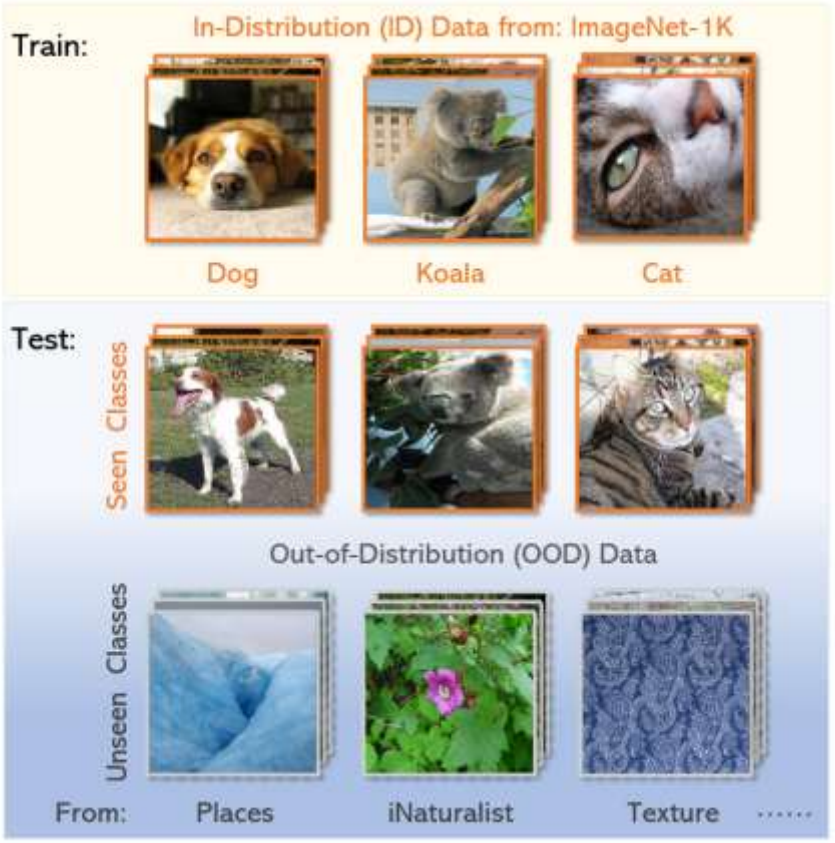
南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Delving into Out-of-Distribution Detection with Vision-Language Representations

Yifei Ming[1] **Ziyang Cai**[1] **Jiuxiang Gu**[2] **Yiyou Sun**[1] **Wei Li**[3] **Yixuan Li**[1]
[1]Department of Computer Sciences, University of Wisconsin-Madison
[2]Adobe [3]Google Research
{alvinming,ziyangc,sunyiyou,sharonli}@cs.wisc.edu
jigu@adobe.com mweili@google.com

