Envisioning Outlier Exposure by Large Language Models for Out-of-Distribution Detection

Chentao Cao

Hong Kong Baptist University

with Zhun Zhong, Zhanke Zhou, Yang Liu, Tongliang Liu, and Bo Han

ICML 2024 Paper: https://arxiv.org/pdf/2406.00806.pdf Code: https://github.com/tmlr-group/EOE

Outline

- Background
- Method
- Experiments
- Summary

Background

Out-of-distribution (OOD) detection is essential to ensuring the reliability of machine learning systems

It can be viewed as a binary classification problem:

$$G_\lambda(x) = \begin{cases} \text{ID} & S(\mathbf{x}) \geq \lambda \\ \text{OOD} & S(\mathbf{x}) < \lambda \end{cases},$$

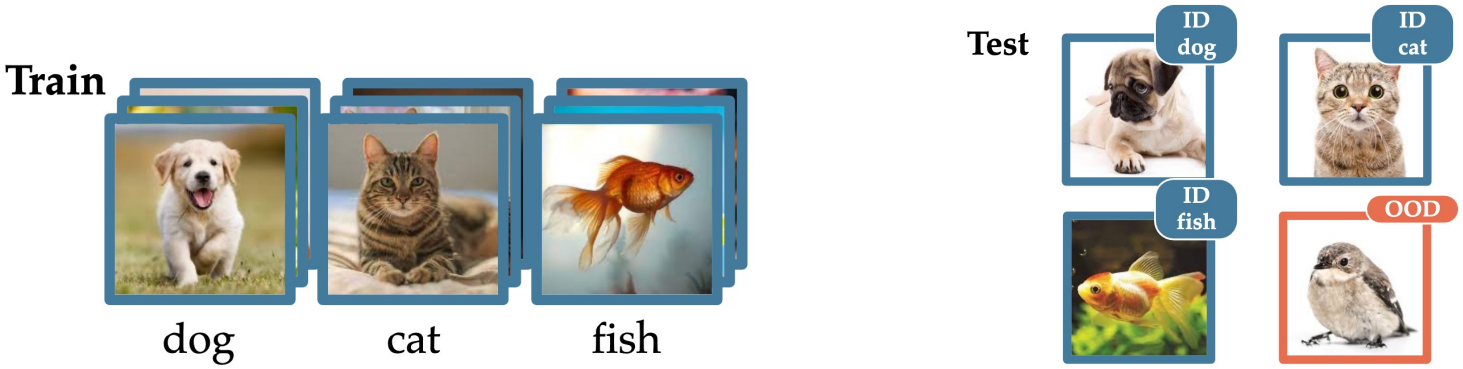
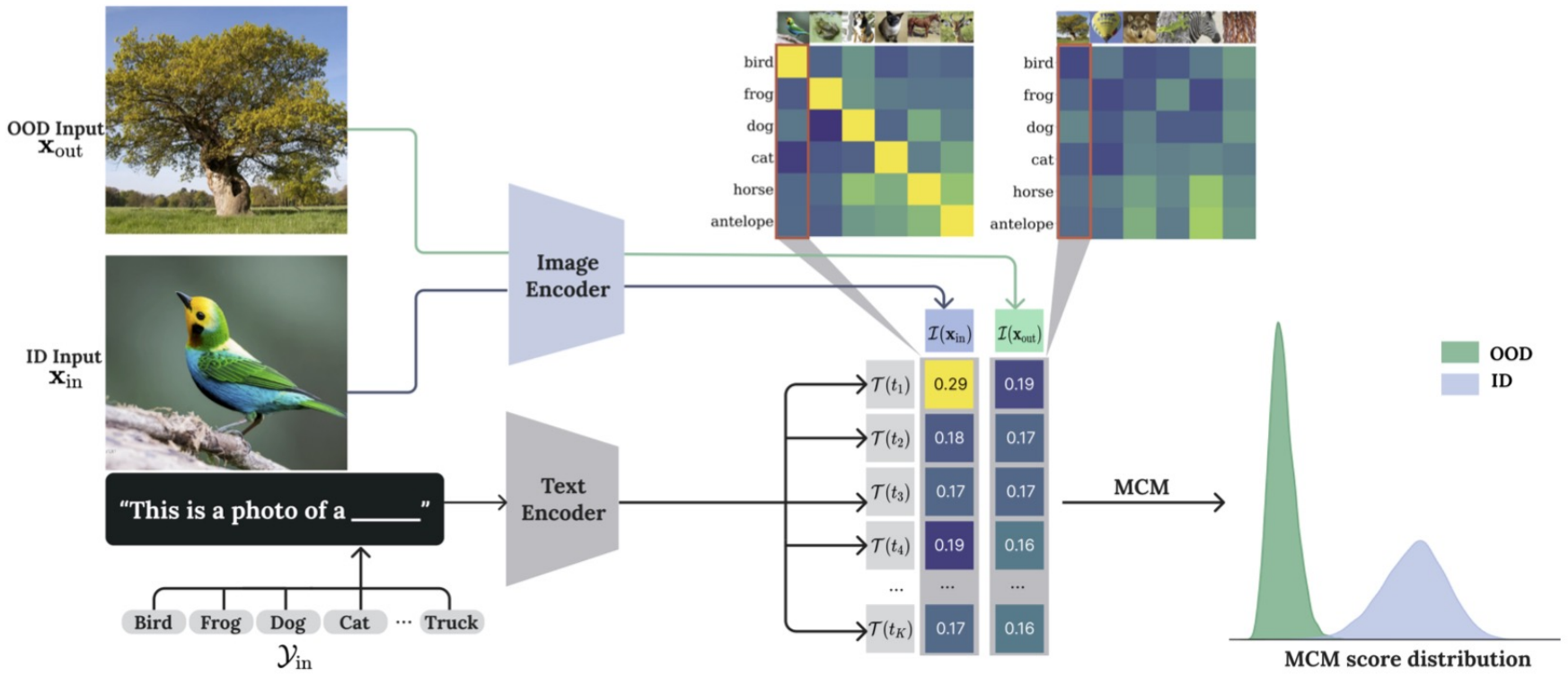


