On the Robustness of Open-World Test-Time Training: Self-Training with Dynamic Prototype Expansion

South China University of Technology

Agency for Science, Technology and Research SINGAPORE

Yushu Li[1], Xun Xu[2†], Yongyi Su[1], Kui Jia[1†]

[1]South China University of Technology, [2]I2R, A*STAR

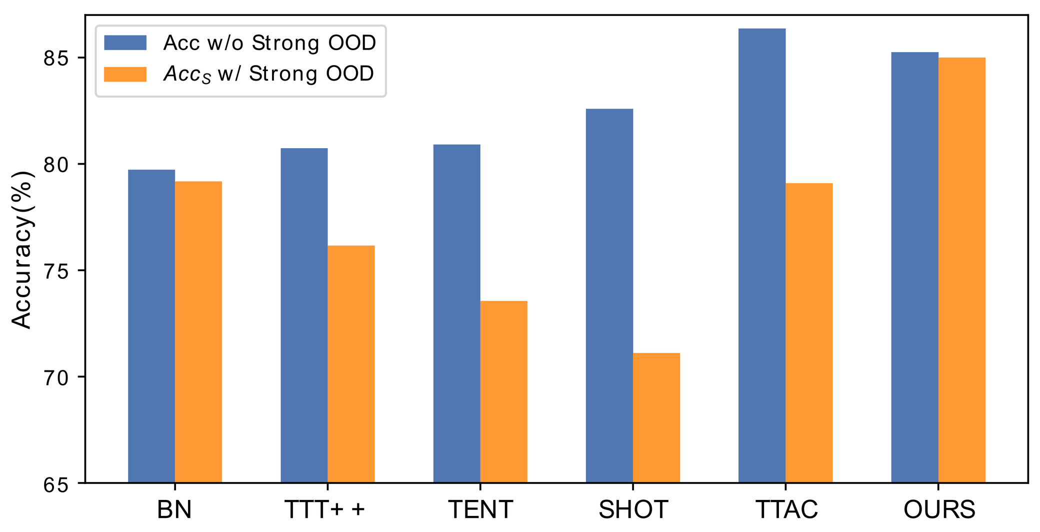
Project page: https://yushu-li.github.io/owttt-site/

Motivation

The existing TTT method only takes into account covariate shift (weak OOD) and does not consider semantic changes (strong OOD). When facing an open-world environment, existing methods primarily encounter two issues:

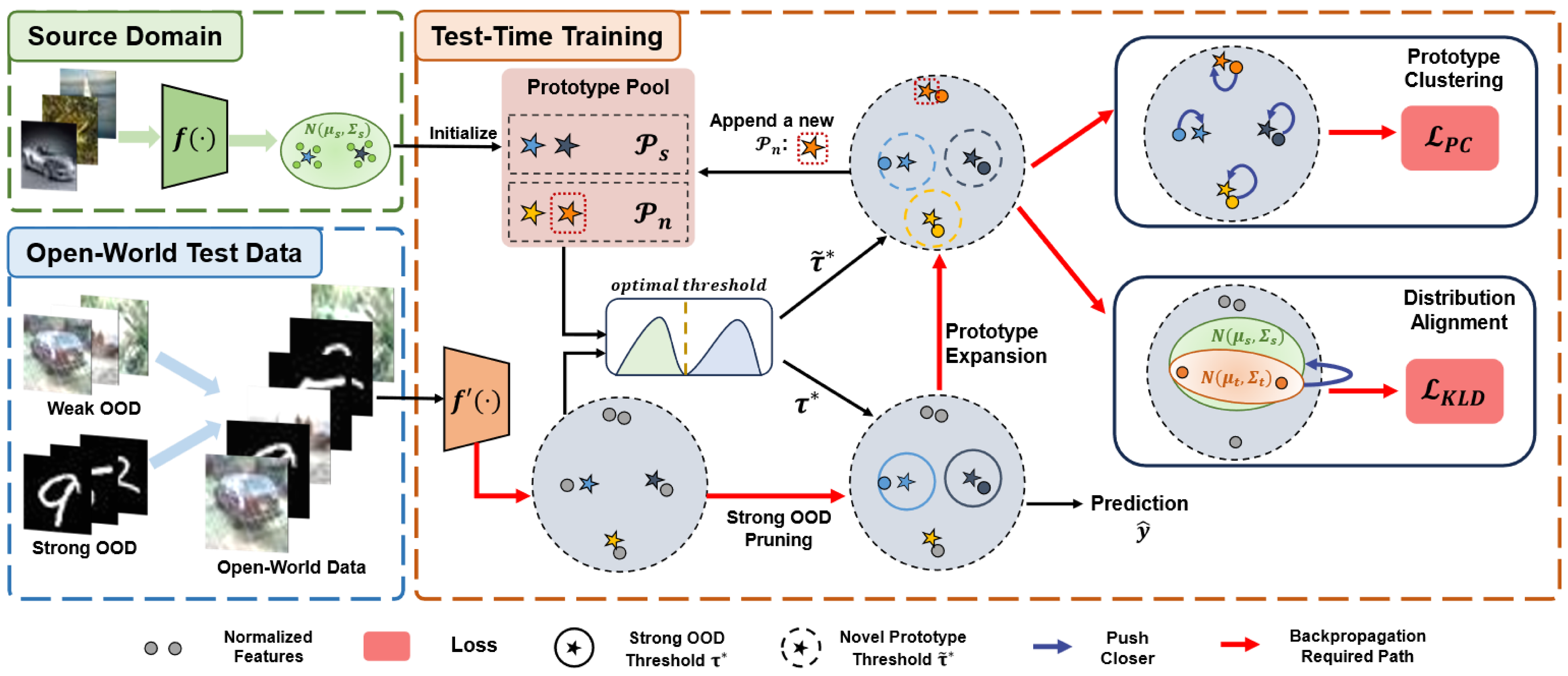
- **Fail to distinguish and reject strong OOD**
- **Severe performance drop on weak OOD**