NeurIPS 2022

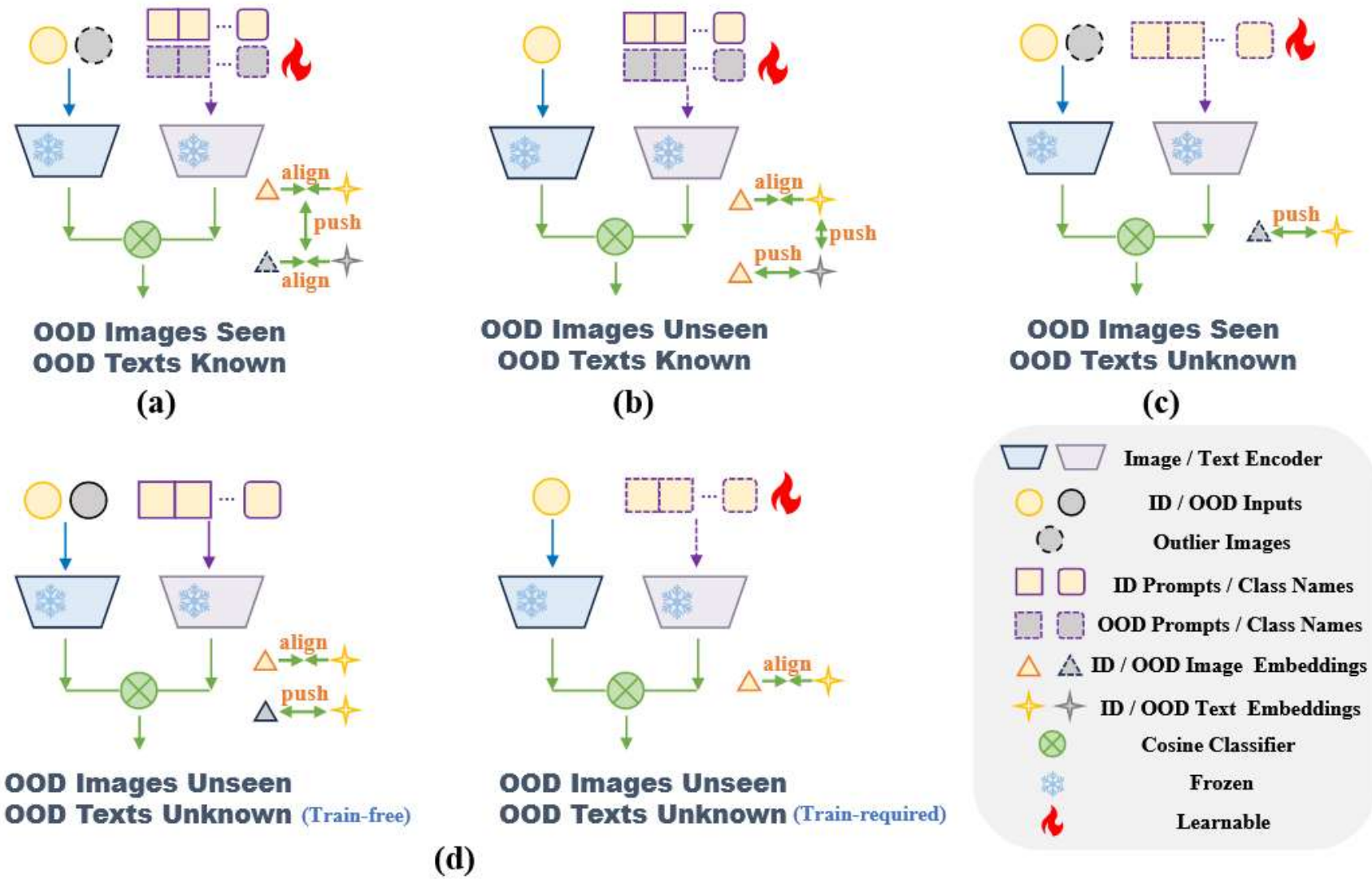
Background



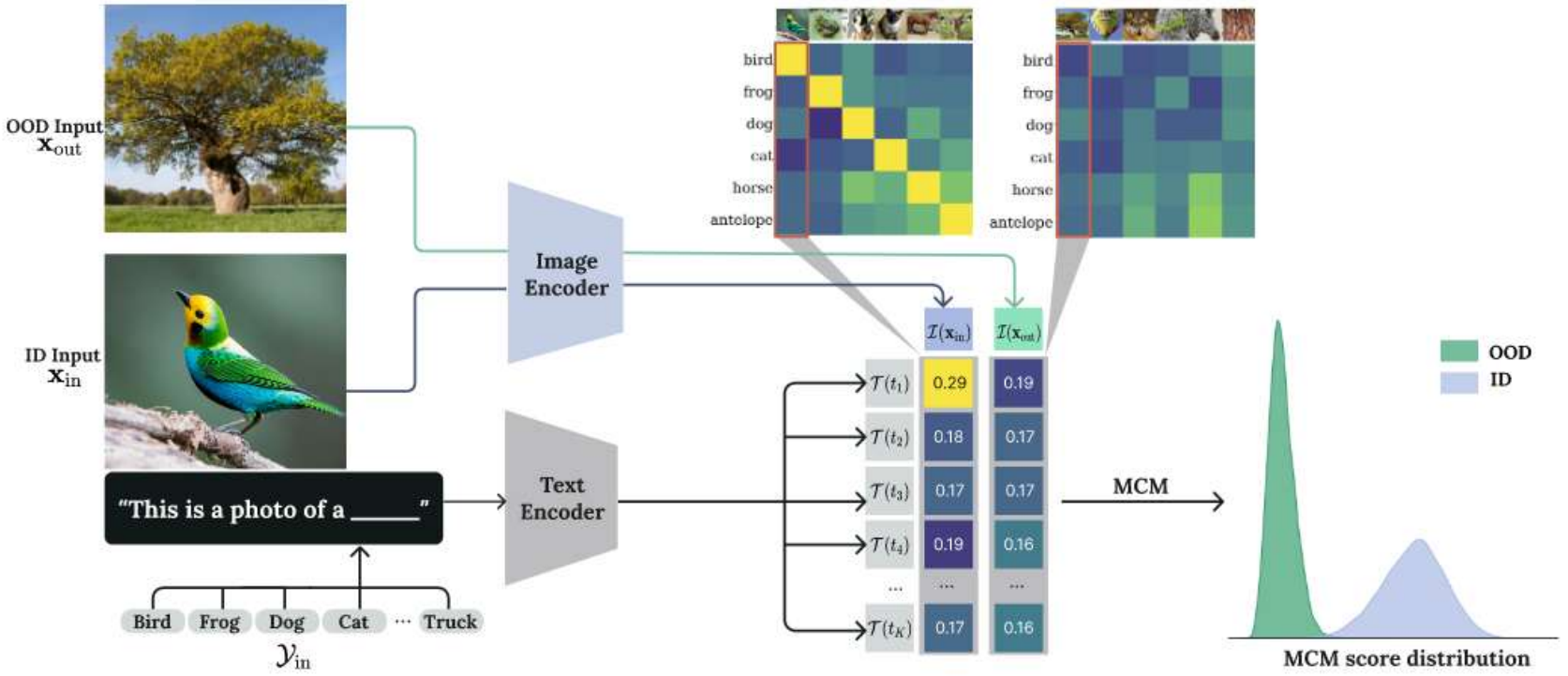
- the vast majority of OOD detection methods are driven by **single-modal learning**.For example, labels are typically encoded as **one-hot vectors** in image classification, leaving the semantic information encapsulated in texts largely unexploited.

- OOD detection relying on **pure visual information** can inherit the limitations.when an OOD input may be visually similar to in-distribution (ID) data **yet semantically different** from any ID class.

Motivation


align
push
align
OOD Images Seen
OOD Texts Known
(a)
align
push
push
OOD Images Unseen
OOD Texts Known
(b)
push
OOD Images Seen
OOD Texts Unknown
(c)
align
push
OOD Images Unseen
OOD Texts Unknown (Train-free)
align
OOD Images Unseen
OOD Texts Unknown (Train-required)
(d)
Image / Text Encoder
ID / OOD Inputs
Outlier Images
ID Prompts / Class Names
OOD Prompts / Class Names
ID / OOD Image Embeddings
ID / OOD Text Embeddings
Cosine Classifier
Frozen
Learnable

Methods



$$s_i(\mathbf{x}') = \frac{\mathcal{I}(\mathbf{x}') \cdot \mathcal{T}(t_i)}{\|\mathcal{I}(\mathbf{x}')\| \cdot \|\mathcal{T}(t_i)\|}$$

$$S_{\text{MCM}}(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) = \max_i \frac{e^{s_i(\mathbf{x}')/\tau}}{\sum_{j=1}^{K} e^{s_j(\mathbf{x}')/\tau}},$$

$$G(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) = \begin{cases} 1 & S_{\text{MCM}}(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) \geq \lambda \\ 0 & S_{\text{MCM}}(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) < \lambda \end{cases},$$

Experiments

- **New insights on softmax scaling for zero-shot OOD detection.**

We provide theoretical justifications that **softmax scaling improves the separability** between ID and OOD data for CLIP-based OOD detection, which is contrary to models trained with cross-entropy (CE) loss.