Image Source: J. Yang, et al, Generalized Out-of-Distribution Detection: A Survey. arXiv preprint arXiv:2110.11334, 2021.

Background

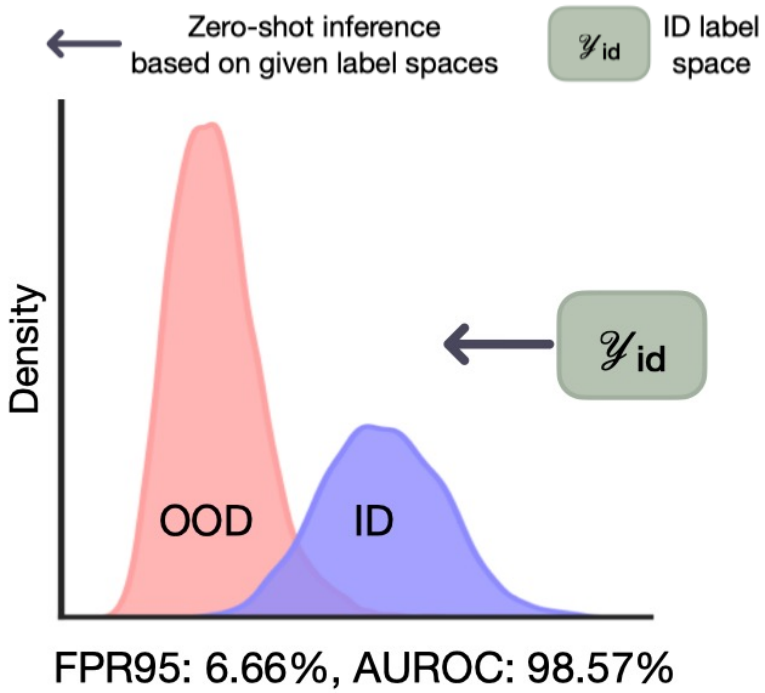
Pre-trained VLMs (e.g., CLIP) enable zero-shot OOD Detection



