7. Appendix

To further demonstrate the effectiveness and robustness of our method, we perform additional experiments and analyses.

1. First, we present the details of the segmentation task in Section 7.2.
2. Next, we evaluate the robustness of our method under various settings, including imbalanced csID and csOOD samples in Section 7.3, the absence of csOOD samples in Section 7.4, and varying numbers of unknown classes within each epoch in Section 7.5.
3. We further analyze the use of the l_1-norm and an l_1-norm-based OOD score NAN [32], as described in Section 7.7. Our findings reveal that the l_1-norm for csID samples is initially smaller than that for csOOD samples at the beginning of adaptation, which adversely affects OOD detection performance during the early stages. The underlying mechanism behind this unexpected phenomenon warrants further investigation. Finally, we visualize the feature distribution using t-SNE in Section 7.8.

7.1. Prototype Update Stability

The exponential moving average (EMA) update used for prototype maintenance is a standard estimator for time-varying means. Under bounded feature drift and finite variance assumptions, the update is stable and tracks the moving target with a small bias–variance trade-off controlled by the momentum α [13, 20].

7.2. Implementation Details of Segmentation Task

In both the Cityscapes and ACDC datasets, pixels are categorized as either classed or void. Classified pixels correspond to ID in Cityscapes and csID in ACDC, while void pixels are treated as inherent OOD, respectively. To integrate our method into a semantic segmentation model, we pass the pixel-wise classification results through the distribution-aware filter by computing the entropy of each pixel. Since the distribution-aware filter produces labels at the pixel level, our method can be seamlessly incorporated into the semantic segmentation adaptation pipeline.

7.3. Imbalanced csID and csOOD Samples

The main experiments use balanced batches of 100 csID and 100 csOOD samples. To assess robustness under class imbalance, we further evaluate on CIFAR-10-C with csOOD-to-csID ratios ranging from 0.2 to 1.0. As reported in Table 6, our model demonstrates robust performance on imbalanced inputs, achieving better mean performance across all ratios.

Method	0.2	0.4	0.6	0.8	1.0	$\Delta\downarrow$	Mean
Source	40.00	40.03	39.98	39.92	39.87	**0.16**	39.96
BN Adapt	49.91	49.55	48.92	47.97	47.10	2.81	48.69
TENT	47.68	44.12	44.06	42.90	42.16	5.52	44.18
+ UniEnt	56.84	57.48	57.13	56.77	56.26	1.22	56.90
+ UniEnt+	**57.15**	**57.59**	57.24	56.88	56.33	1.26	57.04
+ Ours	56.63	57.51	**57.51**	**57.13**	**56.80**	0.88	**57.12**

Table 6. OSCR performance under different csOOD-to-csID imbalance ratios on CIFAR-100-C. The Δ column indicates the performance variation, calculated as the difference between the maximum and minimum values across the ratios. ROSETTA achieves the best overall accuracy and the most consistent performance across all imbalance settings.

7.4. Robustness to the Absence of csOOD Samples

In our setting, we assume the target domain contains two distinct distributions (csID and csOOD). However, csOOD samples may not always be present in practical deployment scenarios. To evaluate the robustness in such cases, we evaluate the model on purely csID target data. As shown in Table 7, our method outperforms the state-of-the-art baselines UniEnt and UniEnt+ on all four benchmarks. We attribute this advantage to eliminating entropy maximization. State-of-the-art methods apply entropy maximization to input without csOOD, driving the model toward uncertain predictions and degrading classification accuracy. By instead optimizing angular and logit-norm, ROSETTA preserves csID performance when no csOOD data are present and outperforms UniEnt by 2.51% on ImageNet-C.

