EFFICIENT OPEN-SET TEST TIME ADAPTATION OF VISION LANGUAGE MODELS

Manogna Sreenivas & Soma Biswas
Department of Electrical Engineering,
Indian Institute of Science,
Bengaluru, India
{manognas, somabiswas}@iisc.ac.in

ABSTRACT

In dynamic real-world settings, models must adapt to changing data distributions, a challenge known as Test Time Adaptation (TTA). This becomes even more challenging in scenarios where test samples arrive sequentially, and the model must handle open-set conditions by distinguishing between known and unknown classes. Towards this goal, we propose ROSITA, a novel framework for Open set Single Image Test Time Adaptation using Vision-Language Models (VLMs). To enable the separation of known and unknown classes, ROSITA employs a specific contrastive loss, termed ReDUCe loss, which leverages feature banks storing reliable test samples. This approach facilitates efficient adaptation of known class samples to domain shifts while equipping the model to accurately reject unfamiliar samples. Our method sets a new benchmark for this problem, validated through extensive experiments across diverse real-world test environments.

1 INTRODUCTION

Over the past decade, computer vision has made remarkable progress Deng et al. (2009); Ren et al. (2015); He et al. (2017); Everingham et al. (2010), primarily under the assumption that training and test data come from the same distribution. However, real-world applications are dynamic, where distribution gaps between training and test data arise due to domain shifts (e.g., lighting, weather, or camera variations Hendrycks & Dietterich (2019)) and semantic shifts (e.g., encountering unseen classes). This necessitates adapting deep learning models to dynamic test environments.

Test Time Adaptation (TTA)Wang et al. (2021); Schneider et al. (2020); Niu et al. (2022) addresses this challenge by adapting models without source data or ground truth labels, where test samples are seen only once. A more challenging scenario is Continuous TTA (CTTA)Döbler et al. (2023), where test domains change over time. However, these approaches predominantly operate in closed-set settings, assuming test data belong to known categories. In real-world scenarios, semantic shifts frequently occur, requiring models to identify and handle unknown classes. A classic example is that of autonomous driving Wang et al. (2022), where models trained for specific geographical locations are deployed elsewhere. For instance, a model trained to recognize only vehicles commonly seen in urban areas—such as *car, truck, motorcycle*—may incorrectly classify a *bicycle* as a *motorcycle* when deployed in rural settings. Open Set Adaptation aims to address this, but existing methods Li et al. (2023) often rely on batch-based updates, limiting their applicability to single-image streams.

Parallel to the recent advances in TTA, there has been tremendous progress in the development of large scale Vision Language Models (VLM) like CLIP Radford et al. (2021). Having trained on large scale web scrapped image-text pairs, these VLMs Radford et al. (2021) have demonstrated impressive zero shot generalization capabilities, making it a natural candidate for TTA. Recent works Shu et al. (2022); Samadh et al. (2023); Karmanov et al. (2024) adapt VLMs for single-image TTA in closed-set scenarios, but their utility in open-set scenarios remains underexplored. Addressing this gap, we propose a benchmark for *Open set Single Image Test Time Adaptation* using VLMs.

We define classes relevant to the downstream task as *desired*, and others as *undesired*. To prevent undesired samples from corrupting adaptation, we employ a Linear Discriminant Analysis (LDA) Fisher (1936); Li et al. (2023)-based identifier to filter undesired samples and classify desired ones accurately. To tackle open-set single-image TTA, we introduce the **Re**DUCe loss, which leverages reliable samples to dynamically contrast desired and undesired classes. This forms the foundation of our proposed framework, **ROSITA** for **O**pen set **S**ingle **I**mage **T**est time **A**daptation.

Our contributions are as follows:

- To the best of our knowledge, we are the first to tackle the challenging and realistic problem of *Open set Single Image Test Time Adaptation using VLMs*, setting a new benchmark.
- Our framework, **ROSITA**, adapts a VLM to recognize desired class samples with domain shifts while enabling it to effectively differentiate unfamiliar samples by saying "I don't know." This distinction between desired and undesired class samples is achieved using our ReDUCe loss, which dynamically contrasts these classes to enhance separability.
- We demonstrate the effectiveness of our method through extensive experiments across a diverse array of domain adaptation benchmarks, simulating various real-world test environments, with samples from single domain, continuous and frequently changing domains. We also experiment varying the ratio of desired and undesired class samples in the test stream.

2 Open Set Single Image Test Time Adaptation

2.1 Problem Setup

Test stream. The model encounters a single test sample x_t at time t, sampled from $\mathcal{D}_t = \mathcal{D}_d \cup \mathcal{D}_u$ comprising of: (i) Desired class samples: $\mathcal{D}_d = \{x_t; y_t \in C_d\}$, with domain shift and belonging to one of the $\mathcal{C}_d$ desired classes, for example, $C_d = \{car, bus, ..., motorcycle\}$; (ii) Undesired class samples: $\mathcal{D}_u = \{x_t; y_t \in C_u\}$, which have semantic shift (irrelevant classes) such that $C_d \cap C_u = \phi$.

Goal. Given a test sample x_t arriving at time t, the goal is to be first recognize if it belongs to a desired class or not, constituting a binary classification task. If x_t is identified as a desired class sample, a subsequent $|C_d|$-way classification is performed, else the prediction is "*I don't know*". In essence, the overall process can be viewed as a $|C_d| + 1$ way classification problem.

Open set Single Image TTA scenarios. We simulate several test scenarios inspired from the real world to evaluate the effectiveness of our method. (1) *Single domain*: We extend the standard TTA scenario where the test samples come from an unseen domain D_d (say *snow* corruption of CIFAR-10C) by incorporating undesired samples D_u (say MNIST). (2) *Continuously changing domains*: Here, D_t changes with time as $(D_d^1 \cup D_u) \rightarrow (D_d^2 \cup D_u) \ldots \rightarrow (D_d^n \cup D_u)$, where D_d^i is the i^{th} domain encountered. (3) *Frequently changing domains*: Here, we significantly reduce the number of samples per domain in continuous open set TTA. Lesser the samples per domain, more frequently the domain of the test stream changes, simulating very dynamic open set test scenarios. (4) *Vary the ratio of samples from C_d to C_u* in the test stream.

2.2 Baselines

We perform experiments using CLIP Radford et al. (2021) and MaPLe Khattak et al. (2023) backbones. CLIP consists of a Vision ($\mathcal{F}_V$) and Text ($\mathcal{F}_T$) encoder, trained using contrastive learning on image-text pairs. MaPLe backbone uses multimodal prompts to adapt CLIP for downstream tasks.