Y. Ming, et al. Delving into Out-of-Distribution Detection with Vision-Language Representations, In NeurIPS, 2022.

Background

Such an approach often fails when encountering hard OOD samples



(a) **Closed-Set:** Using only closed-set ID classes

We wonder 🤔

1) if this issue arises because the pre-trained VLMs are not strong enough

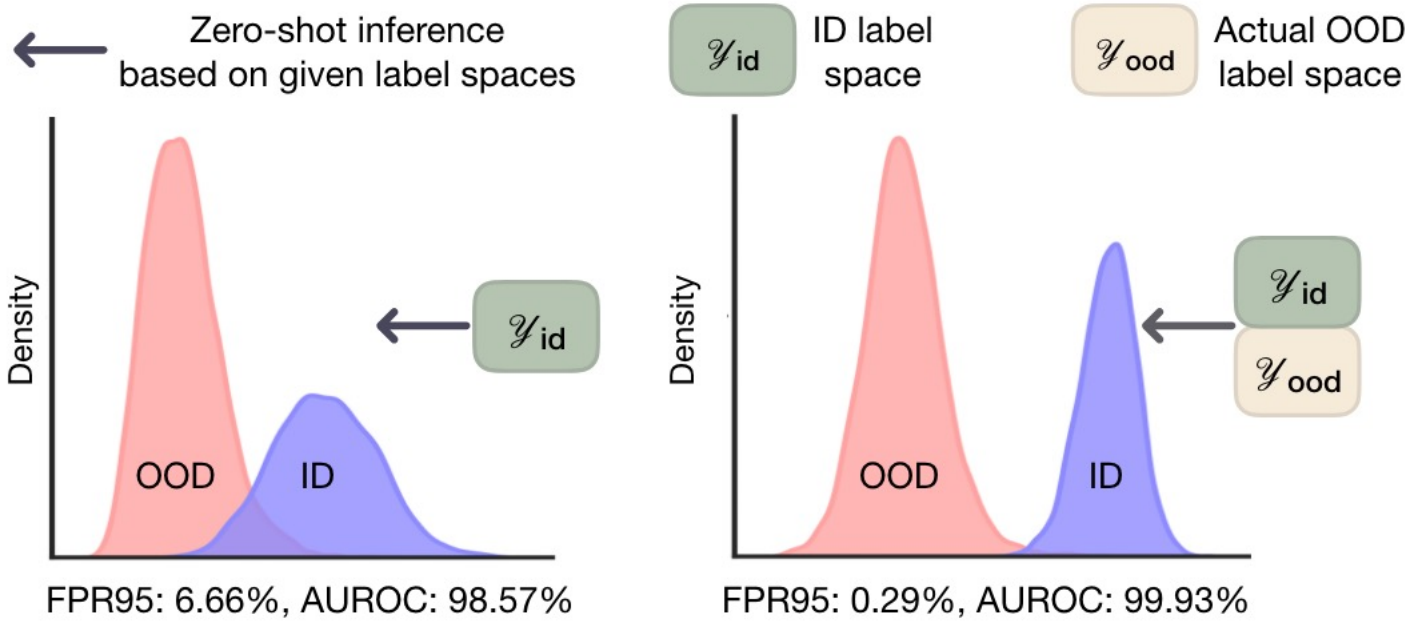
or

2) if it is attributable to the usages of these pretrained models, e.g., an exclusive reliance on closed-set ID classes

