Learning from Unknown for Open-Set Test-Time Adaptation

Supplementary Material

Taki Hasan Rafi[1] Amit Agarwal[2] Hitesh L. Patel[2] Dong-Kyu Chae[1*]
[1]Hanyang University, Seoul, South Korea
[2]Oracle AI, USA
{takihr, dongkyu}@hanyang.ac.kr

1. Evaluation Baselines

We mainly focused on comparing our proposed method with three types of other methods. (1) Entropy-free method: the source model trained with clean datasets is directly tested under an open-set setting. (2) Entropy-based/continual TTA methods: **TENT** [7] estimates the normalization statistics and optimizes the model parameters based on entropy minimization. **CoTTA** [8] adopts the mean-teacher method to improve pseudo labels and provide a weighted average of these labels to mitigate error accumulation. It also introduces a stochastic restoration module to enforce continual adaptation by avoiding catastrophic forgetting. **EATA** [5] reduces the effect of noisy samples with high entropy by employing an active sample selection criterion. To alleviate the issue of forgetting, they introduce a Fisher regularizer to constrain model parameters. (3) Open-set TTA methods: **OSTTA** [4] uses a filtering technique based on the confidence values of the adopted model compared with the original source model, where low confidence samples appear to be noisy. **UniEnt** [2] uses entropy minimization and maximization with a distribution-aware filtering method for both covariate shifted in-and out-of-distribution samples. Furthermore, **UniEnt+** [2] alleviates the noisy samples by leveraging sample-level confidence. Lastly, **Stabilized OSTTA** [3] uses an auxiliary filtering method to validate data from the primary filtering mechanism and also employs knowledge-integrated prediction to calibrate the output of the adopted model.

2. More Results

Additional Results on CIFAR Benchmarks. We perform additional experiments with Places365-C [10] and Texture-C [1] datasets. We follow the same setup and evaluation metric from [3], we further add harmonic mean (H-S) of accuracy and AUROC. In Tab. 1, we demonstrate the performance with CIFAR-10-C by adding different open-set environments. Our method consistently outperforms S-OSTTA method in all metrics. Existing open-set TTA methods exhibit considerable performance, but lack achieving higher performance compared to our method. But S-OSTTA performed closely with our method, but the margin is significant. On the other hand, in Tab. 2, we observe a similar trend as our method outperforms other methods with CIFAR-100-C benchmark as well. Similarly, S-OSTTA achieved the second best score in all metrics.

Additional Results on Tiny-Imagenet Benchmark. In Tab. 3, we perform experiment with Places365-C [10] and Texture-C [1] datasets as open-set environment and Tiny-ImageNet-C as the close-set environment. Tiny-Imagenet-C poses more challenging tasks as it has 200 classes. In both open-set datasets, our method significantly performs other methods, and demonstrates its capability to handle open-set environments.

Performance under Continual Settings. We follow a similar setup [6, 8]. In standard TTA setting, corruption types change abruptly in the highest severity level (e.g. 1-5), where 1 is the lowest and 5 is the highest. However, in continual setting, we experiment this severity level under a sequence by gradually changing severity for the 15 different corruption types. And the we change the corruption types gradually from lowest to highest, so that the distribution shift within each corruption is also gradual. Following previous method [8], we randomly shuffle 10 different corruption types then report average error rate over ten different sequences, shown in Tab. 4. We can see, our method outperforms both TTA and OSTTA settings by a significant margin in CIFAR-10-C dataset.

References

[1] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the

*Corresponding author.

Table 1. Results of different methods on CIFAR-10-C benchmark. ↑ indicates that larger values are better. All values are percentages. We present Source , TTA methods , OSTTA methods , and Our method respectively. We underline the second best score, and best scores are in **bold**. Improvements (±) compared to the second best score are also presented.

Method	Places365-C			Textures-C		
	Acc↑	AUROC↑	H-S↑	Acc↑	AUROC↑	H-S↑
Source [9]	82.46	83.32	82.89	82.46	82.51	82.48
TENT [7]	55.31	54.23	54.76	70.23	68.45	69.33
CoTTA [8]	84.67	82.34	83.49	84.10	79.78	81.88
EATA [5]	84.78	80.35	82.51	81.87	78.24	80.01
OSTTA [4]	84.56	76.45	80.30	81.56	69.43	75.01
UniEnt [2]	84.78	88.64	86.67	82.75	84.43	83.58
UniEnt+ [2]	84.56	89.57	86.99	80.45	89.65	84.80
S-OSTTA [3]	88.23	94.12	91.08	87.56	97.51	92.27
Ours	**90.41**$_{(+1.85)}$	**94.78**$_{(+0.66)}$	**92.54**$_{(+1.46)}$	**90.43**$_{(+2.87)}$	**98.74**$_{(+1.23)}$	**94.40**$_{(+2.13)}$

Table 2. Results of different methods on CIFAR-100-C benchmark. ↑ indicates that larger values are better. All values are percentages.