Classification using VLMs. Given a test image x_t and a set of desired classes $C_d = \{c_1, c_2, \ldots c_N\}$, we construct the text-based classifier by first prepending each class name with a predefined text prompt $\boldsymbol{p}_T$ = "A photo of a". This forms class-specific text inputs $\{\boldsymbol{p}_T, c_i\}$, which are then passed through the text encoder to obtain text embeddings $\boldsymbol{t}_i = \mathcal{F}_T(\{\boldsymbol{p}_T; c_i\})$ for each $c_i \in C_d$. As a result, we get the text-based classifier $\{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots \boldsymbol{t}_2\}$. Finally, the class prediction is made by identifying the text embedding $\boldsymbol{t}_i$ that has the highest similarity to the image feature f_t.

Desired vs Undesired Class Identifier. In real-world, a deployed model may encounter instances from both desired and undesired classes. *We equip all methods Shu et al. (2022); Karmanov et al. (2024); Zhang et al. (2024) with an LDA based parameter-free class identifier Fisher (1936); Li et al. (2023) to reject undesired class samples.* Subsequently, the model is adapted during test time.

Benchmark for Open-set Single Image TTA. We adapt the methods proposed for single image closed-set TTA such as ZSEval Radford et al. (2021), TPT Shu et al. (2022), PAlign Samadh et al.

(2023), TDA Karmanov et al. (2024) for our problem setting. We also adapt TPT and PAlign for continuous model update by adapting prompts, which we refer as TPT-C and PAlign-C respectively. We adapt the recent CNN based open-set TTA works (K+1)PC Li et al. (2023) for VLMs. These methods are described in detail in Appendix A.2.

2.3 Desired vs Undesired class Identifier

Contrary to closed-set TTA setting, updating the model using all the test samples is not desirable in the open-set scenario, where test samples can come from either C_d or C_u. It is hence imperative to equip the model with the ability to say *I don't know* by rejecting samples which do not belong to C_d. In the context of VLMs, we define a score (s_t) of a test sample to be the maximum cosine similarity with the text embeddings as given below:

$$s_t = \max_k \text{sim}(f_t, t_k); \quad k \in \{1, \dots C\} \tag{1}$$

This problem can be viewed as a binary classification problem between desired and undesired samples based on the score s_t. Defining a threshold to discriminate between the two can be particularly challenging in the TTA scenario as the samples are only accessible in an online manner. To circumvent this issue, following Li et al. (2023), we store the scores in a score bank $\mathcal{S}$, which is continuously updated in an online manner to store the latest $|S|$ scores, approximating the latest distribution of scores of the test data. Given this, the optimal threshold can be estimated by performing 1D LDA Fisher (1936). A simple linear search over a range of thresholds is done to identify the best threshold that minimizes the variance of scores of samples from C_d and C_u. For a threshold τ, let $\mathcal{S}_d = \{s_i | s_i > \tau, s_i \in S\}$ and $\mathcal{S}_u = \{s_i | s_i < \tau, s_i \in S\}$ denote the scores of samples identified to belong to C_d and C_u respectively. The optimal threshold τ_t^* at time t is identified as the one that minimizes the intra class variance as follows

$$\tau_t^* = \arg\min_\tau \frac{1}{|\mathcal{S}_d|} \sum_{s \in \mathcal{S}_d} (s - \mu_d)^2 + \frac{1}{|\mathcal{S}_u|} \sum_{s \in \mathcal{S}_u} (s - \mu_u)^2 \tag{2}$$

where μ_d and μ_u are the means estimated from $\mathcal{S}_d$ and $\mathcal{S}_u$ respectively. The test sample x_t is classified as desired if $s_t \geq \tau_t^*$ and undesired otherwise.s *We establish a strong benchmark for Open set Single Image TTA by equipping all the baseline methods (Section 2.2) with this simple and efficient LDA based class identifier.* We now describe the proposed framework **ROSITA**.

3 Proposed ROSITA Framework

Given a single test sample x_t at time t, it is first identified as a desired or undesired class sample as described above. This is important, since, using undesired class samples can have a negative impact on model adaptation. In this work, we propose a test time objective that can leverage both desired and undesired class samples through feature banks to enhance the discriminability between them.

Reliable samples for TTA. We first identify a test sample x_t as a *reliable desired or undesired class sample* based on its score s_t. As we have access to an approximate distribution of the scores as described in Section 2.3, we leverage the statistics μ_d and μ_u estimated through LDA to identify reliable samples. A test sample x_t is said to be a reliable sample belonging to desired classes C_d if its score $s_t > \mu_d$ and a reliable sample from any of the other classes C_u if its score $s_t < \mu_u$. We leverage **Re**liable samples to differentiate **D**esired vs **U**ndesired class samples through a **C**ontrastiv**e** (**ReDUCe**) Loss for Open-set Single Image Test time Adaptation, illustrated in Figure 1.

ReDUCe Loss. A contrastive objective typically needs positive and negative features, the goal being to maximize the similarity between a sample and its positive (could be augmentation Chen et al. (2020) or nearest neighbours Dwibedi et al. (2021)), while minimizing its similarity with the negatives. Such objectives (Chen et al., 2020; He et al., 2020; Khosla et al., 2020; Dwibedi et al., 2021) have been extensively used to learn good image representations in a self-supervised way. While self-supervised learning assumes access to abundant data in an offline manner giving the freedom to carefully choose positives and negatives, this problem is set in an online scenario, where the test samples arrive one at a time and are accessible only at that instant. This challenging setting makes it non trivial to use objectives by Dwibedi et al. (2021). To circumvent this issue of lack of abundant test data, we propose to store two dynamically updated feature banks $\mathcal{M}_d$ and $\mathcal{M}_u$ of sizes N_d and N_u, to store the features of reliable samples from C_d and C_u respectively. We propose a ReDUCe

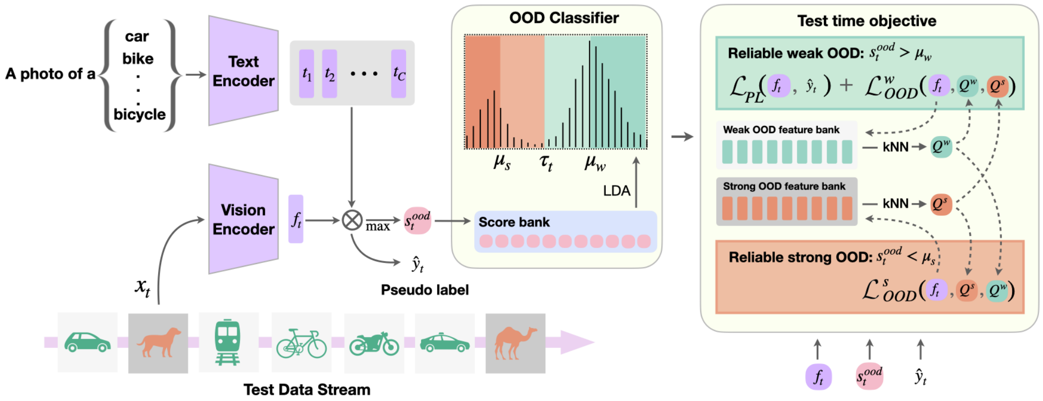


Figure 1: **ROSITA framework:** The test stream with samples from C_d and C_u arrive one at a time. The image is recognized as a sample from C_d and C_u through an LDA based class identifier. Further, if a test sample is reliable, the respective feature banks are updated and the proposed ReDUCe loss is optimized to update the LayerNorm parameters of the Vision Encoder.