☐ Performance w/o & w/ strong OOD



Method

☐ An overview of the proposed method



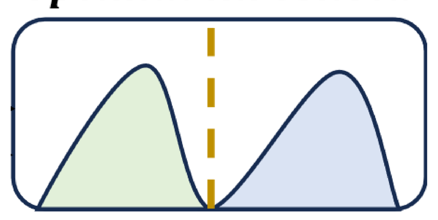
- Prototype Clustering

$$\mathcal{L}_{PC} = -\sum_{k \in \mathcal{C}_s} \mathbb{1}(\hat{y} = k) \log \frac{\exp(\frac{<p_k, z_i>}{\delta})}{\sum_l \exp(\frac{<p_l, z_i>}{\delta})}$$

- Hyper-parameter-free strong OOD detector & prototype expansion

$$os_i = 1 - \max_{p_k \in \mathcal{P}_s} < f(x_i), p_k >$$

optimal threshold



- Distribution Alignment

$$\mathcal{L}_{KLD} = D_{KL}(\mathcal{N}(\mu_s, \Sigma_s) || \mathcal{N}(\mu_t, \Sigma_t))$$

- Total Loss

$$\mathcal{L}_{total} = \mathcal{L}_{PC} + \lambda \mathcal{L}_{KLD}$$

Contribution

- Overlooked by existing studies into test-time training, we argue that **open-world test-time training (OWTTT)** could be spoiled by strong OOD testing data.
- We introduce a baseline method by **prototype clustering** with distribution alignment regularization. A **strong OOD detector** and **prototype expansion** are further developed to improve the robustness of the baseline under OWTTT protocol.
- We established a **benchmark** for evaluating OWTTT protocol covering multiple types of domain shift, including common corruptions and style transfer. Our approach achieves **state-of-the-art** performance on the proposed benchmark.

Empirical Results

☐ Evaluation Metric

- Acc_S: Accuracy of the weak OOD samples.
- Acc_N: Accuracy of the strong OOD samples.
- Acc_H: Harmonic mean between Acc_S and Acc_N.

$$Acc_H = 2 \cdot \frac{Acc_S \cdot Acc_N}{Acc_S + Acc_N}$$

☐ ImageNet-C performance

Method	noise			MNIST			SVHN		
	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H
TEST	18.51	**100.00**	31.24	18.66	**98.27**	31.36	18.94	87.75	31.15
BN	36.34	99.97	53.31	30.77	74.53	43.55	33.26	84.54	47.74
TENT	22.54	10.47	14.29	27.53	10.01	14.68	41.16	45.51	43.22
SHOT	**46.79**	**100.00**	**63.75**	27.47	55.25	36.70	34.00	75.94	46.97
TTAC	42.60	94.52	58.73	30.43	72.11	42.80	31.59	74.07	44.29
OURS	41.40	**100.00**	58.56	**38.86**	93.35	**54.87**	**38.60**	**98.06**	**55.40**