Method	Cifar-10-C	Cifar-100-C	Tiny-ImageNet-C	ImageNet-C
TENT	**85.85**	**62.86**	**33.13**	**45.93**
+UniEnt	83.26 (-2.59)	60.01 (-2.85)	36.60 (+3.47)	44.48 (-1.45)
+UniEnt+	83.93 (-1.92)	59.73 (-3.13)	35.98 (+2.85)	39.96 (-5.97)
+Ours	**84.01 (-1.84)**	**60.39 (-2.47)**	**36.62 (+3.49)**	**46.99 (+1.06)**

Table 7. Experiments without csOOD data. Our method is more robust than UniEnt and UniEnt+, which alternate in yielding their highest accuracy. Tent occasionally achieves higher Acc since it does not account for csOOD.

7.5. Varying Numbers of Unknown Classes

To evaluate the robustness of ROSETTA across varying levels of OOD dataset difficulty, we conduct experiments using CIFAR-10-C as csID and SVHN-C as csOOD, with different numbers of unknown classes. The number of classes in SVHN-C ranges from 2 to 10, representing progressively easier to more challenging csOOD datasets. As shown in Table 8, ROSETTA demonstrates consistent robustness across different numbers of unknown classes while achieving superior overall performance.

Method	2	4	6	8	10	Δ ↓	Mean
Source	70.84	69.28	69.32	69.18	68.44	2.40	69.41
BN Adapt	72.56	72.48	72.52	72.44	72.14	0.42	72.43
TENT	49.51	48.29	51.74	49.53	50.97	3.45	50.01
+ UniEnt	78.71	78.39	78.28	78.13	77.82	0.89	78.27
+ UniEnt+	78.65	78.23	78.23	78.07	77.68	0.97	78.17
+ Ours	**81.50**	**81.37**	**81.84**	**81.87**	**81.37**	**0.50**	**81.59**

Table 8. OSCR performance with increasing numbers of unknown classes in the csOOD set. As the number of unknown classes grows from 2 to 10, the task becomes more challenging. ROSETTA maintains consistently strong results and shows minimal performance degradation, as reflected by the lowest Δ value.

7.6. Computational Cost of ROSETTA

To evaluate the computational cost introduced by ROSETTA, we measure the average running time per batch of our method and AEO [11]. The experiments are conducted using HAC as the csID dataset and EPIC-KITCHENS as the OOD dataset, adapting from the human domain to the animal domain. Each experiment is repeated three times, and we report the average results.

The results show that the average running time per batch for AEO is 2.29 ± 0.026 seconds, while our method takes 2.52 ± 0.0047 seconds. This indicates that ROSETTA introduces only a marginal increase in computational cost.

7.7. Feature Norm as OOD Detection Score

We conducted experiments using l_1-norm and NAN [32] score as alternatives to the energy score for OOD detection on CIFAR-10-C, CIFAR-100-C and Tiny-ImageNet-C with ROSETTA based on TENT [35], as shown in Table. 9. Results on CIFAR-10-C suggest that l_1-norm is effective for OOD detection, improving FPR95 by 4.62. However, as dataset complexity and model size increase, l_1-norm shows decreased performance compared to the energy score, particularly on Tiny-ImageNet-C.

Our analysis reveal that initially, AUROC for l_1-norm-based detection is below 50, as shown in Table 10, indicating a tendency to misclassify csID as csOOD. Further investigation shows this anomaly may arise from model behavior: while OOD samples are generally expected to have lower l_1-norms than ID samples, csID samples initially exhibited lower norms than csOOD samples at the start of adaptation. Consequently, smaller models (e.g., WideResNet) adapt to covariate-shifted data faster than larger models (e.g., ResNet50), resulting in rapid AUROC gains on CIFAR-10-C but slower improvements on Tiny-ImageNet-C. The reason for csID samples having lower feature norms than csOOD samples at the start of adaptation warrants further study.