objective to contrast a reliable sample from C_d by choosing its positives and negatives as the K nearest neighbours from $\mathcal{M}_d$ and $\mathcal{M}_u$ respectively and vice versa for a reliable sample from C_u. The buffer size for $\mathcal{M}_d$ is set as $|C_d| \times K$, where $|C_d|$ is the number of desired classes and K is the number of neighbours retrieved. The feature banks $\mathcal{M}_d$ or $\mathcal{M}_u$ are updated with a feature f_t if it is detected as a reliable sample from C_d and C_u respectively.

We fetch the K nearest neighbours of a reliable test sample x_t from each feature bank as follows.

$$Q_d = \text{kNN}(f_t; \mathcal{M}_d); \quad Q_u = \text{kNN}(f_t; \mathcal{M}_u) \tag{3}$$

Case 1: Reliable sample from C_d. If a test sample is identified as a reliable sample from C_d, we use a reliable pseudo-label loss on the sample x_t and its augmentation $\tilde{x}_t$ as follows:

$$\mathcal{L}_{Re} = \mathcal{L}_{CE}(x_t, \hat{y}_t) + \mathcal{L}_{CE}(\tilde{x}_t, \hat{y}_t); \quad \hat{y}_t = \text{argmax}_i \, \text{sim}(f_t, t_i) \tag{4}$$

where sim represents cosine similarity. Further, we also propose to use a contrastive objective to enhance the clustering of desired class samples while pushing them apart from the undesired class samples.

As we aim to correctly classify the desired class samples, we select positives z^+ from Q_d if its prediction y^+ matches with $\hat{y}_t$. The features Q_u consisting of its kNN from M_u act as its negatives. The following is the ReDUCe loss for a reliable sample from C_d:

$$\mathcal{L}_D = -\frac{1}{K^+} \sum_{z^+ \in Q^d} \mathbf{1}(y^+ = \hat{y}_t) \log \frac{\exp\left(\text{sim}\left(f_t, z^+\right)/\tau\right)}{\sum\limits_{z^- \in Q^u} \exp(\text{sim}(f_t, z^-)/\tau)} \tag{5}$$

where $K^+ = \sum_{z^+ \in Q^d} \mathbf{1}(y^+ = \hat{y}_t)$, is the number of neighbours positively matched with $\hat{y}_t$.

Case 2: Reliable sample from C_u. If a test sample is identified as a reliable sample from C_u, we use the following contrastive objective by selecting positives z^+ from Q_u and negatives z^- from Q_d:

$$\mathcal{L}_U = -\frac{1}{K} \sum_{z^+ \in Q_u} \log \frac{\exp\left(\text{sim}\left(f_t, z^+\right)/\tau\right)}{\sum\limits_{z^- \in Q_d} \exp(\text{sim}(f_t, z^-)/\tau)} \tag{6}$$

The LayerNorm parameters of the Vision Encoder are updated to minimize the following test time objective to adapt the model one sample at a time in an online manner:

$$\mathcal{L}_{ReDUCe} = \begin{cases} \mathcal{L}_{Re} + \mathcal{L}_D & \text{if} \quad s_t > \mu_d \\ \mathcal{L}_U & \text{if} \quad s_t < \mu_u \end{cases} \tag{7}$$

This objective improves the proximity between the test sample and its positives, suitably chosen based on its score s_t, while also pushing apart the test sample and its negatives. This collectively encourages the model to adapt such that each of the desired classes and undesired classes are clustered and farther apart from each other, improving the overall classification performance of C_d and C_u.

4 Experiments

Table 1: Results with ImageNet-C/R as desired class data D_d, MNIST and SVHN for D_u.

	Method	IN-C/MNIST			IN-C/SVHN			IN-R/MNIST			IN-R/SVHN		
		AUC ↑	FPR ↓	HM ↑	AUC ↑	FPR ↓	HM ↑	AUC ↑	FPR ↓	HM ↑	AUC ↑	FPR ↓	HM ↑
CLIP	ZS-Eval	93.39	55.52	41.43	85.89	72.91	40.83	91.27	91.09	71.50	90.43	75.04	71.66
	TPT	93.12	58.01	42.21	85.43	74.47	40.95	91.25	91.23	71.98	90.43	74.98	72.36
	TPT-C	56.57	99.12	6.19	11.38	100.00	7.24	82.81	85.79	68.25	80.94	80.03	69.18
	(K+1) PC	95.76	10.43	42.95	87.75	26.23	38.50	97.46	11.78	81.51	97.55	11.17	80.39
	TDA	90.54	76.23	43.66	86.76	75.45	43.07	91.79	87.83	71.56	90.67	75.41	71.48
	ROSITA	**99.52** +6.13	**4.06** +51.46	**48.53** +7.10	**98.34** +12.45	**10.21** +62.70	**46.32** +5.49	**99.44** +8.17	**4.29** +86.80	**83.53** +12.03	**98.62** +8.19	**9.08** +65.96	**80.75** +9.09

Implementation Details. We experiment with a diverse set of datasets to choose desired class data D_d and undesired class data D_u. For D_d, we use ImageNet-C Hendrycks & Dietterich (2019), ImageNet-R Hendrycks et al. (2021), VisDA Peng et al. (2017) and the Clipart, Painting, Sketch domains from DomainNet as style transfer datasets. We introduce samples from MNIST LeCun et al. (1998), SVHN Netzer et al. (2011) datasets as D_u in the test stream. We describe these, additional datasets, baseline methods and experimental details in the Appendix A.2.1 in a detailed manner.

Table 2: Acc_{HM} on VisDA and Clipart, Painting, Sketch from DomainNet as Weak OOD and MNIST as strong OOD.

