

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

Wenjun Miao^{1,3}, Guansong Pang^{2*}, Zihan Wang^{1,3}, Jin Zheng^{1*}, Xiao Bai^{1,3}

¹School of Computer Science and Engineering, Beihang University

²School of Computing and Information Systems, Singapore Management University

³State Key Laboratory of Software Development Environment, Jiangxi Research Institute, Beihang University

{miaowenjun, wangzihan1118, jinzheng, baixiao}@buaa.edu.cn, gspang@smu.edu.sg

Abstract

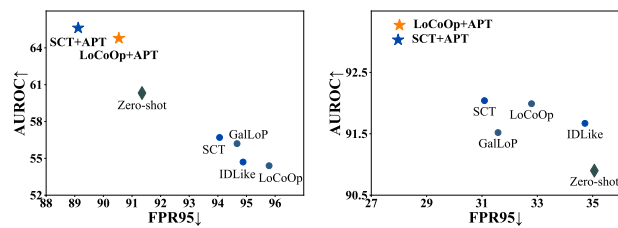
Recent advancements in CLIP-based out-of-distribution (OOD) detection have shown promising results via regularization on prompt tuning, leveraging background features extracted from a few in-distribution (ID) samples as proxies for OOD features. However, these methods suffer from an inherent limitation: a lack of diversity in the extracted OOD features from the few-shot ID data. To address this issue, we propose to leverage external datasets as auxiliary outlier data (i.e., pseudo OOD samples) to extract rich, diverse OOD features, with the features from not only background regions but also foreground object regions, thereby supporting more discriminative prompt tuning for OOD detection. We further introduce **Auxiliary Prompt Tuning (APT)**, a novel framework that can be used as a plug-in module to enable existing prompt tuning-based methods to utilize the auxiliary data for more accurate OOD detection. There are two key challenges of utilizing those auxiliary data in prompt tuning, including I) foreground-background decomposition of unlabeled auxiliary data with diverse outlying objects and II) optimization of foreground OOD features. APT tackles challenge I with an adaptive logit-based Kullback–Leibler divergence method and challenge II by constructing foreground-background pairs for each foreground region to enable effective exploitation of foreground OOD features. Extensive experiments on standard and hard OOD benchmarks show that APT achieves state-of-the-art performance, obtaining significant improvements in challenging scenarios, e.g., hard OOD and 1-shot detection.

1. Introduction

Deep neural networks (DNNs) are widely known to be overconfident on out-of-distribution (OOD) data [9, 16, 20, 28],

*Corresponding authors: G. Pang and J. Zheng.

The code is available at <https://github.com/mala-lab/APT>.



(a) Hard OOD benchmark.

(b) Standard OOD benchmarks.

Figure 1. OOD detection performance in 1-shot training with ImageNet-1K [4] as ID data. (a) presents evaluation results of existing state-of-the-art (SotA) few-shot methods on the large-scale hard OOD test set ImageNet-1k-OOD [32] involving categories semantically similar to ID data, where they even underperform zero-shot methods. (b) shows average performance across four standard benchmarks following a popular protocol [11, 36]. These SotA methods suffer significant performance degradation on hard OOD scenarios, whereas our proposed APT can substantially enhance these detectors in both standard and hard scenarios.

often misclassifying OOD samples from unknown classes as one of the known classes. This phenomenon poses critical risks in safety-sensitive applications such as autonomous driving [12] and medical diagnosis [14]. One notorious problem in OOD detection is the lack of ground-truth information on test-time OOD samples, as they can be drawn from any unknown distribution [11, 19]. Recent works [2, 22, 36] tackle this problem by leveraging background features extracted from a few in-distribution (ID) samples as proxies for OOD features to facilitate prompt tuning of vision-language models (VLMs), such as CLIP [26], for the OOD detection task. Despite showing good performance on existing benchmarks, these methods suffer from the lack of diversity in the extracted OOD features from those few-shot ID data, leading to significant performance degradation on challenging scenarios, e.g., detection of hard OODs, where they can even underperform zero-shot baselines (see Fig. 1).

To address this issue, we propose to leverage samples

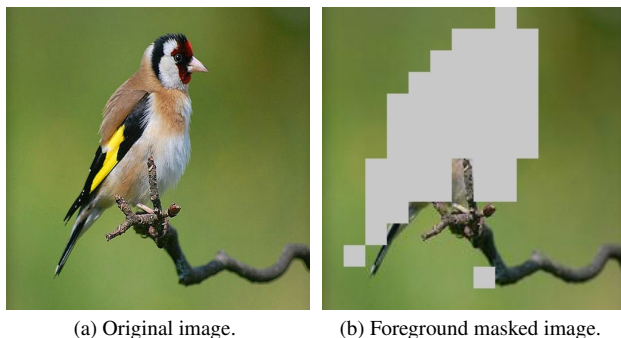


Figure 2. Foreground-background region decomposition via our proposed adaptive logit-based KL divergence. The masked regions have higher semantic similarity to the global/whole image.