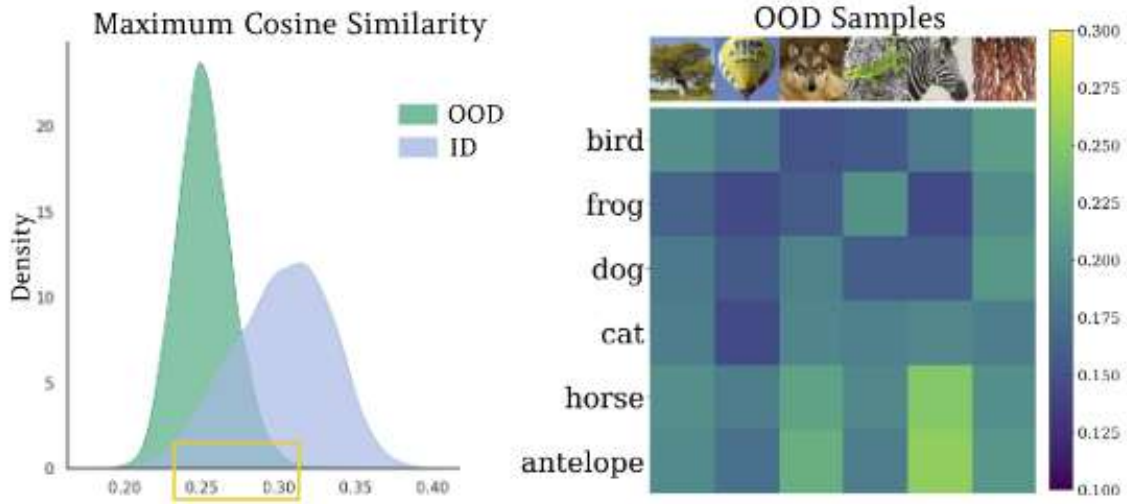


Figure 2: Left: Maximum cosine similarity for ID and OOD inputs. There exists overlapping regions (shown in yellow); Right: Cosine similarities between OOD inputs and ID concept vectors. For OOD inputs, the cosine similarities display uniformity.

Experiments

Metrics

(1) the false positive rate (FPR95) of OOD samples when the true positive rate of in-distribution samples is at 95%
(2) the area under the receiver operating characteristic curve (AUROC)
(3) ID classification accuracy (ID ACC)

Table 1: Zero-shot OOD detection with MCM score based on CLIP-B/16 with various ID datasets.

ID Dataset	OOD Dataset								Average	
	iNaturalist		SUN		Places		Texture			
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CUB-200 [80]	9.83	98.24	4.93	99.10	6.65	98.57	6.97	98.75	7.09	98.66
Stanford-Cars [39]	0.05	99.77	0.02	99.95	0.24	99.89	0.02	99.96	0.08	99.89
Food-101 [6]	0.64	99.78	0.90	99.75	1.86	99.58	4.04	98.62	1.86	99.43
Oxford-Pet [57]	2.85	99.38	1.06	99.73	2.11	99.56	0.80	99.81	1.70	99.62
ImageNet-10	0.12	99.80	0.29	99.79	0.88	99.62	0.04	99.90	0.33	99.78
ImageNet-20	1.02	99.66	2.55	99.50	4.40	99.11	2.43	99.03	2.60	99.32
ImageNet-100	18.13	96.77	36.45	94.54	34.52	94.36	41.22	92.25	32.58	94.48

Table 2: OOD detection performance for ImageNet-1k [11] as ID.

Method	OOD Dataset								Average	
	iNaturalist		SUN		Places		Texture			
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Requires training (or w. fine-tuning)										
MOS [32] (BiT)	9.28	98.15	40.63	92.01	49.54	89.06	60.43	81.23	39.97	90.11
Fort et al. [19] (ViT-B)	15.07	96.64	54.12	86.37	57.99	85.24	53.32	84.77	45.12	88.25
Fort et al. [19] (ViT-L)	15.74	96.51	52.34	87.32	55.14	86.48	51.38	85.54	43.65	88.96
Energy [48] (CLIP-B)	21.59	95.99	34.28	93.15	36.64	91.82	51.18	88.09	35.92	92.26
Energy [48] (CLIP-L)	10.62	97.52	30.46	93.83	32.25	93.01	44.35	89.64	29.42	93.50
MSP [25] (CLIP-B)	40.89	88.63	65.81	81.24	67.90	80.14	64.96	78.16	59.89	82.04
MSP [25] (CLIP-L)	34.54	92.62	61.18	83.68	59.86	84.10	59.27	82.31	53.71	85.68
Zero-shot (no training required)										
MCM (CLIP-B)	30.91	94.61	37.59	92.57	44.69	89.77	57.77	86.11	42.74	90.77
MCM (CLIP-L)	28.38	94.95	29.00	94.14	35.42	92.00	59.88	84.88	38.17	91.49

Experiments

Table 3: Performance comparison on **hard OOD detection** tasks. MCM is competitive on all three hard OOD tasks without training involved. MSP (based on fine-tuned CLIP) does not further improve performance.

Method	**ID** **OOD**	ImageNet-10 ImageNet-20	ImageNet-20 ImageNet-10	Waterbirds Spurious OOD
		FPR95 / AUROC	FPR95 / AUROC	FPR95 / AUROC
MSP [25] (fine-tuning)		9.38 / 98.31	12.51 / 97.70	39.57 / 90.99
Mahalanobis [42] (visual only)		78.32 / 85.60	43.03 / 89.94	2.21 / 99.55
MCM (zero-shot)		5.00 / 98.71	12.91 / 98.09	5.87 / 98.36