Method	VisDA	Clipart	Painting	Sketch
ZSEval	78.28	50.22	47.81	48.59
TPT	78.42	57.71	49.73	54.67
TPT-C	75.35	57.57	49.31	54.41
(K+1)PC	90.35	71.21	70.61	67.21
TDA	76.85	61.04	51.20	55.26
ROSITA	90.64 +12.36	71.40 +21.18	70.89 +23.08	67.35 +18.76

Comparison with prior methods. We observe, from Table 1, 2 that TPT and PAlign perform similar to ZSEval in most datasets, as the prompts are reset after every single image update. On continuously updating prompts in TPT-C and PAlign-C, we observe a reduction in HM compared to ZS-Eval. The effect is more severe with CLIP when compared to MaPLe, as only the text prompts are updated keeping the vision encoder fixed. **ROSITA**, being equipped with a carefully designed objective to better discriminate between samples from C_d and C_u samples (Figure 3), results in overall better metrics in general. We report the results for different open-set scenarios in B.1. Further, we study the need for reliable samples in B.5, analyse the sensitivity of ROSITA's performance for different random seeds in B.2, choice of parameter K in B.3.

Table 3: Ablation study on loss components.

$\mathcal{L}_{Re}$	$\mathcal{L}_D$	$\mathcal{L}_U$	IN-R/MNIST		
			AUC ↑	FPR ↓	HM ↑
✗	✗	✗	91.27	91.09	71.5
✓	✗	✗	81.07	99.02	64.32
✗	✓	✗	87.73	94.67	67.28
✗	✗	✓	99.39	4.81	80.82
✗	✓	✓	99.48	4.40	81.92
✓	✓	✓	**99.44**	**4.29**	**83.53**

Loss Ablation. We observe that only using $\mathcal{L}_{Re}$ or $\mathcal{L}_D$ improves the metrics for CIFAR-10C dataset. For ImageNet-R (IN-R) as D_d, using $\mathcal{L}_{Re}$ or $\mathcal{L}_D$ is observed to increase FPR and decrease HM. IN-R has 200 classes making it a more challenging and confusing task compared to CIFAR-10C. This decrease in performance for IN-R can be attributed to the misclassification of some samples from C_u as reliable desired class samples, increasing the confusion between C_d and C_u classes. Using $\mathcal{L}_U$ significantly reduces the confusion between samples from C_d and C_u, shown by the significant drop in FPR compared to ZSEval. The contrastive objectives $\mathcal{L}_D$ and $\mathcal{L}_U$ to separate the two types of samples, in conjunction with reliable pseudo label loss $\mathcal{L}_{Re}$ which aids to improve the $|C_d|$-way classification of desired class samples, gives the overall best results.

Table 4: Memory overhead in ROSITA due to feature banks.

Dataset	C	**No. of features**	**Memory (in MB)**
CIFAR-10C	10	5x10+64	0.758
VisDA	12	5x12+64	0.778
CIFAR-100C	100	5x100+64	1.679
ImageNet-R	200	5x200+64	2.703
ImageNet-C	1000	5x1000+64	10.89

Memory buffer. Prior prompt tuning methods like TPT Shu et al. (2022), Samadh et al. (2023) do not require any memory buffer. TDA Karmanov et al. (2024) requires a memory buffer of size $(|C_d| \times (3+2)) \times F$ to store 3 features per desired class in the positive cache and 2 features per class in the negative cache. DPE Zhang et al. (2024) requires a memory buffer of size

$(|C_d| \times 3) \times F$ to store 3 features per desired class. ROSITA requires a small memory buffer of size 512 for the score bank S and $(|C_d| \times K + |M_u|) \times F$ for the feature banks. For a ViT-B16 ($F = 512$) model with ImageNet-C ($|C_d| = 1000$), the required memory buffer size is $5 \times 1000 \times 512 + 64 \times 512$ (10.89MB). *The memory to store them and computation required to compute feature similarity is as lightweight as performing a forward pass through a simple linear layer, demonstrating the memory and computational efficiency of ROSITA for real time applications.*

GPU Memory. For prompt tuning methods TPT/-C and PAlign/-C, the GPU memory and time taken (secs/image) scales with the number of classes, as it requires more memory to store the intermediate activations and gradients. The time taken to perform forward and backward pass through the text encoder also depends on the number of classes. On the other hand, ROSITA requires two forward passes and one backward pass through the vision encoder for reliable test samples. For e.g., for ImageNet-C dataset with 1000 classes, ZSEval, TPT, TDA and ROSITA require 5.71 GB, 23.24 GB, 5.71 GB and 5.73 GB GPU memory to perform a single image based model update. Hence, ROSITA is computationally very efficient.

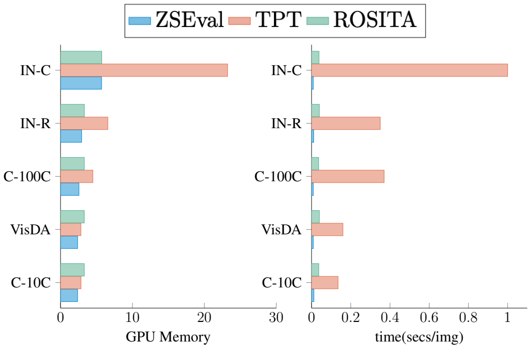


Figure 2: Complexity Analysis of different methods using CLIP backbone.

5 Conclusion

In this work, we propose **ROSITA**, a novel framework to address the challenging problem Open set Test Time Adaptation (TTA) on a single image basis. ROSITA effectively distinguishes between samples from desired classes vs others by leveraging two dynamically updated feature banks. The proposed ReDUCe loss facilitates effective model adaptation by using reliable, while mitigating any negative impact of undesirable samples in the test stream. Through extensive experimentation on diverse domain adaptation benchmarks, we demonstrate the effectiveness of ROSITA in several scenarios inspired by the dynamic real world environment.

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *CVPR*, 2023.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *ICCV*, 2021.

Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7 (2):179–188, 1936.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out1-of-distribution generalization. In *CVPR*, 2021.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14162–14171, 2024.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, 2023.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yushu Li, Xun Xu, Yongyi Su, and Kui Jia. On the robustness of open-world test-time training: Self-training with dynamic prototype expansion. In *ICCV*, 2023.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 7. Granada, Spain, 2011.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *ICML*, 2022.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

Jameel Hassan Abdul Samadh, Hanan Gani, Noor Hazim Hussein, Muhammad Uzair Khattak, Muzammal Naseer, Fahad Khan, and Salman Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *NeurIPS*, 2023.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *NeurIPS*, 2020.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *CVPR*, 2022.

Ce Zhang, Simon Stepputtis, Katia P Sycara, and Yaqi Xie. Dual prototype evolving for test-time generalization of vision-language models. In *NeurIPS*, 2024.