from external datasets as auxiliary outlier data (*i.e.*, samples of classes that do not overlap with ID and ground-truth OOD classes, a.k.a. pseudo OOD samples) to obtain diverse OOD features from abundant background and foreground features. A similar approach, known as outlier exposure (OE) [9], is widely used in traditional OOD detection [16, 18, 30]. But to our knowledge, how to exploit such auxiliary data in prompt tuning of VLMs for OOD detection has not been explored in the literature due to the following two key challenges. One is the foreground-background region decomposition of large-scale auxiliary data of diverse unknown foreground objects. Existing methods [5, 13, 22, 36] rely on ground-truth labels in the ID data to decompose the foreground-background regions, which is infeasible for unlabeled auxiliary data with unknown outlying objects. The other lies in the optimization of foreground OOD features. Although simple fitting of those foreground OOD features in the optimization can improve the hard OOD performance, it can overfit these pseudo foreground OOD objects, impairing the OOD capability on large scene datasets (*e.g.*, SUN [34] and Places [38]) where OOD samples do not have clear foreground objects (see Table 4a).

In this work, we tackle these two challenges with a novel framework, **Auxiliary Prompt Tuning (APT)**. APT can serve as a general plug-in module for enabling existing prompt tuning methods to obtain diverse OOD features from the auxiliary data and then leverage them for enhanced OOD detection models. APT consists of two novel components: 1) foreground-background region decomposition via adaptive logit-based Kullback–Leibler (KL) divergence (namely **ALK**) and 2) foreground-background pairing for superior foreground OOD feature regularization (namely **PairReg**). In ALK, we measure the KL divergence between the classification logits of local image regions and their corresponding global image to quantify their semantic similarity for the decomposition. The key insight is that, due to the superior local region’s visual and textual alignment in VLMs like CLIP [22, 39], the foreground re-

gions often have high semantic similarity to the global image, exhibiting classification logits more closely aligned with those of the corresponding global image, compared to the background regions that have low semantic similarity. As a result, regardless of the diverse unknown objects in the auxiliary data, regions with higher logit-based semantic similarity can be identified as foreground, while regions with lower semantic similarity are considered as background, enabling label-free decomposition of foreground-background regions (see Fig. 2). Subsequently, PairReg constructs foreground-background pairs for each foreground region for joint regularization, enabling knowledge transfer from background to foreground representations. This allows effective foreground OOD feature optimization while at the same time balancing foreground-background OOD features, rather than overfitting the foreground features. Our contributions are summarized as:

- We reveal the problem of lacking diverse OOD features in current prompt tuning-based OOD detectors, which can lead to significant performance degradation on challenging OOD detection scenarios, *e.g.*, hard OOD tasks.
- We then propose to leverage auxiliary data to incorporate diverse OOD features into prompt tuning for OOD detection. To make it possible, we curate an auxiliary dataset for well-established OOD detection benchmarks.
- We further introduce a novel Auxiliary Prompt Tuning (APT) framework. It consists of two novel components and can be used as a plug-in module to enable existing prompt tuning-based methods to effectively extract diverse OOD features from auxiliary data and leverage them for more accurate OOD detection (see Fig. 1).
- Comprehensive results show that APT achieves SotA performance on standard and hard OOD benchmarks, with significant improvements in challenging scenarios, *e.g.*, reducing 4.94% FPR95 in 1-shot hard OOD tasks.

2. Related Work

OOD Detection in VLMs. Current VLM/CLIP-based OOD detection focuses on two primary paradigms: zero-shot methods and prompt tuning techniques. Zero-shot methods operate without requiring ID images during training or inference. MCM [21] is an early method that aligns visual features with textual concepts through softmax probability maximization. GL-MCM [23] extends MCM [21] by incorporating local region scoring to improve fine-grained detection. CLIPN [33] trains an auxiliary text encoder with large-scale external datasets, but it cannot be optimized for specific ID data and requires extensive training overhead. OLE ding2024zero explores the use of outlier label data through CLIP. In contrast, prompt tuning techniques use limited labeled ID data for optimization. LSN [24] and NegPrompt [15] employ negative prompts to capture ID

sample semantics, but they need carefully prompt design. Recently, LoCoOp [22] achieves significant success by using extracted background features as OOD features to perform OOD regularization. IDLike [2] randomly crops training ID data to obtain enhanced OOD features from augmented background features. SCT [36] further extends LoCoOp [22] by calibrating the extracted background features to mitigate inaccurate foreground-background decomposition. However, they exhibit limited effectiveness in challenging scenarios, *e.g.*, hard OOD scenarios, due to the lack of diversity for OOD features extracted from background representations under few-shot ID data. Therefore, we propose to leverage auxiliary data to extract diverse OOD features for enhanced prompt tuning.