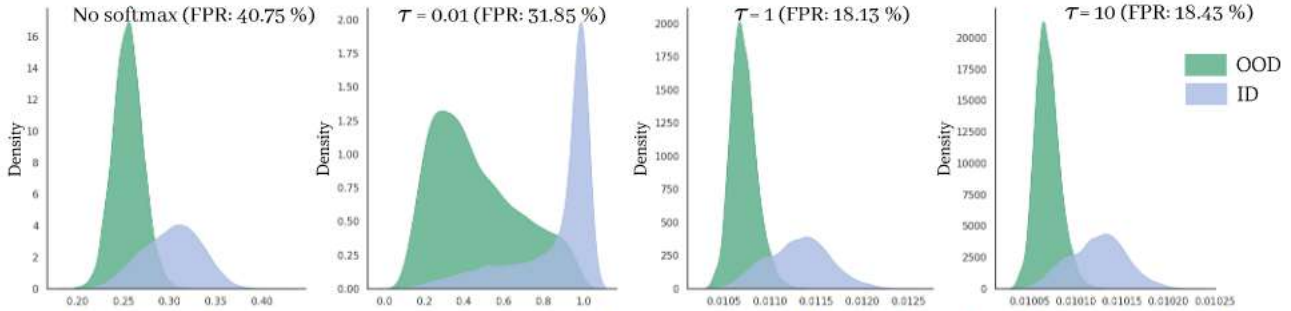


Figure 4: The influence of softmax scaling and temperature. We use ImgeNet-100 (ID) vs. iNaturalist (OOD). Softmax scaling with a moderate temperature significantly improves FPR95.

Table 7: ID classification accuracy on ImageNet-1k (%)

Method	**ID ACC**
zero-shot	
MCM (CLIP-B/16)	67.01
MCM (CLIP-L/14)	73.28
w. fine-tuning	
MSP (CLIP-B/16)	79.39
MSP (CLIP-L/14)	84.12
Energy [48] (CLIP-B/16)	79.39
Energy [48] (CLIP-L/14)	84.12
Fort et al. [19] (ViT-B/16)	81.25
Fort et al. [19] (ViT-L/14)	84.05
MOS [32] (BiT)	75.16

Experiments

Table 1: Zero-shot OOD detection with MCM score based on CLIP-B/16 with various ID datasets.

ID Dataset	OOD Dataset								Average	
	iNaturalist		SUN		Places		Texture			
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CUB-200 [80]	9.83	98.24	4.93	99.10	6.65	98.57	6.97	98.75	7.09	98.66
Stanford-Cars [39]	0.05	99.77	0.02	99.95	0.24	99.89	0.02	99.96	0.08	99.89
Food-101 [6]	0.64	99.78	0.90	99.75	1.86	99.58	4.04	98.62	1.86	99.43
Oxford-Pet [57]	2.85	99.38	1.06	99.73	2.11	99.56	0.80	99.81	1.70	99.62
ImageNet-10	0.12	99.80	0.29	99.79	0.88	99.62	0.04	99.90	0.33	99.78
ImageNet-20	1.02	99.66	2.55	99.50	4.40	99.11	2.43	99.03	2.60	99.32
ImageNet-100	18.13	96.77	36.45	94.54	34.52	94.36	41.22	92.25	32.58	94.48

Table 4: Zero-shot OOD detection of $S^{\text{wo}}_{\text{MCM}}$ based on CLIP-B/16.

ID Dataset	OOD Dataset								Average	
	iNaturalist		SUN		Places		Texture			
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Stanford-Cars [39]	0.00	100	0.02	99.99	0.26	99.94	0.00	100	0.07	99.98
Food-101 [6]	0.56	99.86	0.09	99.95	0.49	99.88	8.33	97.44	2.37	99.28
Oxford-Pet [57]	0.02	99.98	0.05	99.97	0.20	99.94	0.27	99.91	0.14	99.95
ImageNet-10	2.40	99.42	1.79	99.55	2.83	99.32	1.86	99.56	2.22	99.46
ImageNet-20	14.96	97.87	13.10	97.97	14.21	97.67	13.46	97.32	13.93	97.71
ImageNet-1k	61.66	89.31	64.39	87.43	63.67	85.95	86.61	71.68	69.08	83.59

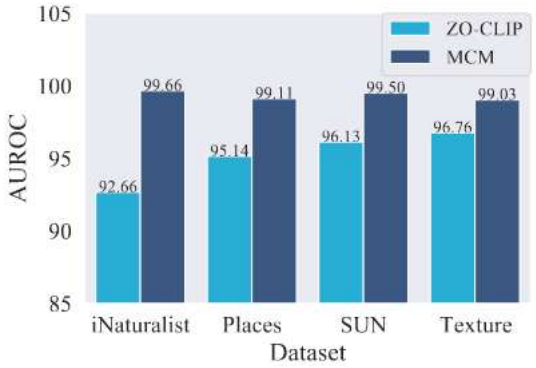


Figure 3: Comparison with a candidate label-based score ZO-CLIP on ImageNet-20, based on our implementation of [16]. Implementation details are deferred to Appendix E.1.

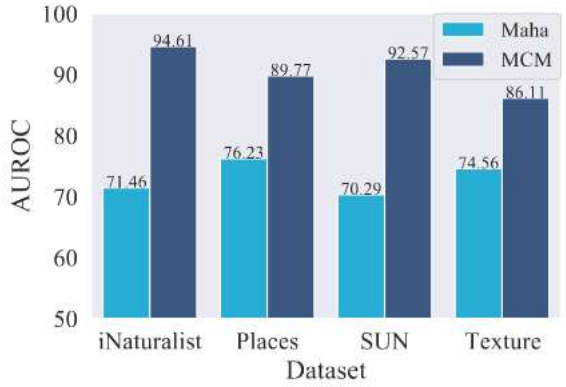


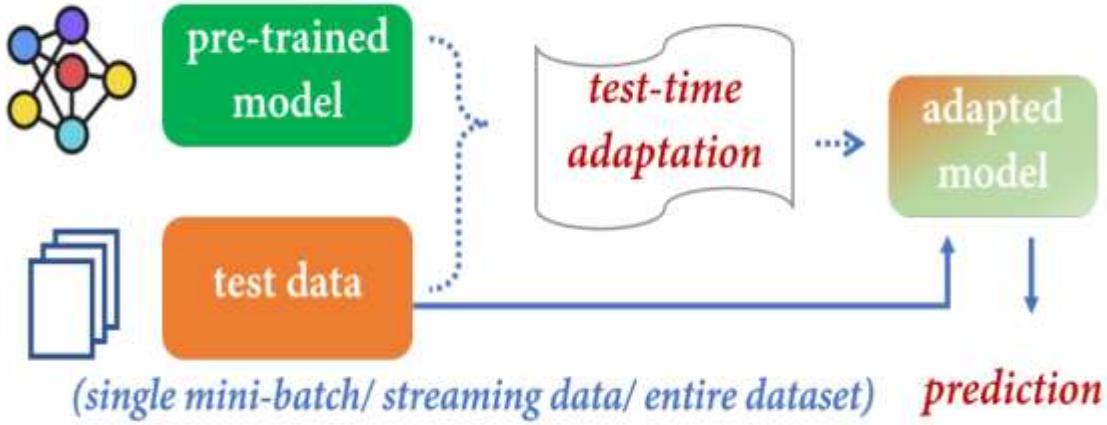
Figure 5: Comparison with Mahalanobis (Maha) score on ImageNet-1k.