ID dataset: CUB-200-2011
OOD dataset: Places

Background

Incorporating with actual OOD class labels



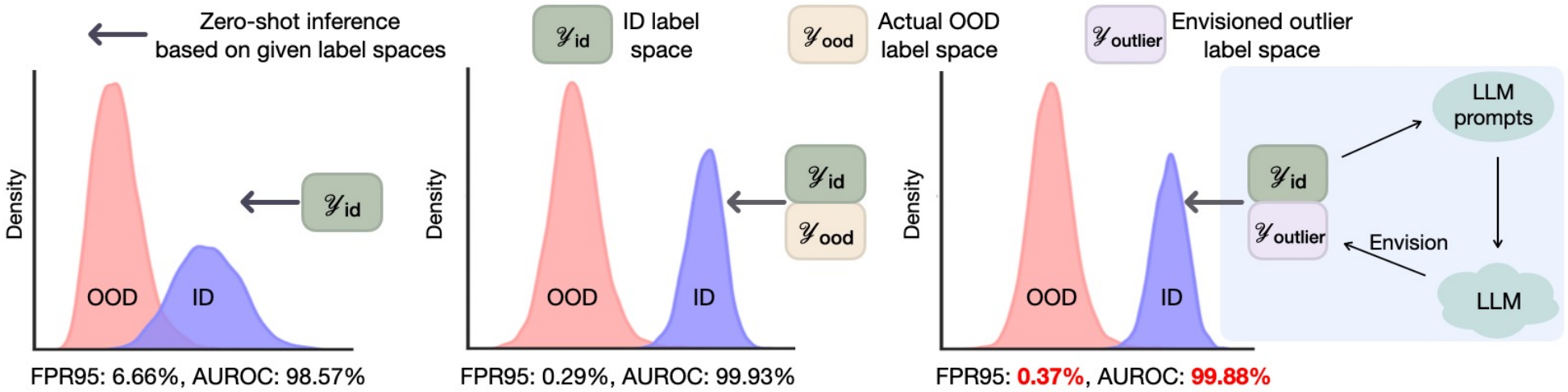
(a) **Closed-Set:** Using only closed-set ID classes

(b) **Ground Truth:** Incorporating with actual OOD class labels (unavailable)

➔ Building a text-based classifier with only closed-set labels largely restricts the inherent capability of CLIP

Research Problem

Is it possible to generate the potential outlier class labels for OOD detection without access to test-time data? 🤔



(a) **Closed-Set:** Using only closed-set ID classes

(b) **Ground Truth:** Incorporating with actual OOD classes (unavailable)

(c) **EOE (Ours):** Incorporating with envisioned outlier classes

Outline

- Background
- Method
- Experiments
- Summary

Motivation

LLMs possess a wealth of expert knowledge and strong reasoning capabilities

- LLMs know the visual features of lots of categories
- and then can utilize the visual features to envision outlier classes

Image generated by DALLE2

Technical Challenge

How to guide LLMs to generate the desired outlier class label? Since the OOD classes are unknown

Observations:

- MCM can easily distinguish visually distinct ID and OOD samples
- Indistinguishable ID/OOD samples are often visually similar

husky

wolf

visual similarity rule!