Dataset	OOD Score	Acc↑	AUROC↑	FPR95↓	OSCR↑
CIFAR-10-C	Energy	84.34	93.75	29.91	**81.37**
	l_1-norm	84.34	**94.01**	**25.19**	81.04
	NAN	84.34	92.87	35.88	80.50
CIFAR-100-C	Energy	59.20	91.80	35.89	**56.80**
	l_1-norm	59.20	**92.40**	**29.89**	56.01
	NAN	59.20	91.95	31.91	55.88
Tiny-ImageNet-C	Energy	37.32	**64.01**	**88.53**	**29.75**
	l_1-norm	37.32	61.54	89.01	27.61
	NAN	37.32	61.45	89.83	27.57

Table 9. OOD detection performance comparison across different OOD scores on CIFAR-10-C, CIFAR-100-C, and Tiny-ImageNet-C datasets.

Corruption	Acc↑	AUROC↑	FPR95↓	OSCR↑
Gaussian noise	37.99	48.25	95.41	18.54
Shot noise	38.83	52.32	93.90	24.32
Impulse noise	30.75	51.41	94.62	20.19
Defocus blur	38.89	58.68	92.09	27.52
Glass blur	28.96	58.14	91.67	20.76
Motion blur	43.60	65.47	88.92	34.01
Zoom blur	45.42	65.79	88.45	35.49
Snow	37.63	64.07	88.44	29.12
Frost	39.66	64.41	86.94	30.53
Fog	33.68	61.65	89.62	25.69
Brightness	41.78	66.21	86.55	33.23
Contrast	13.11	57.79	89.30	8.90
Elastic	39.17	66.30	86.10	30.85
Pixelate	45.06	71.57	81.01	37.60
JPEG	45.26	71.07	82.11	37.46
Mean	37.32	61.54	89.01	27.61

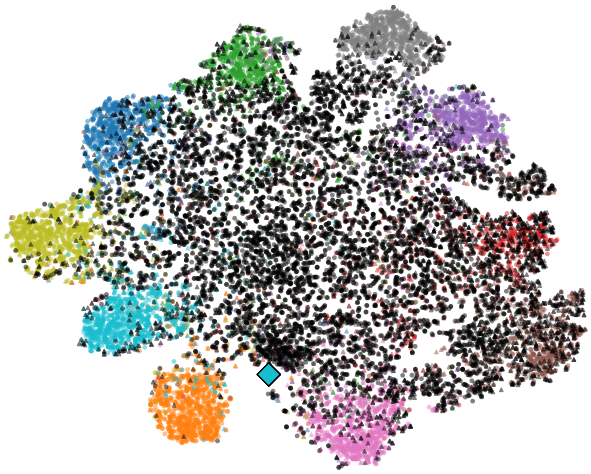
Table 10. Performance metrics for each task using l_1-norm as the OOD score of our method on Tiny-ImageNet-C, ordered by task sequence during adaptation. An AUROC below 50 indicates a tendency to classify OOD samples as ID more frequently than random guessing.

7.8. T-SNE Visualization

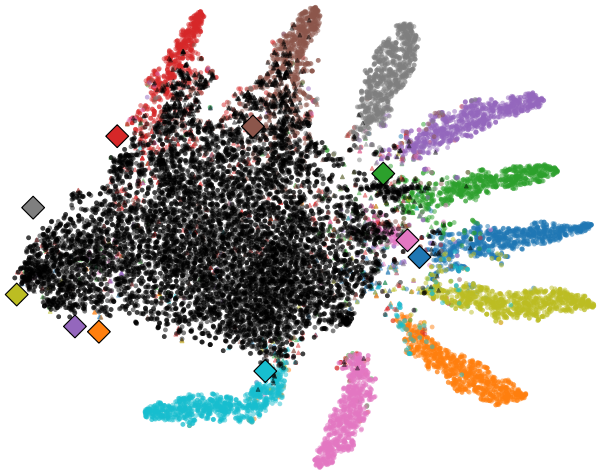
To further illustrate the effectiveness of our loss function, we visualize the feature vectors of CIFAR-10-C test samples with SVHN-C test samples as csOOD using t-SNE [34], as shown in Fig. 8a, 8b, 8c, 8d. In the t-SNE plot, black triangles indicate misclassified csOOD samples, black circles represent correctly detected csOOD samples, colored triangles denote misclassified csID samples, and colored circles represent correctly classified csID samples. Colored stars mark the embeddings of the updated class prototypes, and colored diamonds represent the fixed weights of the linear classification layer for each class.