APPENDIX

A Implementation Details

A.1 Vision Language Models

CLIP Radford et al. (2021) is a multimodal VLM consisting of two modules: Vision encoder and Text encoder denoted as $\mathcal{F}_V$ and $\mathcal{F}_T$ respectively. During pre-training, the two modules are jointly trained in a contrastive self-supervised fashion to align massive amounts of web-scrapped image-text pairs. CLIP has demonstrated impressive zero-shot generalization ability across a wide variety of datasets.

MaPLe Khattak et al. (2023) is a multimodal prompt learner model that simultaneously adapts both the vision and text encoders while finetuning CLIP for downstream tasks. They use learnable text prompts $\boldsymbol{p}_T$ and bridge the two modalities using visual prompts obtained as $\boldsymbol{p}_V = \mathrm{Proj}(\boldsymbol{p}_T)$. Learnable tokens are also introduced in the deeper layers of both image and text encoders, to enable progressive adaptation of the features. As in Samadh et al. (2023), we use MaPLe as an additional VLM backbone to test our approach.

A.2 Baseline Methods

ZSEval: Given a test image x_t, the image feature is extracted from the vision encoder as $f_t = \mathcal{F}_V(x_t)$. For a C-class classification problem, the classifier is obtained by prepending a predefined text prompt $\boldsymbol{p}_T$="A photo of a", with the class names $\{c_1, c_2, \ldots c_C\}$ to form class specific text inputs $\{\boldsymbol{p}_T, c_i\}$ for $i \in \{1, \ldots C\}$. These texts are then embedded through the text encoder as $\boldsymbol{t}_i = \mathcal{F}_T(\{\boldsymbol{p}_T; c_i\})$ to get the text classifiers $\{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots \boldsymbol{t}_C\}$. The class prediction is made by identifying the text feature $\boldsymbol{t}_i$ which has the highest similarity with the image feature f_t.

TPT Shu et al. (2022) aims to improve the zero shot generalization ability of CLIP by providing custom adaptable context for each image. This is done by prepending learnable text prompts $\boldsymbol{p}_T$ to the class names instead of a predefined text prompt. The text classifiers $\boldsymbol{t}_i = \mathcal{F}_T(\{\boldsymbol{p}_T; c_i\}), i \in \{1, 2, \ldots C\}$ are now a function of these learnable prompts, which are specially adapted for each test image using an entropy minimization objective as $\arg\min_{\boldsymbol{p}_T} \mathcal{L}_{\text{ent}}$. The entropy is obtained using the average score vector of the filtered augmented views.

PromptAlign Samadh et al. (2023) (PAlign) leverages multimodal prompt learner model MaPLe Khattak et al. (2023) to facilitate the adaptation of both vision and language encoders for each test sample. Inspired by earlier TTA works Schneider et al. (2020); Wang et al. (2021), they propose to align the token distributions of source and target domains, considering ImageNet as a proxy for the source dataset of CLIP. The vision and language prompts of MaPLe are optimized with the objective $\arg\min_{\{\boldsymbol{p}_V, \boldsymbol{p}_T\}} \mathcal{L}_{ent} + \mathcal{L}_{align}$ for each sample x_t.

TPT-C/PAlign-C: We adapt TPT and PAlign for continuous model update, which we refer as TPT-C and PAlign-C respectively. The prompts $\{\boldsymbol{p}_T\}$ and $\{\boldsymbol{p}_V, \boldsymbol{p}_T\}$ in TPT and PAlign are continuously updated with the test stream with their respective test objectives for this purpose.

(K+1)PC (Li et al., 2023): This was the first work exploring open world TTA, however it was done in the context of CNNs and not VLMs. Also, the test samples come in batches, while we perform single image TTA. We adapt this method for our problem setting as follows: As we use VLMs, we use the text prototypes (instead of the source prototypes). The prototype pool is dynamically updated by adding features of reliable test samples recognized to belong to undesired classes. The vision encoder is updated using a (K+1) way prototypical cross entropy loss.

TDA (Karmanov et al., 2024): TDA is a training-free dynamic adapter for test-time adaptation in vision-language models, utilizing a lightweight key-value cache for efficient pseudo label refinement without backpropagation.

Table 5: Results with CIFAR-10C and CIFAR-100C for desired classes D_d and four other datasets (MNIST, SVHN, Tiny-ImageNet, CIFAR-100C/10-C respectively) for D_U. *All methods use the same OOD detector described in Section* 2.3