Framework

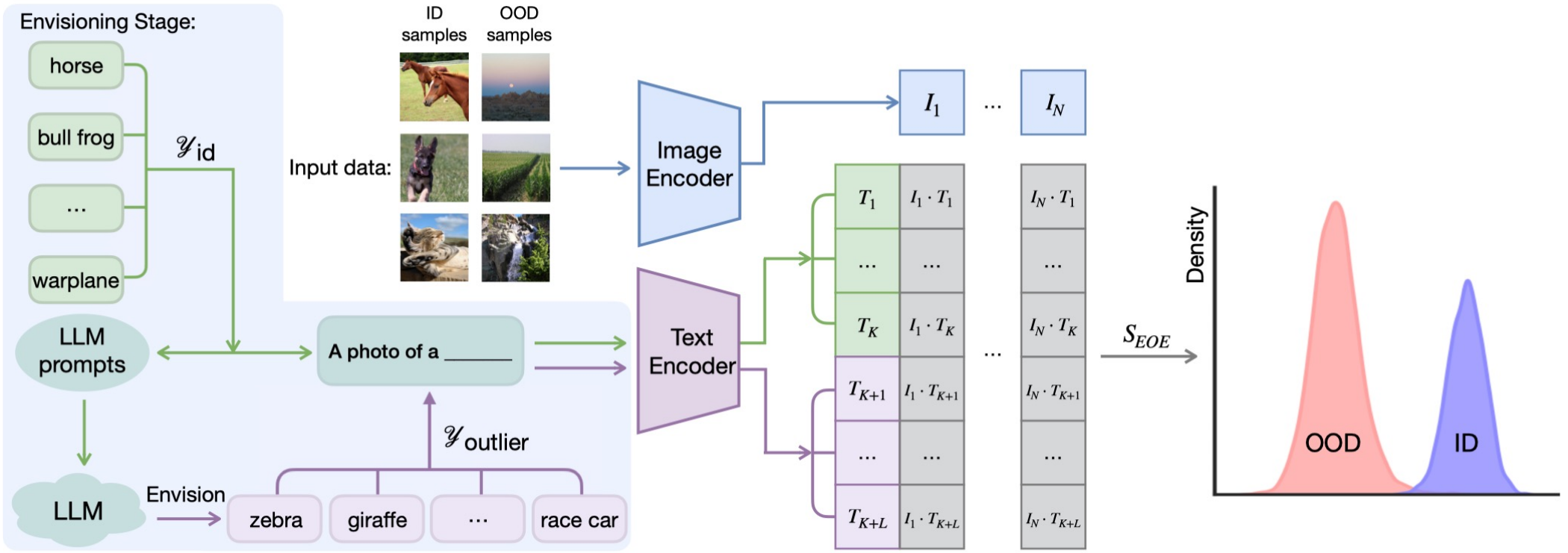


Figure 2: The framework of the proposed EOE. Given a set of ID class labels $\mathcal{Y}_{\text{id}}$, we first leverage the designed prompts to generate a set of outlier class labels, $\mathcal{Y}_{\text{outlier}}$, by using a LLM. Then, we input both the ID and generated OOD class labels into the text encoder for building the textual classifier. During the test stage, given an input image, we obtain the visual feature by the image encoder and calculate the similarities between the visual feature and the textual classifier. Finally, the OOD score is obtained by scaling the similarities with the proposed detection score function S_{EOE}.

Envision Outlier Exposure

We categorize OOD detection tasks into three types: far, near, and fine-grained OOD detection

Q: I have gathered images of K distinct categories: $\mathcal{Y}_{\text{id}}$. Summarize what broad categories these categories might fall into based on visual features. Now, I am looking to identify L classes that visually resemble these broad categories but have no direct relation to these broad categories. Please list these L categories for me.

Far OOD prompt

A: These L categories are:

Figure 3: LLM prompt for far OOD detection, consisting of both the contents of **Q** and **A**.

Q: Given the image category y_i, please suggest visually similar categories that are not directly related or belong to the same primary group as y_i. Provide suggestions that share visual characteristics but are from broader and different domains than y_i.

Near OOD prompt

A: There are l classes similar to y_i, and they are from broader and different domains than y_i:

Figure 4: LLM prompt for near OOD detection.

Q: I have a dataset containing K different species of *class-type*. I need a list of L distinct *class-type* species that are NOT present in my dataset, and ensure there are no repetitions in the list you provide. For context, the species in my dataset are: $\mathcal{Y}_{\text{id}}$.

Fine-grained OOD prompt

A: The other L *class-type* species not in the dataset are:

Figure 5: LLM prompt for fine-grained OOD Detection.

Implementation | the full algorithm

$$S_{\text{EOE}}(x; \mathcal{Y}_{\text{id}}, \mathcal{Y}_{\text{outlier}}, \mathcal{T}, \mathcal{I}) = \max_{i \in [1,K]} \frac{e^{s_i(x)}}{\sum_{j=1}^{K+L} e^{s_j(x)}} - \beta \cdot \max_{k \in (K,K+L]} \frac{e^{s_k(x)}}{\sum_{j=1}^{K+L} e^{s_j(x)}}$$