Effectiveness of Vision Language Models for Open-world Single Image Test Time Adaptation

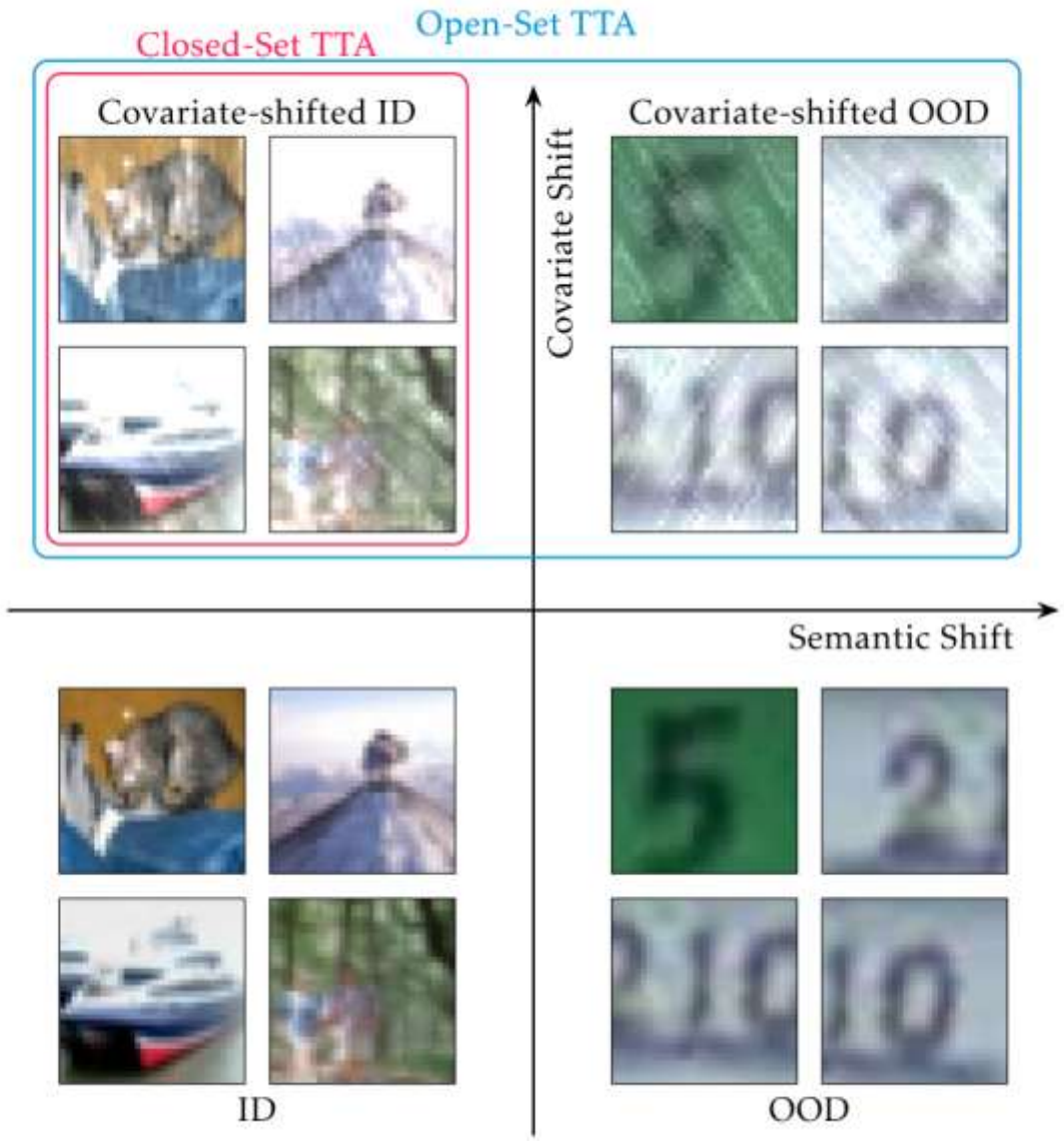
Manogna Sreenivas, Soma Biswas
Indian Institute of Science
Bengaluru, India
{manognas, somabiswas}@iisc.ac.in

arxiv 2024

Background



test-time adaptation(TTA)



