	98.1	98.0	98.1	97.9	98.1	92.9
LAME [1]	74.8	48.2	**21.1**	**30.6**	43.4	**32.5**	**21.6**	**23.0**	39.6	**33.3**	72.7	66.5	51.5	**20.7**	**27.5**	40.5
CoTTA [6]	49.3	55.1	49.1	52.9	56.8	55.7	49.5	50.0	53.6	53.4	54.9	53.9	53.8	50.1	53.5	52.8
NOTE [2]	52.2	64.9	47.5	57.0	61.9	67.3	60.4	67.8	77.4	90.6	97.1	96.8	92.8	95.9	96.6	75.1
RoTTA	**36.4**	**44.4**	29.7	36.5	**41.0**	44.1	26.8	29.5	**33.0**	40.3	**40.3**	**38.2**	**33.9**	28.5	34.9	**35.8**(+4.7)

Table S. Average classification error of the task CIFAR100 → CIFAR100-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	t ⟶															
Method	motion	snow	fog	shot	defocus	contrast	zoom	brightness	frost	elastic	glass	gaussian	pixelate	jpeg	impulse	Avg.
Source	30.8	39.5	50.3	68.0	29.3	55.1	28.8	29.5	45.8	37.2	54.1	73.0	74.7	41.2	39.4	46.4
BN [4]	48.5	54.0	58.9	56.2	46.4	48.0	47.0	45.4	52.9	53.4	57.1	58.2	51.7	57.1	58.8	52.9
PL [3]	50.6	62.1	73.9	87.8	90.8	96.0	94.8	96.4	97.4	97.2	97.4	97.4	97.3	97.4	97.4	88.9
TENT [5]	53.3	77.6	93.0	96.5	96.7	97.5	97.1	97.5	97.3	97.2	97.1	97.7	97.6	98.0	98.3	92.8
LAME [1]	**22.4**	**30.4**	43.9	66.3	**21.3**	51.7	**20.6**	**21.8**	39.6	**28.0**	48.7	72.8	74.6	**33.1**	**32.3**	40.5
CoTTA [6]	49.2	52.7	56.8	53.0	48.7	51.7	49.4	48.7	52.5	52.2	54.3	54.9	49.6	53.4	56.2	52.2
NOTE [2]	45.7	53.0	58.2	65.6	54.2	52.0	59.8	63.5	74.8	91.8	98.1	98.3	96.8	97.0	98.2	73.8
RoTTA	31.8	36.7	**40.9**	**42.1**	30.0	**33.6**	27.9	25.4	**32.3**	34.0	**38.8**	**38.7**	**31.3**	38.0	42.9	**35.0**(+5.5)

Table T. Average classification error of the task CIFAR100 → CIFAR100-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	t ⟶															
Method	frost	impulse	jpeg	contrast	zoom	glass	pixelate	snow	defocus	motion	brightness	elastic	shot	fog	gaussian	Avg.
Source	45.8	39.4	41.2	55.1	28.8	54.1	74.7	39.5	29.3	30.8	29.5	37.2	68.0	50.3	73.0	46.4
BN [4]	52.9	58.8	57.6	48.2	47.4	57.6	50.9	52.4	47.0	47.2	45.1	54.0	56.4	57.7	58.2	52.8
PL [3]	56.9	73.3	86.7	94.4	95.8	97.3	97.2	97.4	97.6	97.4	97.7	97.6	97.8	98.3	98.1	92.2
TENT [5]	60.1	84.2	95.7	97.2	97.4	97.9	97.8	98.0	98.1	98.2	98.3	98.4	98.4	98.4	98.4	94.4
LAME [1]	**39.9**	**32.4**	**33.4**	51.4	**20.6**	49.0	74.4	**31.3**	**21.2**	**22.6**	**21.9**	**28.1**	66.9	43.9	72.5	40.6
CoTTA [6]	51.5	55.3	54.3	51.8	49.4	55.3	50.7	54.2	51.4	50.6	49.5	53.6	55.0	57.1	55.8	53.0
NOTE [2]	51.6	60.9	60.3	45.4	54.3	70.8	68.8	75.0	75.7	87.1	94.7	95.6	96.7	96.4	97.2	75.4
RoTTA	40.0	46.3	42.8	**36.4**	29.2	**42.3**	**33.2**	34.4	28.4	29.2	26.4	34.5	**38.5**	**39.8**	**39.3**	**36.0**(+4.6)