Algorithm 1 Zero-shot OOD detection with envisioned outlier class labels

Input: ID class labels $\mathcal{Y}_{\text{id}}$, test sample x, text encoder $\mathcal{T}$, image encoder $\mathcal{I}$, LLM, β, threshold λ;

Envisioning Stage:

Given $\mathcal{Y}_{\text{id}}$, $\mathcal{Y}_{\text{outlier}} = \text{LLM}(\text{prompt}(\mathcal{Y}_{\text{id}}))$;

Testing Stage:

$K = \text{len}(\mathcal{Y}_{\text{id}}), L = \text{len}(\mathcal{Y}_{\text{outlier}})$;

// Compute label-wise matching score

$\{s_i(x) = \frac{\mathcal{I}(x) \cdot \mathcal{T}(t_i)}{\|\mathcal{I}(x)\| \cdot \|\mathcal{T}(t_i)\|}\}_{i=1}^{K+L}; \quad t_i \in \mathcal{Y}_{\text{id}} \cup \mathcal{Y}_{\text{outlier}}$;

// Compute OOD detection score

$S_{\text{EOE}}(x) = \max_{i \in [1,K]} \frac{e^{s_i(x)}}{\sum_{j=1}^{K+L} e^{s_j(x)}} - \beta \max_{k \in (K,K+L]} \frac{e^{s_k(x)}}{\sum_{j=1}^{K+L} e^{s_j(x)}}$;

Output: OOD detection decision $\mathbf{1}\{S_{\text{EOE}} \geq \lambda\}$.

Outline

- Background
- Method
- Experiments
- Summary

Experiments | main results

Table 2: Zero-shot **far** OOD detection results for ImageNet-1K as ID dataset. The **black bold** indicates the best performance. The gray indicates that the comparative methods require training or an additional massive auxiliary dataset. Energy (FT) requires fine-tuning, while Energy is post-hoc.

Method	**OOD Dataset**								**Average**	
	iNaturalist		SUN		Places		Texture			
	FPR95↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑
MOS (BiT)	9.28	98.15	40.63	92.01	49.54	89.06	60.43	81.23	39.97	90.11
Fort et al.	15.07	96.64	54.12	86.37	57.99	85.24	53.32	84.77	45.12	88.25
Energy(FT)	21.59	95.99	34.28	93.15	36.64	91.82	51.18	88.09	35.92	92.26
MSP	40.89	88.63	65.81	81.24	67.90	80.14	64.96	78.16	59.89	82.04
CLIPN	19.13	96.20	25.69	94.18	32.14	92.26	44.60	88.93	30.39	92.89
Energy	81.08	85.09	79.02	84.24	75.08	83.38	93.65	65.56	82.21	79.57
MaxLogit	61.66	89.31	64.39	87.43	63.67	85.95	86.61	71.68	69.08	83.59
MCM	30.92	94.61	37.59	92.57	44.71	89.77	57.85	86.11	42.77	90.77
EOE (Ours)	12.29	97.52	20.40	95.73	30.16	92.95	57.53	85.64	**30.09**	**92.96**
Ground Truth	-	-	-	-	13.24	96.96	24.29	95.04	-	-

Experiments | main results

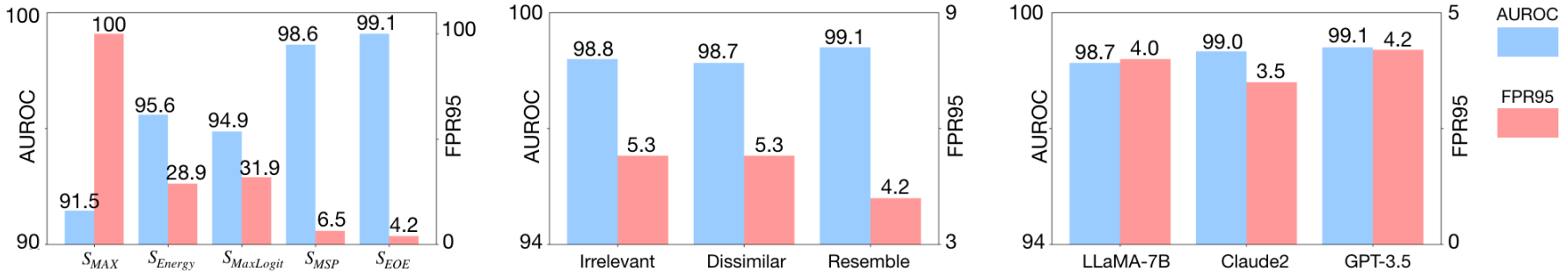
Table 3: Zero-shot **near** OOD detection results. The **bold** indicates the best performance on each dataset, and the gray indicates methods requiring an additional massive auxiliary dataset.

