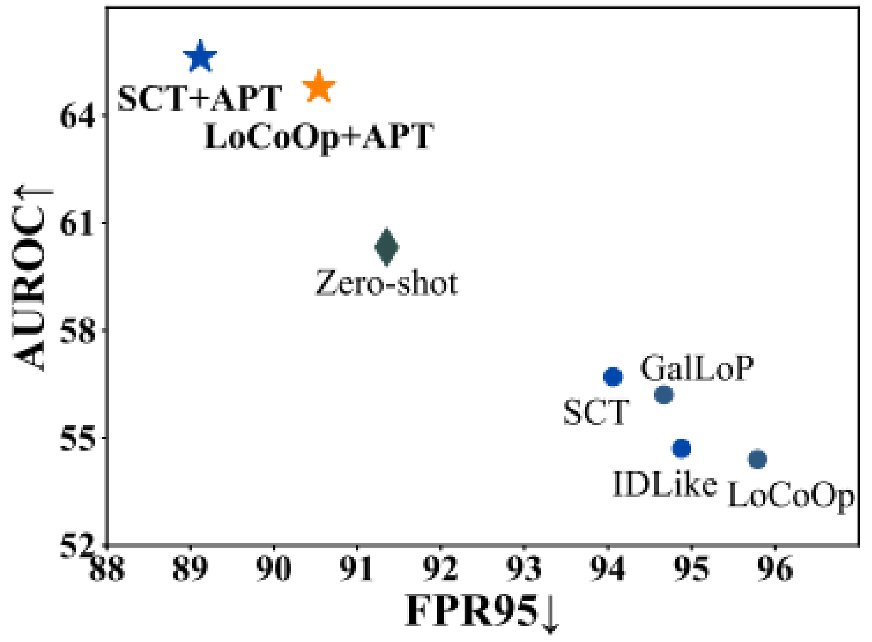
Auxiliary Prompt Tuning of Vision-Language Models for Few-Shot Out-of-Distribution Detection