Table U. Average classification error of the task CIFAR100 → CIFAR100-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	t ⟶															
Method	defocus	motion	zoom	shot	gaussian	glass	jpeg	fog	contrast	pixelate	frost	snow	brightness	elastic	impulse	Avg.
Source	29.3	30.8	28.8	68.0	73.0	54.1	41.2	50.3	55.1	74.7	45.8	39.5	29.5	37.2	39.4	46.4
BN [4]	47.1	48.6	47.8	56.2	57.6	57.6	57.6	57.5	48.7	50.6	51.8	53.2	46.9	53.5	58.8	52.9
PL [3]	48.8	58.7	69.9	88.0	95.1	96.6	96.7	96.9	97.4	97.4	98.2	98.2	98.2	98.3	98.5	89.1
TENT [5]	51.0	67.6	85.8	95.9	97.2	97.5	97.2	97.7	98.1	97.9	97.7	97.7	98.0	98.0	98.2	91.7
LAME [1]	**21.2**	**22.8**	**21.1**	66.3	72.8	49.0	**33.3**	44.8	51.7	74.9	39.8	**31.2**	**21.3**	**27.3**	**32.3**	40.6
CoTTA [6]	48.4	48.8	48.2	52.9	54.0	53.8	52.7	57.2	52.6	48.6	51.8	53.9	49.4	52.3	56.0	52.0
NOTE [2]	45.1	46.7	49.1	67.3	65.5	69.4	75.5	80.3	83.8	96.0	97.6	97.1	96.1	97.9	98.7	77.7
RoTTA	29.6	31.3	28.8	**43.9**	**41.5**	**41.3**	40.9	**39.8**	**32.1**	**32.6**	**33.1**	33.0	26.5	34.5	42.9	**35.4**(+5.2)

Table V. Average classification error of the task CIFAR100 → CIFAR100-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	t ⟶															
Method	glass	zoom	impulse	fog	snow	jpeg	gaussian	frost	shot	brightness	contrast	motion	pixelate	defocus	elastic	Avg.
Source	54.1	28.8	39.4	50.3	39.5	41.2	73.0	45.8	68.0	29.5	55.1	30.8	74.7	29.3	37.2	46.4
BN [4]	58.8	47.7	59.2	57.6	52.7	56.9	58.2	52.0	56.7	45.5	47.8	48.2	51.7	46.1	54.0	52.9
PL [3]	60.1	59.5	75.1	85.7	91.5	94.6	96.5	97.1	97.4	97.3	98.0	97.7	97.9	97.8	97.7	89.6
TENT [5]	61.6	71.5	91.0	95.9	96.6	97.1	96.9	97.3	97.4	97.2	97.9	98.0	98.1	97.9	97.8	92.8
LAME [1]	48.6	**20.6**	**32.3**	44.4	**30.2**	**33.6**	72.4	40.0	66.3	**21.6**	52.0	**22.8**	74.6	**20.7**	**27.5**	40.5
CoTTA [6]	56.4	48.9	56.1	57.8	54.1	54.2	56.2	53.6	55.4	50.0	53.6	51.6	51.2	50.7	54.4	53.6
NOTE [2]	62.5	46.3	61.5	61.1	58.6	68.4	76.1	78.3	92.0	93.4	96.1	95.4	96.2	95.8	96.4	78.5
RoTTA	**45.5**	30.0	45.9	**42.6**	35.3	41.8	**42.2**	**34.5**	**40.2**	27.3	**31.3**	30.2	**32.7**	28.1	34.9	**36.2**(+4.3)

Table W. Average classification error of the task CIFAR100 → CIFAR100-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	t ⟶															
Method	contrast	gaussian	defocus	zoom	frost	glass	jpeg	fog	pixelate	elastic	shot	impulse	snow	motion	brightness	Avg.
Source	55.1	73.0	29.3	28.8	45.8	54.1	41.2	50.3	74.7	37.2	68.0	39.4	39.5	30.8	29.5	46.4
BN [4]	49.5	58.8	47.0	46.5	52.2	57.6	57.6	57.6	51.7	53.5	56.0	58.5	53.1	47.6	46.3	52.9
PL [3]	53.6	70.4	76.0	85.1	91.2	95.2	96.0	97.0	96.9	97.3	97.3	97.6	97.5	97.6	97.7	89.8
TENT [5]	60.2	89.1	95.0	96.2	96.9	97.0	96.5	97.0	97.0	97.2	97.6	97.8	97.5	97.9	97.7	94.0
LAME [1]	51.3	72.5	**21.5**	**21.0**	39.6	49.0	**33.3**	44.8	74.8	**28.0**	66.8	**32.5**	**30.6**	**22.5**	**21.4**	40.6
CoTTA [6]	52.3	55.3	49.5	48.1	52.1	54.8	52.7	56.9	50.6	52.6	53.7	55.8	54.6	50.6	50.5	52.7
NOTE [2]	**39.1**	64.7	48.9	50.6	59.1	70.1	71.7	75.0	85.2	95.7	96.9	98.4	96.0	95.9	94.9	76.1
RoTTA	41.4	**46.2**	30.5	28.5	**36.0**	**40.9**	40.5	**39.6**	**33.0**	35.0	**38.2**	43.1	33.9	30.7	27.1	**36.3**(+4.3)