Method	ID / OOD	ImageNet-10 / ImageNet-20		ImageNet-20 / ImageNet-10		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CLIPN		7.80	98.07	13.67	97.47	10.74	97.77
Energy		10.30	97.94	16.40	97.37	13.35	97.66
MaxLogit		9.70	98.09	14.00	97.81	11.85	97.95
MCM		5.00	98.71	17.40	97.87	11.20	98.29
EOE (Ours)		4.20	99.09	13.93	98.10	**9.07**	**98.59**
Ground Truth		0.20	99.80	0.20	99.93	0.20	99.87

Table 4: Zero-shot **fine-grained** OOD detection results. The **bold** indicates the best performance on each dataset, and the gray indicates methods requiring an additional massive auxiliary dataset.

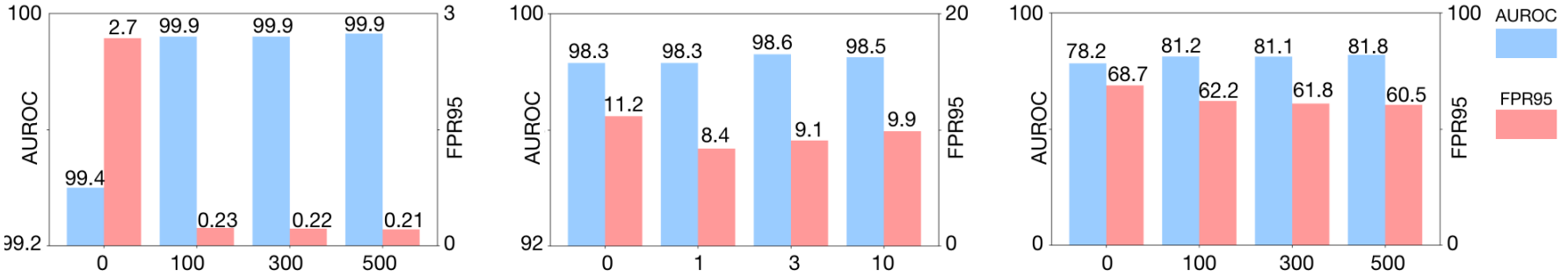
Method	ID / OOD	CUB-100 / CUB-100		Stanford-Cars-98 / Stanford-Cars-98		Food-50 / Food-51		Oxford-Pet-18 / Oxford-Pet-19		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CLIPN		73.54	74.65	53.33	82.25	43.33	88.89	53.90	86.92	56.05	83.18
Energy		76.13	72.11	73.78	73.82	44.95	89.97	68.51	88.34	65.84	81.06
MaxLogit		76.89	73.00	72.18	74.80	41.73	90.79	65.66	88.49	64.11	81.77
MCM		83.58	67.51	83.99	68.71	43.38	91.75	63.92	84.88	68.72	78.21
EOE (Ours)		74.74	73.41	76.83	71.60	37.95	91.96	52.55	90.33	**60.52**	**81.82**
Ground Truth		61.23	81.42	58.31	83.71	11.34	97.79	29.17	95.58	40.01	89.63

Experiments | ablation study



(a) Score Function (b) LLM Prompts (c) Various LLMs

Figure 6: Ablation study on (a) score function, (b) LLM prompts, and (c) various LLMs. ID dataset: ImageNet-10; OOD dataset: ImageNet-20.