		Method	MNIST			SVHN			Tiny-ImageNet			CIFAR-100C/10-C		
			AUC ↑	FPR ↓	HM ↑	AUC ↑	FPR ↓	HM ↑	AUC ↑	FPR ↓	HM ↑	AUC ↑	FPR ↓	HM ↑
CIFAR-10C	CLIP	ZS-Eval	91.91	85.04	75.57	89.93	64.20	74.08	91.33	27.07	74.63	82.57	67.92	68.89
		TPT	91.89	85.55	75.81	89.93	64.41	74.36	91.31	27.23	75.17	82.57	68.06	69.17
		TPT-C	81.64	67.53	74.86	58.48	71.72	48.26	74.08	61.45	49.88	61.45	94.30	46.10
		(K+1)PC	98.05	12.50	83.27	80.74	50.33	70.10	87.09	52.29	73.98	62.55	91.68	56.46
		TDA	92.94	71.11	77.06	92.02	52.68	76.64	91.68	25.37	75.94	83.54	66.06	70.13
		ROSITA	**99.10** **+7.19**	**7.63** **+77.41**	**84.17** **+8.60**	**94.79** **+4.86**	**32.59** **+31.61**	**78.80** **+4.72**	**96.43** **+5.10**	**12.10** **+14.97**	**80.06** **+5.43**	**82.99** **+0.42**	**62.89** **+5.03**	**69.56** **+0.6**
	MAPLE	ZS-Eval	98.48	3.77	83.63	98.34	**7.86**	83.57	90.86	27.54	76.04	86.14	52.08	71.76
		TPT	98.15	5.67	81.56	98.34	7.89	82.73	90.86	27.61	75.46	86.15	52.14	70.94
		TPT-C	98.56	**3.74**	83.51	98.32	8.18	83.47	91.18	26.93	76.31	86.50	50.56	71.07
		PAlign	98.15	5.67	82.24	**98.34**	7.90	83.51	90.86	27.60	75.98	86.15	52.18	71.52
		PAlign-C	98.56	3.74	83.49	98.32	8.13	83.46	91.18	26.90	76.30	86.50	50.58	71.04
		ROSITA	**99.34** **+0.86**	5.22 **+1.45**	**87.63** **+4.00**	97.80 **+0.54**	13.15 **+5.29**	**84.17** **+0.60**	**91.67** **+0.81**	**25.31** **+2.23**	**77.67** **+1.63**	**86.82** **+0.68**	**50.33** **+1.75**	**73.15** **+1.39**
CIFAR-100C	CLIP	ZS-Eval	77.78	99.93	48.39	64.70	98.68	45.85	67.31	73.89	45.80	63.28	93.25	44.04
		TPT	77.76	99.94	48.33	64.71	98.63	45.85	67.28	73.82	45.93	63.26	93.20	44.02
		TPT-C	51.57	100.00	27.04	9.40	99.98	5.74	59.74	79.76	18.41	55.86	**86.35**	13.64
		(K+1)PC	96.89	12.15	59.72	75.24	51.64	43.73	41.84	99.61	31.83	54.02	93.93	32.00
		TDA	80.33	99.57	46.52	71.77	96.11	46.01	70.70	69.63	47.52	66.07	91.90	45.79
		ROSITA	**96.07** **+18.29**	**19.28** **+80.65**	**57.34** **+8.95**	**82.09** **+17.39**	**64.64** **+34.04**	**48.17** **+2.32**	**83.55** **+16.24**	**50.76** **+23.13**	**55.88** **+10.08**	**68.54** **+5.26**	89.71 **+3.54**	**47.98** **+3.94**
	MAPLE	ZS-Eval	87.43	64.19	54.97	92.98	40.51	56.42	68.80	74.35	48.24	66.93	87.94	46.06
		TPT	87.42	64.09	53.09	92.97	40.44	54.37	68.80	**74.20**	46.97	66.93	87.95	44.38
		TPT-C	87.65	63.08	55.14	93.09	40.30	56.31	68.85	74.71	48.53	66.97	87.94	46.30
		PAlign	87.42	64.11	53.98	92.97	40.48	55.37	68.80	74.23	47.69	66.93	87.93	45.16
		PAlign-C	88.25	57.31	55.69	93.45	39.39	57.39	68.76	78.12	48.15	66.82	87.80	47.01
		ROSITA	**97.04** **+9.61**	**11.01** **+53.18**	**62.06** **+7.09**	**96.26** **+3.28**	**20.99** **+19.52**	**59.25** **+2.83**	**70.37** **+1.57**	77.00 **+2.65**	**48.68** **+0.44**	**69.57** **+2.64**	**83.61** **+4.33**	**48.80** **+2.74**

A.2.1 Datasets

We experiment with a diverse set of datasets, encompassing corruption datasets, style transfer datasets, and other common datasets.

CIFAR10-C Hendrycks & Dietterich (2019) is a small-scale corruption dataset of 10 classes with 15 common corruption types. It consists of 10,000 images for each corruption.

CIFAR-100C Hendrycks & Dietterich (2019) is also a corruption dataset with 100 classes and 15 corruption types. It also consists of 10,000 images for each corruption.

ImageNet-C Hendrycks & Dietterich (2019) is a large-scale corruption dataset spanning 1000 categories with a total of 50,000 images. 15 types of corruption images are synthesized from these 50,000 images.

ImageNet-R Hendrycks et al. (2021) is a realistic style transfer dataset encompassing interpretations of 200 ImageNet classes, amounting to a total of 30,000 images.

VisDA Peng et al. (2017) is a synthetic-to-real large-scale dataset, comprising of 152,397 synthetic training images and 55,388 real testing images across 12 categories.

MNIST LeCun et al. (1998) is a dataset of handwritten images consisting of 60,000 training and 10,000 testing images.

SVHN Netzer et al. (2011) is also a digits dataset with house numbers captured from real streets. It consists of 50,000 training images and 10,000 testing images.

We perform experiments on five weak OOD datasets. The corresponding strong OOD datasets are chosen such that there is no overlap between weak and strong OOD datasets and is described in Table 6. The 15 corruptions fall into four categories: synthetic weather effects, per-pixel noise, blurring, and digital transforms. *snow* corruption is a synthesized weather effect on which all the main experiments of CIFAR-10C, CIFAR-100C and ImageNet-C are done.

Table 6: Details of Weak and strong OOD dataset combinations

Datasets		# images		
Weak OOD	Strong OOD	weak	strong	total
CIFAR-10C	MNIST, SVHN, Tiny ImageNet, CIFAR-100C	10000	10000	20000
CIFAR-100C	MNIST, SVHN, Tiny ImageNet, CIFAR-10C	10000	10000	20000
ImageNet-C	MNIST, SVHN	50000	50000	100000
ImageNet-R	MNIST, SVHN	30000	30000	60000
VisDA	MNIST, SVHN	50000	50000	100000

B Additional Analysis

Here, we study the robustness of the proposed method ROSITA more extensively, in the terms of (1) Performance in different Open set TTA scenarios. (2) Error bars on different test data streams, (2) Role of the parameter K, the number of neighbours, (4) Analysis of OOD scores on using different combinations of the proposed loss components, (5) Effectiveness of LDA based OOD detector in comparison with simple thresholding, (6) Complexity Analysis.

B.1 Performance in different Open set TTA scenarios.

(a) Continuously changing domains: We sequentially present 15 corruptions from CIFAR-10C, which form the domain D_d, alongside samples from four other datasets D_u. **(b) Frequently changing domains:** To further simulate more dynamic test environments, for CIFAR-10C/MNIST, we reduce the number of samples per corruption to 100, 250, 500, and 1000 in the continuously changing domain open-set TTA scenario. Reducing the sample count per corruption causes more frequent domain changes, increasing the challenge for adaptation. **(c) Varying ratio of samples belonging to classes C_d vs C_u:** We simulate real-world scenarios using the CIFAR-10C/MNIST dataset by varying the ratio of samples from the known classes C_d versus unknown classes C_u in the test stream by varying this ratio as 0.2, 0.4, 0.6, and 0.8. From results in Table 7,we observe that **ROSITA** demonstrates consistent superiority across all three open-set TTA scenarios, showcasing its capability to adapt effectively to both continuously and frequently changing domains, as well as varying class distributions.

Table 7: Performance in different Open set TTA scenarios.