Wenjun Miao, Guansong Pang, Zihan Wang, Jin Zheng, Xiao Bai

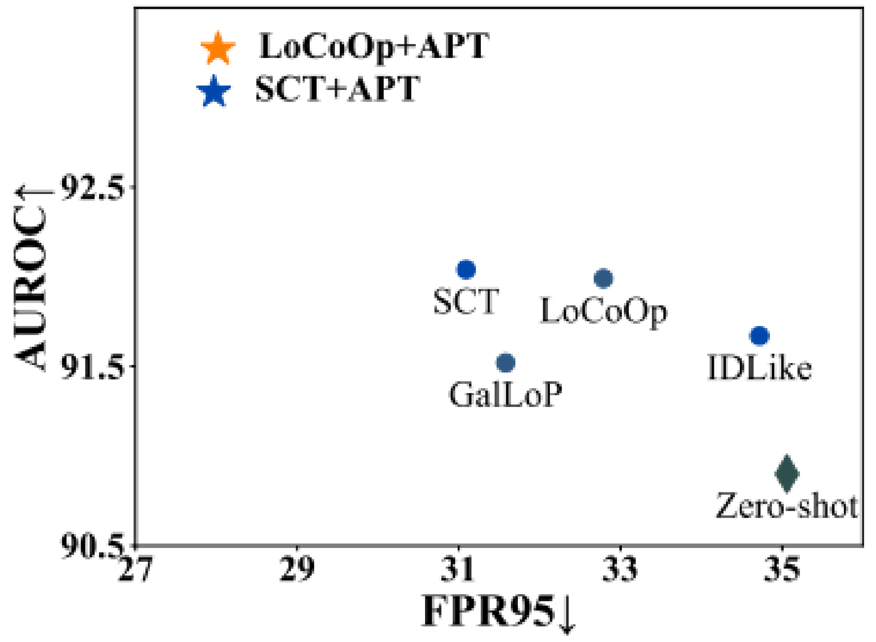
Paper Link

GitHub Page

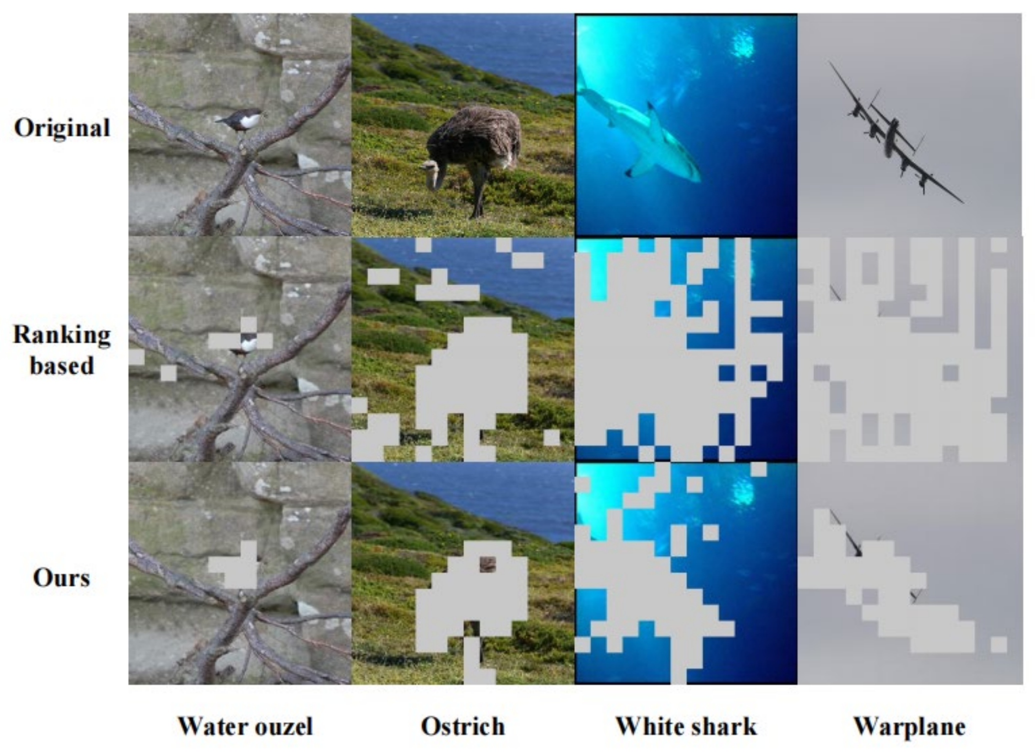
Motivation



(a) Hard OOD benchmark.



(b) Standard OOD benchmarks.



- Current prompt tuning-based OOD detectors rely on the background regions of few-shot ID data to obtain OOD features, which are often not diverse enough and can lead to significant performance drop on challenging scenarios, *e.g.*, hard OOD.
- We propose **APT** to leverage auxiliary/outlier data to incorporate diverse OOD features into prompt tuning for OOD detection. To make it possible, we also curate an auxiliary dataset for well-established OOD detection benchmarks (ID data: ImageNet-1k).