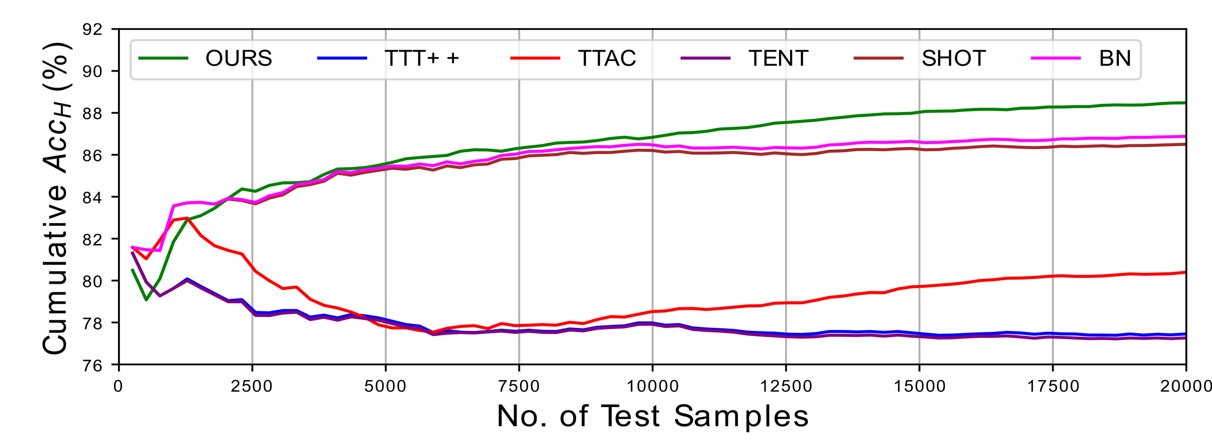
☐ Ablation Study

O.D.	P.C.	P.E.	D.A.	Noise			SVHN			CIFAR100-C		
				Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H
-	-	-	-	70.6	0.0	0.0	70.6	0.0	0.0	70.6	0.0	0.0
✓	-	-	-	68.6	100.0	81.4	60.9	86.4	71.5	52.7	74.2	61.7
✓	$\mathcal{P}_s$	-	-	65.2	91.5	76.1	60.9	90.0	72.7	56.3	69.0	62.0
✓	$\mathcal{P}_s$+$\mathcal{P}_n$	✓	-	68.7	99.8	81.4	65.3	95.0	77.4	52.6	78.9	63.2
✓	-	-	✓	72.9	88.8	80.1	78.1	88.0	82.8	70.5	78.7	74.4
✓	$\mathcal{P}_s$+$\mathcal{P}_n$	✓	✓	85.5	98.6	91.6	85.0	87.9	86.4	74.1	84.6	79.0

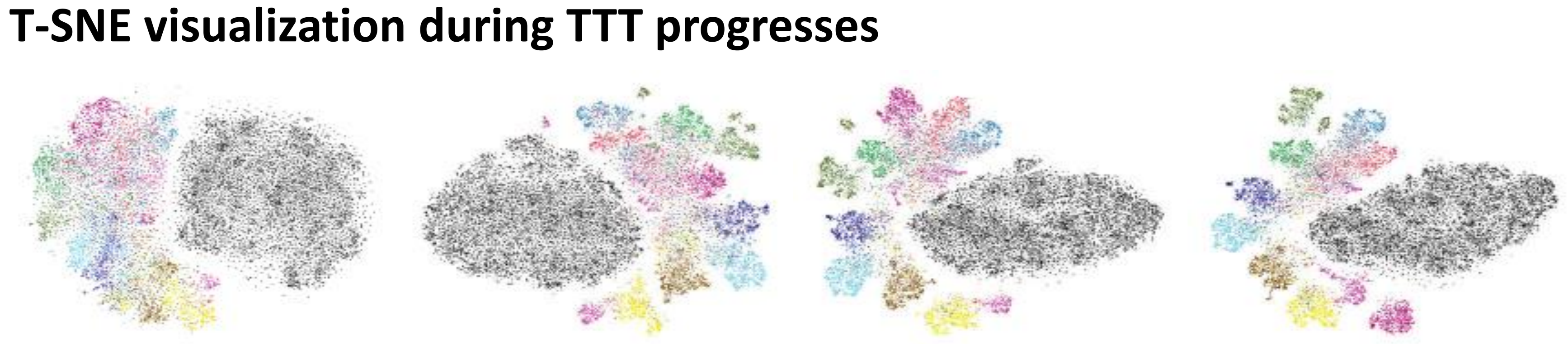
☐ OWTTT results on CIFAR10-C. All numbers are in %.

Method	Noise			MNIST			SVHN			Tiny-ImageNet			CIFAR100-C		
	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H
TEST	68.59	**99.97**	81.36	60.48	88.81	71.96	60.94	86.44	71.48	57.41	79.63	66.72	52.74	74.24	61.67
BN	76.63	95.69	85.11	76.15	95.75	84.83	79.18	**94.71**	86.25	67.66	82.67	74.42	68.44	81.38	74.35
TTT++	41.09	57.31	47.86	59.52	77.52	67.34	68.77	85.80	76.34	66.70	79.28	72.44	65.69	77.47	71.10
TENT	32.24	33.30	32.77	55.64	68.27	61.31	66.70	82.50	73.77	66.54	79.32	72.37	64.80	76.40	70.12
SHOT	63.54	71.37	67.23	56.92	53.26	55.03	70.01	72.58	71.27	67.78	82.25	74.32	67.73	72.87	70.21
TTAC	64.46	77.42	70.35	77.60	84.53	80.92	77.30	81.10	79.16	71.64	77.14	74.29	71.94	75.44	73.65
OURS	**85.46**	98.60	**91.56**	**83.89**	**97.83**	**90.32**	**84.99**	87.94	**86.44**	**71.77**	**84.71**	**77.70**	**74.08**	**84.64**	**79.01**

☐ Cumulative performance during TTT progresses



☐ T-SNE visualization during TTT progresses



(a) 0% (b) 33% (c) 67% (d) 100%

Web Page

PDF

Try it!