Motivation

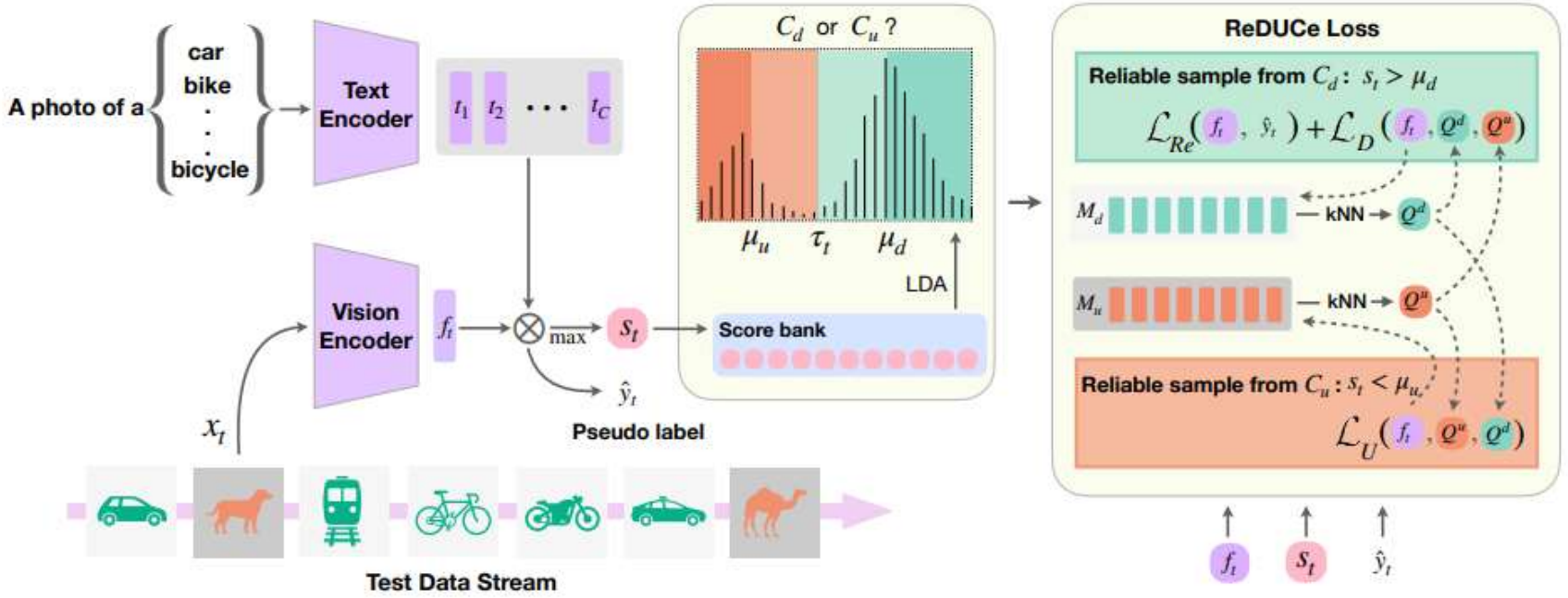
setting

test data contains both **covariate shift and semantic shift**.

- **covariate shift：** the test images differ from the training data not only in appearance factors like lighting and background
- **semantic shift：** but also may include new classes that were unseen during training

ROSITA aims to address both challenges simultaneously, enabling the model to adapt online to changing environments while also **detecting and rejecting unknown classes.**

Methods



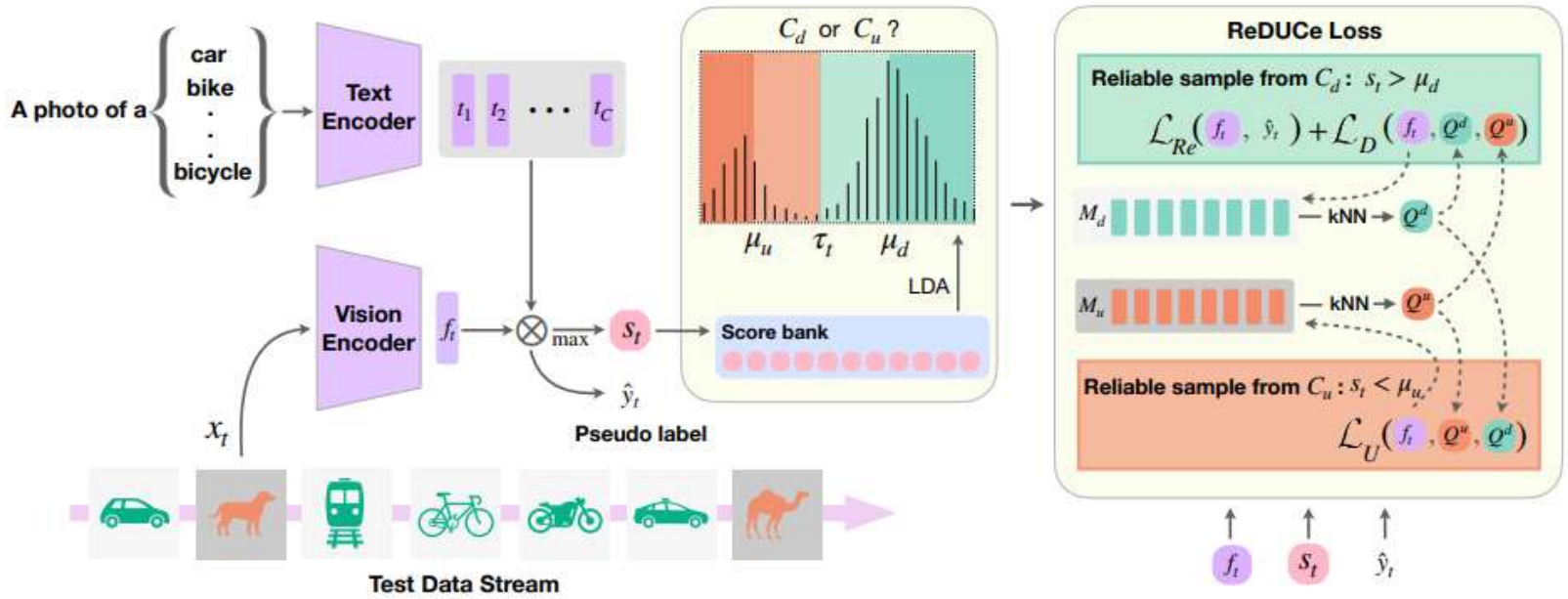
$$s_t^{ood} = \max_k \text{sim}(f_t, t_k); \quad k \in \{1, \dots C\}$$

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

where $\mathcal{S}_d = \{s_i | s_i > \tau, s_i \in S\}$ and $\mathcal{S}_u = \{s_i | s_i < \tau, s_i \in S\}$ represent the scores of samples classified as desired and undesired, respectively, and μ_d and μ_u are their means. Using this threshold, the test sample x_t is classified as:

$$\tilde{y}_t = \begin{cases} \text{desired} & \text{if } s_t \geq \tau_t^* \\ \text{undesired} & \text{if } s_t < \tau_t^* \end{cases} \tag{3}$$