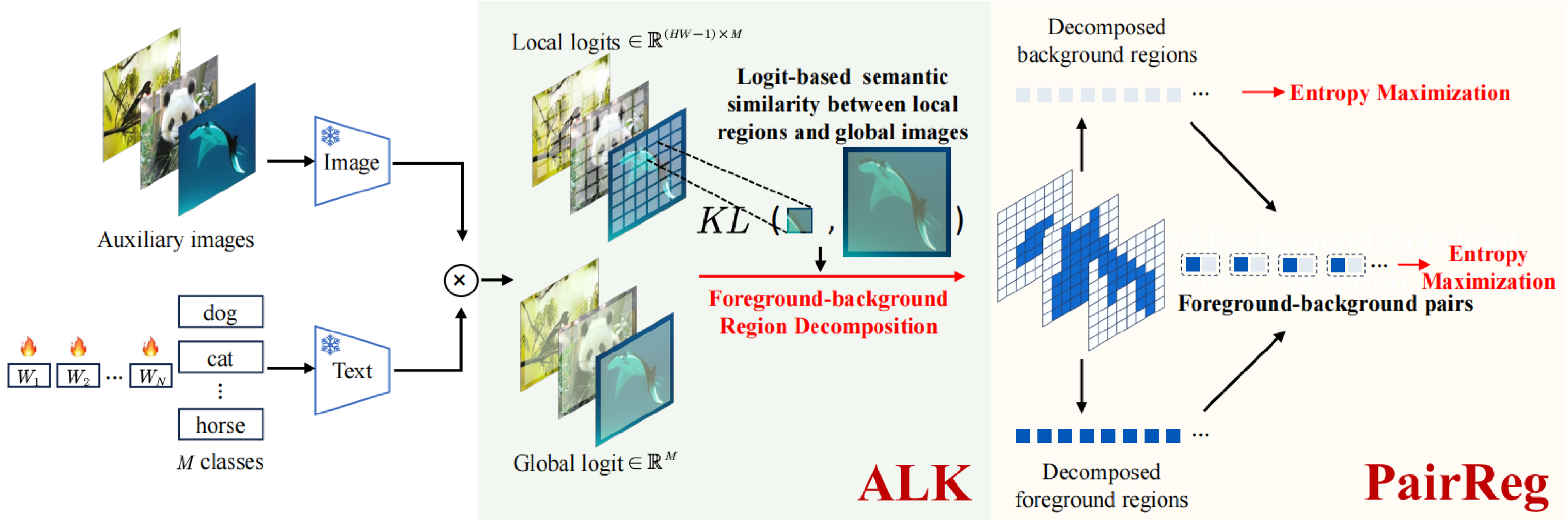
Experiments

OOD Data	Standard OOD Data										Hard OOD Data	
Method	iNaturalist		SUN		Places365		Textures		Average		ImageNet-1k-OOD	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
1-shot												
LoCoOp [21]	23.53	94.89	24.15	94.55	32.84	91.51	50.67	87.01	32.79	91.99	95.79	54.40
LoCoOp+APT	13.53	97.09	19.62	95.81	29.08	92.95	49.84	87.24	**28.02**	**93.27**	**90.54**	**64.78**
IDLike [2]	12.07	97.65	40.55	91.07	47.94	88.31	38.34	89.67	34.72	91.67	94.88	54.70
IDLike+APT	10.14	98.22	33.82	92.49	41.54	90.07	38.71	90.02	**31.05**	**92.70**	**90.13**	**63.49**
SCT [35]	19.16	95.70	23.52	94.58	32.81	91.23	48.87	86.66	31.09	92.04	94.06	56.70
SCT+APT	13.94	96.89	18.93	95.80	28.04	92.96	50.96	86.49	**27.97**	**93.03**	**89.12**	**65.62**
16-shot												
LoCoOp [21]	17.58	96.30	22.82	95.20	32.21	92.03	45.27	88.86	29.47	93.10	93.86	57.29
LoCoOp+APT	13.29	97.05	20.43	95.56	28.65	92.98	46.40	87.98	**27.19**	**93.39**	**89.11**	**65.24**
IDLike [2]	9.71	98.05	38.93	90.54	47.06	88.06	32.82	91.89	32.12	92.14	93.07	57.59
IDLike+APT	8.22	98.46	31.52	93.01	37.57	91.32	33.73	91.22	**27.76**	**93.50**	**89.24**	**65.48**
SCT [35]	13.94	95.86	20.55	95.33	29.86	92.24	41.51	89.06	26.47	93.37	93.43	57.72
SCT+APT	9.70	97.79	20.12	95.52	28.54	92.84	45.78	88.43	**26.03**	**93.64**	**87.64**	**65.86**

- Comprehensive results show that APT achieves SotA performance on standard and hard OOD benchmarks, with significant improvements in challenging scenarios, *e.g.*, reducing ~5% FPR95 in 1-shot hard OOD tasks.

