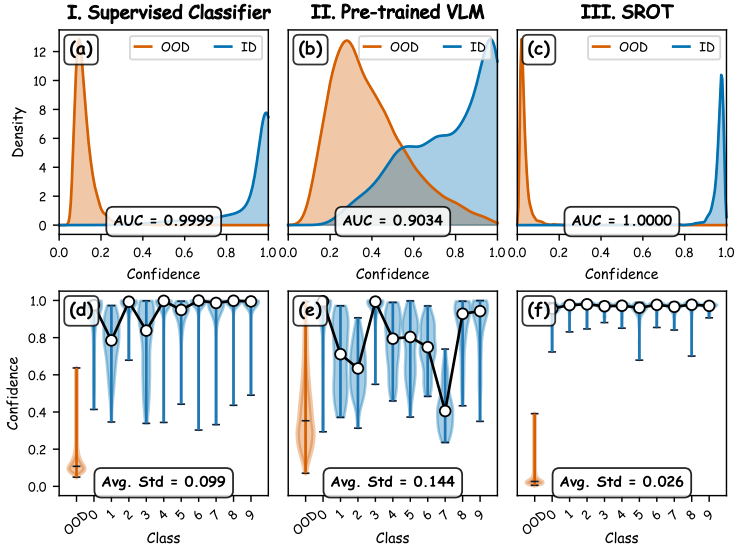
Test-Time Adaptation via Self-Reinforced Optimal Transport for Zero-Shot OOD Detection with Vision-Language Models

Mingcai Chen Heng-yang Lu Yuntao Du
Baoming Zhang Hao Zhou

chenmc@njupt.edu.cn

Motivation: The Confidence Bottleneck



- **Supervised model:** clear ID/OOD separation.
- **Zero-shot CLIP:** strongly overlapping confidence scores.
- **SROT:** restores separation using class-wise posterior structure.

Figure: Oxford-IIIT Pet (ID), iNaturalist (OOD).

Why a Better Threshold Is Not Enough

- Standard zero-shot OOD detection thresholds maximum CLIP confidence:

$$C(x) = \max_k P_{\mathrm{CLIP}}(x)_k.$$

- We analyze single-threshold separation with the Youden/KS index:

$$J^* = \sup_\theta \big[\mathrm{TPR}(\theta) - \mathrm{FPR}(\theta)\big].$$

- Under the stated label-independence and stochastic-monotonicity assumptions for the confidence transformation,

$$\boxed{J_t^* \le J_s^*}.$$

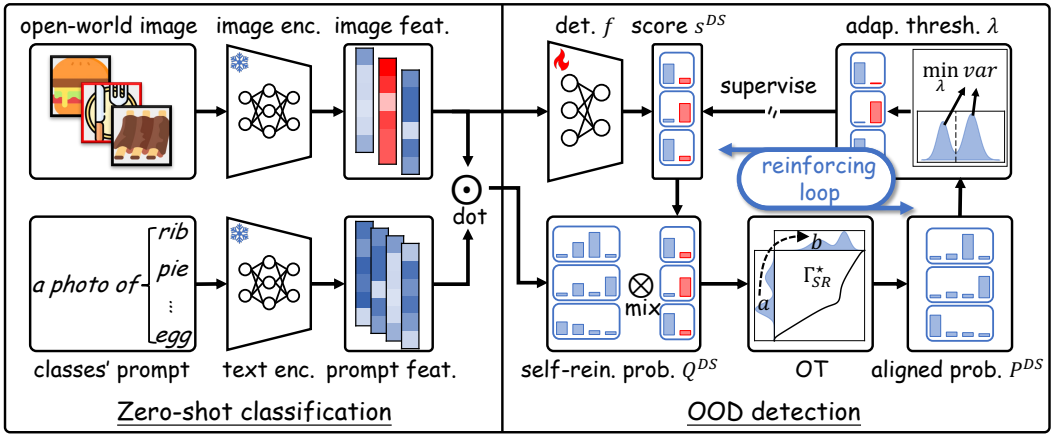
Takeaway: Reshape the full posterior instead of only tuning its threshold.

SROT Step 1: Optimal-Transport Rectification

1. For an unlabeled test batch, obtain CLIP posteriors $Q \in [0,1]^{B\times K}$ and set the transport cost $M = -\log Q$.
2. Apply entropic optimal transport with sample marginal $a = \frac{1}{B}\mathbf{1}_B$ and class prior $b = \frac{1}{K}\mathbf{1}_K$:
$$\Gamma^* = \arg\min_{\Gamma \in \mathcal{U}(a,b)} \langle \Gamma, -\log Q \rangle + \varepsilon\, \mathrm{KL}(\Gamma \| ab^\top).$$
3. Normalize the transport rows and derive an OOD score:
$$\widetilde{P}_{i,:} = \Gamma^*_{i,:}/a_i, \qquad s^{\mathrm{OT}}(x_i) = 1 - \max_k \widetilde{P}_{i,k}.$$

Intuition: Balance batch-level class assignments while preserving CLIP's image–text similarity preferences.

SROT Step 2: Detector and Self-Reinforcing Feedback



*Abbreviations: enc. -> encoder; feat. -> feature; det. -> detector; rein. -> reinforced; adap. thresh. -> adaptive thresholding; prob. -> probability; ❄ -> frozen weights; 🔥 -> trainable weights.

OT scores → adaptive pseudo-labels → lightweight OOD detector → feedback into OT.

Experimental Results

- **Nine ID datasets** and four OOD datasets; all CLIP encoders remain frozen.
- Harmonic accuracy Acc_H balances correct ID recognition and OOD rejection.

Average Acc_H (%)	AdaND	**SROT**
Nine ID datasets	73.13	**75.66**
CUB-200-2011	66.14	**70.01**
ImageNet	73.09	**76.14**
ImageNet-Sketch	53.77	**61.42**

- On ImageNet as ID, average AUROC rises from **95.58%** to **98.40%**, and FPR95 drops from **16.00%** to **6.60%**.

Ablation and Takeaway

CUB-200-2011 ablation	Acc_H (%)
Without OT	66.37
OT without detector feedback	69.12
Full SROT	**70.07**

- **Posterior alignment** improves OOD pseudo-label quality.
- **Feature-level detector feedback** further refines the OT cost.
- SROT needs **no target labels** and **does not update CLIP**.

Full posteriors + optimal transport + feedback yield stronger OOD detection.

Thanks!

chenmc@njupt.edu.cn

Any questions or comments are welcome.