Methods



Case 1: Reliable Desired sample

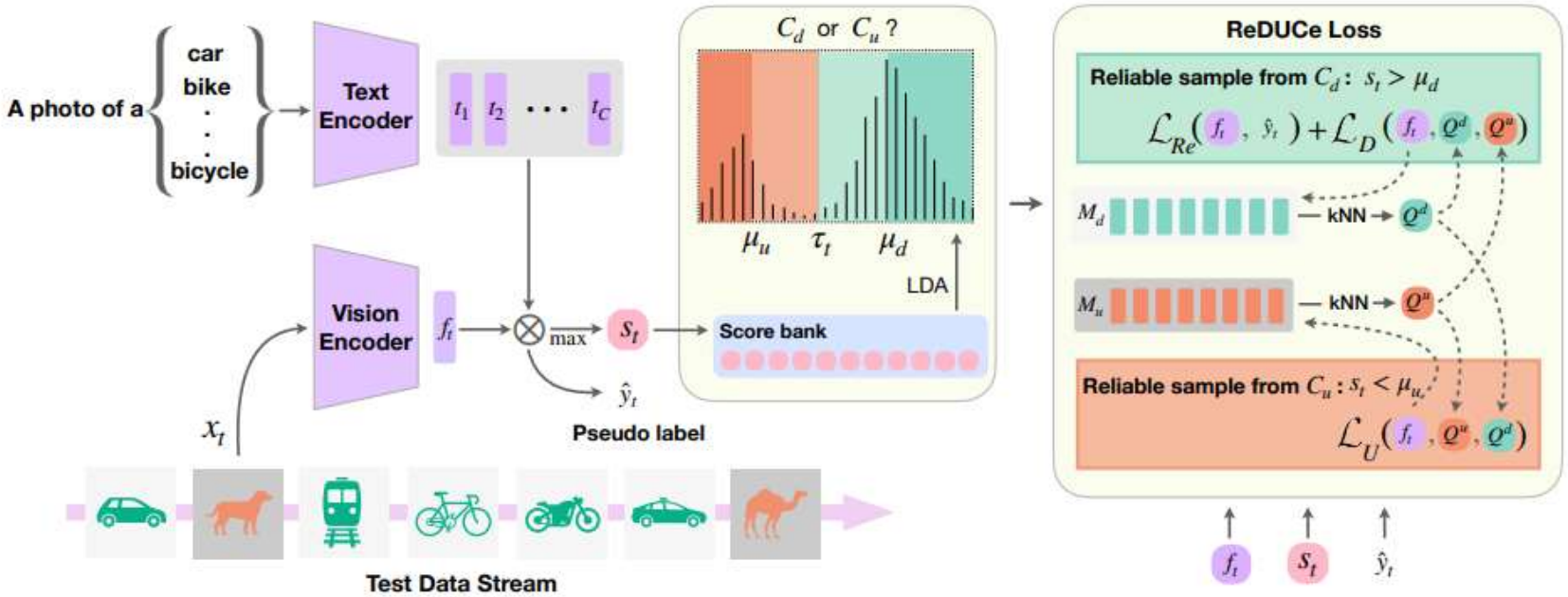
$$Q_w = \mathrm{kNN}(f_t; \mathcal{M}_w); \quad Q_s = \mathrm{kNN}(f_t; \mathcal{M}_s)$$

$$\mathcal{L}_{Re} = \mathcal{L}_{CE}(x_t, \hat{y}_t) + \mathcal{L}_{CE}(\tilde{x}_t, \hat{y}_t); \quad \hat{y}_t = \mathrm{argmax}_i \, \mathrm{sim}(f_t, t_i) \tag{5}$$

$$\mathcal{L}_D = -\frac{1}{K^+} \sum_{z^+ \in Q_d} \mathbf{1}(y^+ = \hat{y}_t) \log \frac{\exp\left(\mathrm{sim}\left(f_t, z^+\right)/\tau\right)}{\sum_{z^- \in Q_u} \exp(\mathrm{sim}(f_t, z^-)/\tau)} \tag{6}$$

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

Experiments



Case 2: Reliable unDesired sample

$$\mathcal{L}_U = -\frac{1}{K} \sum_{z^+ \in Q_u} \log \frac{\exp\left(\mathrm{sim}\left(f_t, z^+\right)/\tau\right)}{\sum_{z^- \in Q_d} \exp(\mathrm{sim}(f_t, z^-)/\tau)} \tag{7}$$

$$\mathcal{L}_{ReDUCe} = \begin{cases} \mathcal{L}_{Re} + \mathcal{L}_D & \text{if} \quad s_t > \mu_d \\ \mathcal{L}_U & \text{if} \quad s_t < \mu_u \end{cases} \tag{8}$$

Experiments

Table 3: Results with CIFAR-10C/100C as desired class data D_d and four other datasets as D_u.

		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
		UniEnt	91.98	85.2	75.62	89.97	64.38	74.18	91.40	26.96	74.73	82.59	68.14	68.98
		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
		DPE	46.97	99.10	27.60	84.15	85.24	68.52	89.92	31.30	69.90	79.18	75.06	62.34
		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
		(K+1)PC	98.34	9.63	86.52	71.01	78.78	68.70	71.20	85.81	68.29	62.35	88.44	61.89
		UniEnt	98.17	5.49	82.64	98.35	7.85	83.65	90.90	27.41	76.08	86.16	51.91	71.72
		TDA	98.42	4.13	81.97	98.60	6.20	83.95	91.27	27.00	76.84	86.72	51.40	72.61
		DPE	83.82	92.73	55.52	97.42	12.95	79.41	89.10	31.13	74.32	73.57	73.67	53.64
		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
		UniEnt	77.94	99.93	48.32	64.78	98.61	45.84	67.40	73.77	45.83	63.28	93.18	44.04
		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
		DPE	67.06	99.88	42.54	43.23	99.79	35.69	61.42	80.62	42.80	60.08	92.80	42.21
		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
		(K+1)PC	96.49	9.42	62.97	65.73	78.63	32.60	42.94	99.95	27.52	53.48	94.26	34.70
		UniEnt	87.40	64.02	54.86	92.99	40.36	56.42	68.84	74.26	48.41	66.93	87.96	46.09
		TDA	89.82	52.24	55.46	95.04	30.76	59.51	72.05	71.83	49.19	69.12	87.36	49.06
		DPE	39.05	98.88	33.66	84.29	76.13	52.20	63.74	82.75	45.74	65.61	90.67	46.36
		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

Experiments

Table 13: Comparison of C_d vs C_u class identifiers: MSP vs DAF vs LDA. The three thresholds for ReDUCe loss correspond to $\tau_u/\tau_t/\tau_d$ where τ_u and τ_d is used to identify reliable test samples and τ_t is used to distinguish between C_d and C_u samples. In the case of DAF with ReDUCe loss, we use the estimated means μ_d^* and μ_u^* of the two Gaussian mixture components to identify reliable samples.