Method	Places365-C			Textures-C		
	Acc↑	AUROC↑	H-S↑	Acc↑	AUROC↑	H-S↑
Source [9]	53.45	65.34	58.92	53.45	62.65	57.55
TENT [7]	26.45	60.10	36.86	29.80	61.56	40.49
CoTTA [8]	55.77	72.81	63.51	51.45	67.68	58.47
EATA [5]	53.90	71.35	61.47	50.45	58.29	53.28
OSTTA [4]	60.32	72.65	66.21	58.76	65.30	61.84
UniEnt [2]	59.39	77.19	67.33	57.78	73.43	64.64
UniEnt+ [2]	58.76	78.67	67.31	56.45	73.89	64.42
S-OSTTA [3]	62.78	85.41	72.07	62.10	93.23	75.17
Ours	**64.32**$_{(+1.54)}$	**88.67**$_{(+3.26)}$	**74.93**$_{(+2.86)}$	**65.23**$_{(+3.13)}$	**95.78**$_{(+2.55)}$	**78.03**$_{(+2.86)}$

Table 3. Results of different methods on Tiny-ImageNet benchmark. ↑ indicates that larger values are better. All values are percentages.

Method	Places365-C			Textures-C		
	Acc↑	AUROC↑	H-S↑	Acc↑	AUROC↑	H-S↑
Source [9]	28.24	67.69	40.05	28.24	71.67	40.49
TENT [7]	40.78	65.76	50.38	34.87	46.67	39.87
CoTTA [8]	41.56	72.45	53.19	56.46	72.45	63.49
EATA [5]	44.32	77.34	56.23	42.87	65.23	51.67
OSTTA [4]	47.67	75.24	58.37	45.72	60.23	51.34
UniEnt [2]	46.87	78.25	58.57	44.45	64.72	51.96
UniEnt+ [2]	45.23	78.13	57.44	44.32	63.52	51.67
S-OSTTA [3]	48.24	84.08	61.23	47.89	82.80	60.45
Ours	**50.76**$_{(+2.52)}$	**86.87**$_{(+2.79)}$	**64.04**$_{(+2.81)}$	**49.90**$_{(+2.01)}$	**85.40**$_{(+2.60)}$	**63.84**$_{(+3.39)}$

wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 1

[2] Zhengqing Gao, Xu-Yao Zhang, and Cheng-Lin Liu. Unified entropy optimization for open-set test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23975–23984, 2024. 1, 2, 3

[3] Byung-Joon Lee, Jin-Seop Lee, and Jee-Hyong Lee. Stabilizing open-set test-time adaptation via primary-auxiliary filtering and knowledge-integrated prediction. *arXiv preprint arXiv:2508.18751*, 2025. 1, 2, 3

[4] Jungsoo Lee, Debasmit Das, Jaegul Choo, and Sungha Choi. Towards open-set test-time adaptation utilizing the wisdom of crowds in entropy minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16380–16389, 2023. 1, 2, 3

[5] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient

Table 4. Experiments on CIFAR-10-to-CIFAR-10-C by gradually changing. The severity level changes from lowest to highest and the corruption type changes when the severity level is lowest. Results are presented in mean over 10 randomly shuffled corruption types. Lower is better.

Avg. Error (%)↓	**Source**	**TENT [7]**	**CoTTA [8]**	**OSTTA [4]**	**UniEnt [2]**	**UniEnt+ [2]**	**S-OSTTA [3]**	**Ours**
CIFAR-10-C	26.5	33.6	12.2	24.6	11.3	11.2	9.4	**8.1**

test-time model adaptation without forgetting. In *International conference on machine learning*, pages 16888–16905. PMLR, 2022. 1, 2

[6] Taki Hasan Rafi, Amit Agarwal, Hitesh L. Patel, and Dong-Kyu Chae. Towards robust continual test-time adaptation via neighbor filtration. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 5161–5165. ACM, 2025. 1

[7] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 1, 2, 3

[8] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 1, 2, 3

[9] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 2

[10] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 1