Method	(a) Continuously changing domains				(b) Frequently changing domains				(c) Varying ratio of C_d/C_u			
	CIFAR-10C				No. of samples per corruption				Ratio			
	SVHN	MNIST	Tiny	C-100C	100	200	500	1000	0.2	0.4	0.6	0.8
ZSEval	64.33	64.04	66.50	58.49	61.41	61.87	61.42	63.30	75.56	75.59	75.57	75.56
TPT	64.26	64.03	66.50	58.47	61.33	62.32	61.59	63.24	75.67	75.75	75.81	75.83
TPT-C	33.05	46.44	59.38	37.24	60.62	61.30	57.16	34.88	72.70	74.31	74.79	75.16
(K+1)PC	65.13	62.52	66.93	57.46	60.90	60.76	61.40	63.26	62.31	68.85	81.70	82.90
TDA	66.02	66.44	67.64	59.44	60.17	61.43	63.22	64.82	72.45	75.04	77.54	77.91
ROSITA	**66.86**	**65.26**	**68.89**	**59.16**	**61.64**	**66.82**	**67.97**	**73.24**	**82.96**	**83.97**	**84.51**	**84.37**

B.2 Analysis on error bars

To study the robustness of our method for differently ordered test streams, we run ROSITA with five random seeds and report the Mean and Standard deviation of the Acc_{HM} in Table 8 for CIFAR-10C/100C as weak OOD data and MNIST, SVHN, Tiny ImageNet, CIFAR-100C/10C as strong OOD data (corresponding to our results in Table 5 in the main paper). We observe that the variance in the performance of ROSITA is very low, reinforcing the robustness of the proposed method for different shuffled datasets and augmentations created.

Table 8: Performance (Mean and Standard deviation of Acc_{HM}) of ROSITA across 5 random seeds for CIFAR-10/100C as weak OOD data with 4 strong OOD datasets.

Dataset	MNIST	SVHN	Tiny	CIFAR-100/10C
CIFAR-10C	84.07 ± 0.023	78.90 ± 0.038	80.10 ± 0.014	69.44 ± 0.018
CIFAR-100C	57.09 ± 0.041	47.90 ± 0.047	55.95 ± 0.051	48.10 ± 0.024

Table 9: Performance (Acc_{HM}) on varying K with MNIST as strong OOD.

Weak OOD Dataset	# Classes	K					
		0	1	3	5	7	9
CIFAR-10C	10	80.97	83.9	84.32	84.17	84.10	84.02
ImageNet-R	200	64.32	83.65	83.87	83.53	83.39	83.42
ImageNet-C	1000	42.05	48.35	47.17	48.53	48.37	47.73

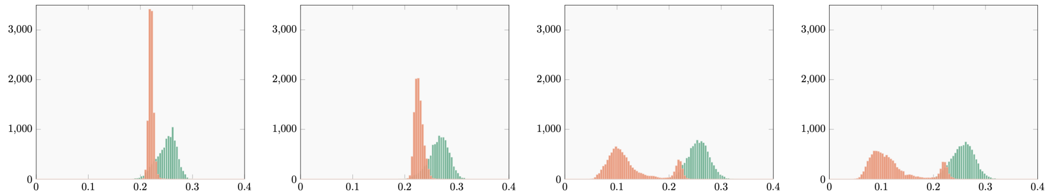
B.3 Analysis on parameter K

We vary the hyperparameter K which represents the number of positives and negatives chosen in Equation 5 and 6 and report the results (Acc_{HM}) in Table 9. The size of the weak OOD feature bank $\mathcal{M}_w$ is set as $N_w = K \times C$. N_s increases with the number of classes as well as the number of neighbours K. We set K to be 5 in all main results reported, which corresponds to feature bank size N_s of 50, 1000, 5000 respectively for the datasets CIFAR-10C, ImageNet-R and ImageNet-C respectively. In Table 9, we abuse the notion $K = 0$ to correspond to the case where only $\mathcal{L}_{PL}$ is used and no contrastive OOD loss is used. The results show that even with $K = 1$, there is a significant improvement in Acc_{HM} when compared to the case where $\mathcal{L}^w_{OOD}, \mathcal{L}^s_{OOD}$ is not used ($K = 0$). On further increasing K, we observe improvement only for the CIFAR-10C weak OOD dataset, but the performance is similar for ImageNet-R and ImageNet-C for higher values of K as well. Further, we investigate this observation that the performance of ROSITA is similar on significantly varying K or the feature bank size. For $K = 5$, we check the average number of positives actually selected for L^w_{OOD} in Equation 5 for each of these datasets. We find this to be $4.1, 2.5$ and 1.5 for CIFAR-10C, ImageNet-R and ImageNet-C respectively. This agrees with the results in Table 9 where K of 3, 5 works better compared to 1 as more neighbours have common pseudo label, aiding the clustering of classes of interest. For CIFAR-10C and ImageNet-R, using $K < 5$ suffices and for ImageNet-C as only 1-2 neighbours are matched for majority of reliable OOD samples, setting $K = 1$ suffices. For practical purposes, this observation suggests that the weak OOD feature buffer size can indeed be reduced based on storage budget available depending on the application and device the model is deployed on. For e.g., if the memory budget available can store only upto 1000 features, K can be set flexibly depending on the number of classes of interest. For ImageNet-C with 1000 classes, K can be set to 1.

B.4 Loss Ablation

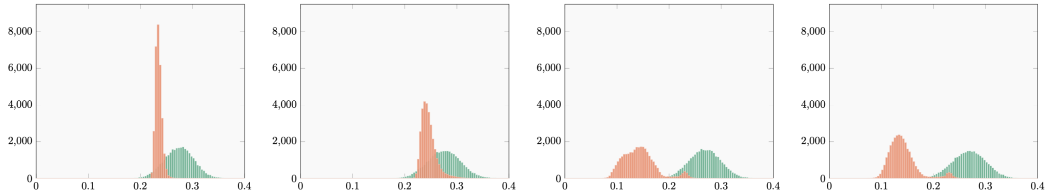
We provide detailed results of Table 3 in Table 10. Additionally, we visualise the histograms of OOD scores on using different combinations of the proposed loss components in the Figures 3, 4, justifying their role in better discrimination of weak and strong OOD sample.

From Figure 3 and 4, we observe that, on using just $\mathcal{L}_{PL}$, the weak and strong OOD scores still sufficiently overlap, similar to the case of ZSEval. The performance purely depends on the quality of pseudo labels of the detected reliable weak OOD samples. In CIFAR-10C, as there are only 10 classes and given that ZSEval performance in CIFAR-10C is fairly good, it ensures good quality pseudo labels, hence resulting in overall better metrics on even using $\mathcal{L}_{PL}$ as shown in Table 10. ImageNet-R dataset inherently has more confusion as it is a 200-way classification problem. This naturally could result in low quality pseudo labels, in turn degrading the performance compared to ZSEval. Alongside, using $\mathcal{L}_{PL}$ for weak OOD samples which are misclassified as strong OOD samples increases the FPR and results in a decrease in metrics overall compared to ZSEval. On the other hand, using $\mathcal{L}^w_{OOD} + \mathcal{L}^s_{OOD}$ separates the OOD scores of weak and strong samples, resulting



(a) ZSEval (b) L_{PL} (c) $L^{w}_{OOD} + L^{s}_{OOD}$ (d) $L_{PL}+L^{w}_{OOD}+L^{s}_{OOD}$

Figure 3: Histograms of Weak and Strong OOD scores for ZS-Eval and on using different loss components of ROSITA on CIFAR-10C/MNIST dataset using CLIP.