C_d vs C_s	Threshold	Test-time objective	D_u: MNIST				
			C-10C	C-100C	IN-C	IN-R	VisDA
MSP	0.4/0.6/0.8	ReDUCe	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
LDA	$s_t > \tau_t$	UniEnt	75.62	48.31	41.53	71.73	78.09
DAF	$\pi(x_t) > 0.5$		79.43	50.12	46.52	79.30	86.79
LDA	$\mu_u/\tau_t/\mu_d$	ReDUCe	**84.17**	**57.34**	**48.53**	**83.53**	90.64
DAF	$\mu_u^*/0.5/\mu_d^*$		83.56	55.37	48.33	83.32	**90.97**

Table 14: Accuracy on updating different parameter groups on CIFAR-10C and ImageNet-R datasets.

Optimizer	Parameters	CIFAR-10C					ImageNet-R				
		$1e^{-6}$	$1e^{-5}$	$1e^{-4}$	$1e^{-3}$	$1e^{-2}$	$1e^{-6}$	$1e^{-5}$	$1e^{-4}$	$1e^{-3}$	$1e^{-2}$
SGD	Prompts	73.40	31.04	12.53	11.18	10.19	73.97	74.17	74.71	25.68	10.63
	Full	10.48	10.44	9.99	10.00	10.01	14.18	7.19	0.65	0.65	0.42
	First Block	75.1	76.12	78.27	13.07	10.01	73.84	74.31	74.91	8.76	0.32
	Last Block	73.45	72.42	59.44	10.17	10.02	75.95	77.93	24.82	0.52	0.67
	Prompts+LN	73.82	46.77	24.71	10.24	10.18	73.76	75.09	76.35	28.72	11.74
	LoRA Adapters	73.86	73.90	75.42	83.15	12.58	73.51	73.57	74.22	77.39	34.83
	LayerNorm	74.35	76.61	80.41	**84.58**	11.69	74.13	74.35	75.23	**76.92**	33.07
AdamW	Prompts	72.40	18.6	12.83	10.04	10.08	74.4	75.17	27.93	6.82	4.37
	Full	10.32	10.03	10.00	10.00	9.97	14.83	0.95	0.28	0.52	0.66
	First Block	79.05	24.70	10.84	10.00	10.00	74.6	74.8	5.68	0.26	0.15
	Last Block	59.23	10.84	10.49	10.00	10.01	77.44	10.67	0.51	0.25	0.33
	Prompts+LN	75.01	72.10	21.92	13.33	10.01	74.52	76.45	12.99	8.87	5.55
	LoRA Adapters	77.64	81.55	14.01	10.25	10.02	74.34	76.14	18.63	2.26	0.62
	LayerNorm	76.10	81.57	**85.9**	85.27	10.03	73.96	75.64	78.28	**78.81**	31.47

原因

①无法准确判断样本是期望类还是非期望类 → 误用熵最小化（本应用于期望类）或熵最大化（用于非期望类）→ 错误更新模型。

UniEnt 不能显式拉开特征空间中不同类的距离，只是对熵进行调控 → 分类边界不清晰 → 高误报率（FPR）

②ReDUCe loss它不依赖得分分布是否清晰，而是直接优化特征空间的几何结构，显式推动非期望类远离期望类和文本原型，避免了重叠问题。

Experiments

Table 16: Open-set CTTA performance for CIFAR-10C and CIFAR-100C as desired datasets.

Method	CIFAR-10C				CIFAR-100C			
	SVHN	MNIST	Tiny	C-100C	MNIST	SVHN	Tiny	C-10C
ZSEval	64.33	64.04	66.50	58.49	39.00	36.29	38.41	35.04
TPT	64.26	64.03	66.50	58.47	39.00	36.24	38.38	34.45
(K+1)PC	65.13	62.52	66.93	57.46	40.64	37.05	38.23	34.55
TDA	66.02	66.44	67.64	59.44	40.49	40.35	39.92	35.42
DPE	23.36	50.12	58.96	35.56	30.75	19.23	35.85	27.62
ROSITA	**66.86**	**65.26**	**68.89**	**59.16**	**41.64**	**38.02**	**40.44**	**36.05**

CCC benchmark (Press et al., 2023) was specifically introduced to assess the long-term continual adaptation behavior of TTA methods in a changing world, covering scenarios such as weather changing from foggy to rainy, day to night. We experiment with the CCC dataset where gradual domain changes are synthesized across 15 corruptions of ImageNet-C as D_d and MNIST as D_u for a sequence length of 300k samples. From Table 17, we observe that ROSITA consistently outperforms prior methods even in this challenging CCC dataset.

Table 17: Results on open-set CCC benchmark.

Method	CCC/MNIST				
	AUC ↑	FPR ↓	Acc_D ↑	Acc_U ↑	Acc_{HM} ↑
ZS-Eval	88.45	88.24	18.37	99.53	31.01
(K+1)PC	95.82	20.16	20.37	99.97	33.84
TDA	87.42	85.83	20.42	99.35	33.87
UniEnt	89.90	82.86	18.99	99.62	31.90
DPE	84.68	87.56	18.47	99.40	31.16
ROSITA	**96.02**	**19.96**	**21.14**	**99.11**	**34.84**

Thanks