(a) Far OOD detection (b) Near OOD detection (c) Fine-grained OOD detection

Figure 7: Evaluation on the number of outlier class labels. When the number of outlier class labels is zero, the method reduces to the baseline MCM.

Understanding | without hitting the GT OOD

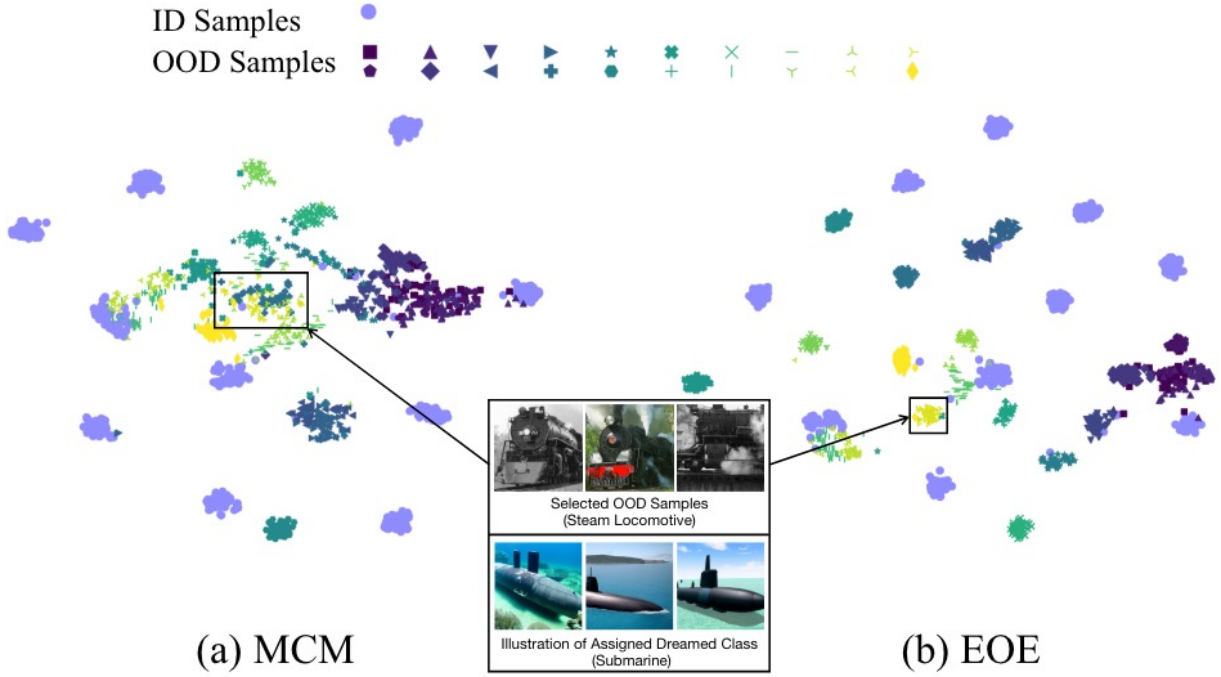


Figure 8: T-SNE visualizations obtained by the classifier output. ID set: ImageNet-10; OOD set: ImageNet-20. We use distinct colors to represent different OOD classes. The illustrated envisioned OOD name is the class assigned with the corresponding cluster, and its examples are generated by Stable Diffusion (Rombach et al., 2022).

Outline

- Background
- Method
- Experiments
- Summary

Summary

Main contributions

- We propose a new perspective that leverages expert knowledge from LLM to envision potential outlier class labels, facilitating OOD detection
- We propose EOE, providing LLM prompts based on the visual similarity rule to envision potential outlier class labels, and design a score function to effectively distinguish between ID samples and OOD samples
- Extensive experiments show that EOE achieves improvements of 2.47%, 2.13%, 3.59%, and 12.68% on the far OOD, near OOD, fine-grained OOD, and ImageNet-1K in terms of FPR95

Thanks for your listening!

Chentao Cao csctcao@comp.hkbu.edu.hk