(a) ZSEval (b) L_{PL} (c) $L^{w}_{OOD} + L^{s}_{OOD}$ (d) $L_{PL}+L^{w}_{OOD}+L^{s}_{OOD}$

Figure 4: Histograms of Weak and Strong OOD scores for ZS-Eval and on using different loss components of ROSITA on ImageNet-R/MNIST dataset using CLIP.

in two distinct peaks as seen in Figure 3 and 4, which in turn results in a significantly low FPR as reported in Table 10. The best results are obtained using all the three proposed loss components $\mathcal{L}_{PL} + \mathcal{L}^{w}_{OOD} + \mathcal{L}^{s}_{OOD}$, which better discriminates weak and strong OOD samples and also helps in selecting weak OOD samples with more accurate pseudo labels. Hence, using pseudo label loss and OOD contrastive losses aid each other, resulting in the best overall metrics as shown in Table 10.

Table 10: Detailed results on Loss Ablation.

$\mathcal{L}_{Re}$	$\mathcal{L}_{D}$	$\mathcal{L}_{U}$	CIFAR-10C/MNIST					ImageNet-R/MNIST				
			AUC	FPR	Acc_D	Acc_U	Acc_{HM}	AUC	FPR	Acc_D	Acc_U	Acc_{HM}
✗	✗	✗	91.91	85.04	60.82	99.77	75.57	91.27	91.09	55.67	99.90	71.50
✓	✗	✗	95.29	30.82	68.36	99.30	80.97	81.07	99.02	48.42	95.76	64.32
✗	✓	✗	95.23	28.91	66.93	98.52	79.71	87.73	94.67	51.13	98.34	67.28
✗	✗	✓	98.61	12.73	66.60	99.68	79.84	99.39	4.81	67.81	99.99	80.82
✗	✓	✓	99.27	4.15	67.76	99.73	80.69	99.48	4.40	69.38	99.98	81.92
✓	✓	✓	99.10	7.63	72.81	99.74	84.17	99.44	4.29	71.73	99.98	83.53

B.5 Analysis on OOD Classifier and Reliable samples

Here, we study the role of OOD classifier in the Open World Single Image Test Time Adaptation setting. We compare the LDA based OOD classifier described in Section 2.3 in comparison with simple confidence thresholding with TTA algorithm of ROSITA described in 3. A test sample is classified as weak OOD if $s_t^{ood} > \tau_t$ and strong OOD if $s_t^{ood} < \tau_t$. Further, in ROSITA, TTA is performed on reliable weak and strong OOD samples based on LDA statistics as described in Section 3. We generalize this and call a test sample as reliable weak OOD sample if $s_t^{ood} > \tau_w$ and strong OOD if $s_t^{ood} > \tau_s$. Here, we perform experiments to understand the role of OOD classifier, reliable samples and the performance of ROSITA with time.

Effectiveness of the LDA based OOD classifier: To study the role of the OOD classifier in ROSITA, we perform the following experiments **(1) Simple thresholding:** We set fixed thresholds τ_w, τ_s to identify reliable weak and strong OOD samples respectively and τ_t to classify a sample into

Table 11: Comparison of Simple threshold (row 1-3) vs LDA based OOD detector (row 5). Comparison of ROSITA using all samples (row 4) vs only reliable samples (row 5) for TTA.

Thresholds	strong OOD dataset: MNIST				
$\tau_s/\tau_t/\tau_w$	C-10C	C-100C	IN-C	IN-R	VisDA
0.4/0.6/0.8	43.44	34.42	1.20	77.12	88.49
0.3/0.5/0.7	33.70	32.60	1.74	80.29	50.87
0.5/0.5/0.5	22.82	37.41	1.91	30.90	32.31
$\tau_t/\tau_t/\tau_t$	**84.99**	55.16	44.05	83.28	**91.24**
$\mu_s/\tau_t/\mu_w$	84.17	**57.34**	**48.53**	**83.53**	90.64

weak or strong OOD . **(2) LDA based:** As described in Section 2.3, we set τ_w to μ_w and τ_s to μ_s to identify reliable weak and strong OOD samples to perform TTA. We report the results(Acc_{HM}) of all five weak OOD datasets with MNIST as strong OOD dataset using CLIP backbone. **Observations:** The first three rows in Table 11 correspond to simple thresholding cases where the thresholds are manually set and kept fixed throughout TTA using ROSITA. We observe that the performance significantly varies for different choice of thresholds, especially in the case of ImageNet-R (IN-R) and VisDA here. This shows that it is not feasible to choose these thresholds apriori in a TTA task as the softmax confidence scores depends on unknown factors like the type, severity of domain shift, confusion of classes etc. Hence, using fixed threshold to discriminate between weak and strong OOD samples is undesirable. In the OOD classifier we use (Section 2.3), a score bank $\mathcal{S}$ is used to track how the OOD scores of the test samples change with time. The statistics μ_w, μ_s are continuously estimated to identify reliable weak and strong OOD samples. From Table 11, we observe that the best results (last row) are obtained on using the thresholds estimated in an online manner.

Need for reliable samples: To understand the role of selecting reliable samples for TTA, we do a simple experiment where we only use the threshold τ_t to distinguish between a weak and strong OOD samples. For all weak OOD samples classified, we perform TTA using the loss defined in Equation 5. Similarly, we use the objective in Equation 6 for all strong OOD samples. The results are reported in the fourth row in Table 11. We see that, for CIFAR-10C and VisDA, this case performs slightly better than our case(last row in Table 11) where TTA is performed only on reliable samples. CIFAR-10C and VisDA dataset have 10 and 12 classes of interest respectively. The zero shot performance of these datasets being good, as the class confusion is less, using all samples for TTA can be helpful. On the other hand, the classification in CIFAR-100C, ImageNet-C and ImageNet-R is harder, due the confusion arising due to the large number of classes. Using non reliable test samples, with scores in the range $\mu_s < s_t^{ood} < \mu_w$ can adversely affect the adaptation process. Hence, using only reliable samples for TTA performs better for these datasets as seen from the last two rows in Table 11). In a general test time adaptation scenario, where we have no prior information about the difficulty of the classification task, in terms of severity of domain shift and class confusion, it is desirable to only use reliable samples for model updates.