In Fig. 8a, csOOD and csID samples are highly intermixed when only the entropy-minimizing loss is applied, highlighting the difficulty of detecting csOOD samples. By comparing Fig. 8a with Fig. 8c, we observe that the introduction of $\mathcal{L}_{\text{ang}}$ limits feature norm growth, thereby making

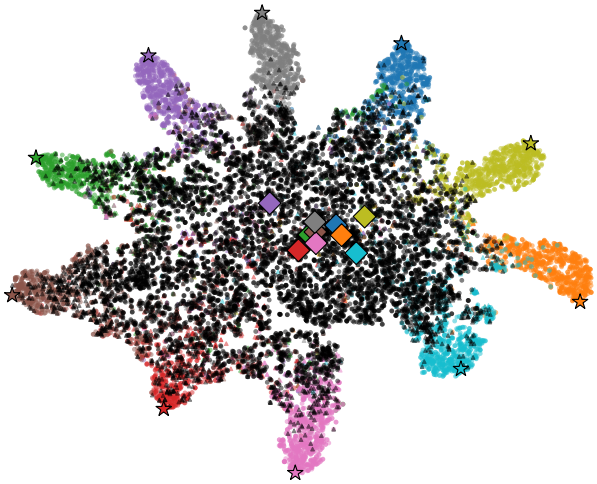
the fixed classification weights more distinguishable. This addition of $\mathcal{L}_{\text{ang}}$ reduces the norm of csOOD features and increases the cosine similarity between csID samples and their updated class prototypes, thereby improving csOOD detection. Furthermore, when comparing Fig. 8b with Fig. 8d, our method achieves a clearer separation between csID and csOOD samples, which is not only beneficial for csID classification but also for csOOD detection in open-set TTA.



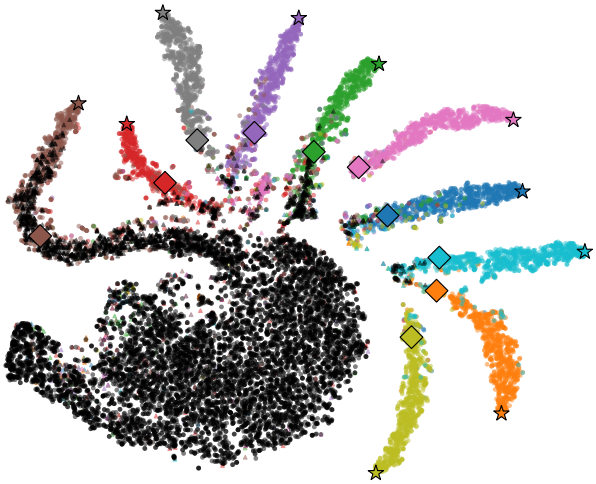
(a) TENT [44] w/ $\mathcal{L}_{t,\text{csID}}$



(b) TENT [44] w/ UniEnt+



(c) TENT [44] w/ $\mathcal{L}_{t,\text{csID}} + \mathcal{L}_{t,\text{ang}}$



(d) TENT [44] w/ $\mathcal{L}_{t,\text{ROSETTA}}$

Figure 8. T-SNE Visualization on CIFAR-10-C Test Set with SVHN-C as csOOD. Black markers represent csOOD samples, while colored markers represent csID samples. Circles indicate correctly classified csID samples or correctly detected csOOD samples, whereas triangles denote misclassified csID samples or undetected csOOD samples. Stars represent the embeddings of the updated class prototypes, and diamonds represent the fixed weights of the linear classification layer.