Miyai, Atsuyuki, et al. "Locoop: Few-shot out-of-distribution detection via prompt learning." Advances in Neural Information Processing Systems 36 (2023): 76298-76310.
Bai, Yichen, et al. "Id-like prompt learning for few-shot out-of-distribution detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
Yu, Geng, et al. "Self-calibrated tuning of vision-language models for out-of-distribution detection." Advances in Neural Information Processing Systems 37 (2024): 56322-56348.

The Proposed Method APT



- Two main modules of APT: Adaptive Logit-based KL Divergence (**ALK**) & Foreground-background Pair Regularization (**PairReg**)
- **ALK** decomposes foreground-background regions of outlier images in an **unsupervised** way (outlier classes are unknown) by measuring the semantic similarities of local regions to the global features via KL divergence between their classification logits:

$$J_x^{back} = \{i \in I : \mathrm{KL}(\mathrm{S}(p(x)) \parallel \mathrm{S}(p(x_i))) > \epsilon\},$$

$$J_x^{fore} = \{i \in I : \mathrm{KL}(\mathrm{S}(p(x)) \parallel \mathrm{S}(p(x_i))) \le \epsilon\},$$

where x is an input image and x_i is its image patch, $\mathrm{S}(\cdot)$ is a Softmax operation, and $p(x_i)$ is defined as follows:

$$p(x_i) = \frac{\exp\left(\mathrm{sim}\left(\boldsymbol{f}^i, \boldsymbol{g}_m\right)/\tau\right)}{\sum_{m'=1}^{M} \exp\left(\mathrm{sim}\left(\boldsymbol{f}^i, \boldsymbol{g}_{m'}\right)/\tau\right)}$$

where $\boldsymbol{f}^i$ are local features from image encoder and $\boldsymbol{g}_m$ are embeddings of a text prompt of an ID class m. $p(x)$ can be obtained in a similar way.

Method	Standard OODs		ImageNet-1k-OOD	
	FPR95	AUROC	FPR95	AUROC
SCT	31.09	92.04	94.06	56.70
SCT+OE [9]	36.47	90.17	96.26	55.13
SCT+EnergyOE [16]	40.52	86.21	96.98	54.22
SCT+APT	**27.97**	**93.03**	**89.12**	**65.62**

- **PairReg** first constructs foreground-background pairs for each foreground region and performs joint regularization on the background regions and paired instances via entropy maximization.

$$p_{pair}(x_i) = p(x_i) + p(\hat{x}_j)$$

$$\mathcal{L}_{APT} = \mathbb{E}_{x \sim X_{aux}}[\alpha \ell_{OOD}(p(J_x^{back})) + \beta \ell_{OOD}(p_{pair}(J_x^{fore}))]$$

$$\mathcal{L} = \mathcal{L}_{ID} + \mathcal{L}_{APT}$$

Shot	*Back*	*Pair*	Standard OODs		ImageNet-1k-OOD	
			FPR95↓	AUROC↑	FPR95↓	AUROC↑
1-shot	✗	✗	31.09	92.04	94.06	56.70
	✓	✗	29.02	92.88	93.14	58.65
	✗	✓	29.57	91.46	90.35	63.10
	✓	✓	**27.97**	**93.03**	**89.12**	**65.62**
16-shot	✗	✗	26.47	93.37	93.43	57.72
	✓	✗	26.08	93.60	92.07	59.41
	✗	✓	26.39	93.44	89.27	64.52
	✓	✓	**26.03**	**93.64**	**87.64**	**65.86**