OOD Detection with Auxiliary Data. A popular branch of OOD detection methods [1, 6–8, 10, 17, 25, 29] is to leverage auxiliary data (*i.e.*, pseudo OOD data, assuming that do not overlap with ID and test-time OOD samples) to enhance the discriminability between OOD and ID data. OE [9] pioneered this paradigm by enforcing uniform prediction on auxiliary data, demonstrating its potential for OOD detection. Subsequently, EnergyOE [16] maximizes the free energy of auxiliary samples, and HB [10] employs energy-based Hopfield boosting to refine OOD discrimination, achieving substantial performance improvement. Despite these advancements, to our knowledge, no work has been done on exploring such auxiliary data to enhance prompt tuning for OOD detection. In addition, multiple auxiliary datasets have emerged. UDG [35] introduces a semantically coherent OOD benchmark but is limited to small-scale ID datasets. DOE [37] utilizes the ImageNet-21K-P [27] as the auxiliary dataset for ImageNet-1K [4] but lacks a challenging OOD test set. PASCL [32] constructs ImageNet-Extra (auxiliary dataset) and ImageNet-1k-OOD (large-scale hard OOD test set) from ImageNet-21K [28] for ImageNet-1K-based ID data, but there are overlapping classes between ImageNet-Extra and the current CLIP-based OOD benchmarks. Therefore, we curate an auxiliary dataset derived from ImageNet-Extra that removes such overlapping to support the utilization of those auxiliary data in OOD detection with VLMs.

3. Problem Statement

Preliminaries. In VLM-based OOD detection, ID classes refer to the classes used in downstream classification tasks, distinct from classes used during pre-training. OOD classes can be any classes that are different from these ID classes. CLIP is commonly used as the instantiation of the VLM. Formally, let X_{in} be the input space of the ID data, $Y_{in} = \{1, 2, \dots, M\}$ be the ID label space, X_{out} and $Y_{out} = \{M + 1, M + 2, \dots\}$ denote input space and label space for OOD data, respectively, with no class overlap between ID

and OOD data ($Y_{in} \cap Y_{out} = \emptyset$), then a popular objective of OOD detection with VLMs is to train a set of prompts such that for any test data $x \in X_{in} \cup X_{out}$: if x drawn from X_{in} , it can be classified into correct ID class; and if x is drawn from X_{out} , it can be detected as OOD data. CLIP-based OOD detection often takes a few-shot setting, where only a limited number of ID images in each class (*e.g.*, 1 or 16 images) are used during training. It is normally assumed that genuine OOD data X_{out} is not available during training since OOD samples are unknown instances. On the other hand, auxiliary data X_{aux} that is not X_{out} but drawn from a different distribution other than X_{in} are often available. These auxiliary samples X_{aux} can be used as pseudo OOD samples to support the training of OOD detectors.

Vanilla Prompt Tuning. Formally, given an ID image x_{in} and its corresponding label y_{in} , we can obtain the global visual feature $\mathbf{f}_{in} = f(x_{in})$ with the CLIP’s visual encoder $f(\cdot)$. The textual prompt vectors for a class can be denoted as $\mathbf{t}_m = \{\omega_1, \omega_2, \dots, \omega_N, \mathbf{c}_m\}$, where \mathbf{c}_m represents the class embedding of a ID class token and $\omega = \{\omega_n |_{n=1}^N\}$ corresponds to N learnable context vectors. Each learnable context vector has the same dimension as the class token embedding. The text encoder $g(\cdot)$ maps a prompt vector \mathbf{t}_m to a textual feature vector $\mathbf{g}_m = g(\mathbf{t}_m)$ for the ID classes. The CLIP-based prediction probability can be computed as:

$$p(y = m | x_{in}) = \frac{\exp(\text{sim}(\mathbf{f}_{in}, \mathbf{g}_m) / \tau)}{\sum_{m'=1}^M \exp(\text{sim}(\mathbf{f}_{in}, \mathbf{g}_{m'}) / \tau)}, \quad (1)$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity function and τ is a temperature parameter.

Prompt Tuning with Background Features for OOD Detection. Compared with vanilla prompt tuning methods like CoOp [40], LoCoOp [22] is an advanced prompt tuning framework that performs regularization on extracted local/background features from ID-irrelevant background regions for OOD detection. Specifically, given ID training data $\mathcal{D}_{in} = (X_{in}, Y_{in})$, LoCoOp [22] minimizes:

$$\mathcal{L}_{\text{LoCoOp}} = \mathbb{E}_{(x,y) \sim \mathcal{D}_{in}} [\ell_{\text{CE}}(p(y|x), y) + \lambda \ell_{\text{OOD}}(p(\mathbf{R}(x)))], \quad (2)$$

where $\ell_{\text{CE}}(\cdot)$ is a cross-entropy loss, $\ell_{\text{OOD}}(\cdot)$ is the negative entropy of the given probability vector, and λ is a hyperparameter. $\mathbf{R}(\cdot)$ is a ranking-based approach to extract ID-irrelevant background regions in LoCoOp [22].

Building upon LoCoOp [22], SCT [36] introduces two calibration factors to refine the regularization of background features, mitigating inaccurate foreground-background decomposition. The objective in SCT [36] is defined as:

$$\mathcal{L}_{\text{SCT}} = \mathbb{E}_{(x,y) \sim \mathcal{D}_m} [\ell_{\text{CE}}(p(y|x), y) * (1 - p(y|x)) + \lambda \ell_{\text{OOD}}(p(\mathbf{R}(x))) * p(y|x)], \quad (3)$$

where SCT [36] retains the same ranking-based approach as LoCoOp to extract background regions. However, this

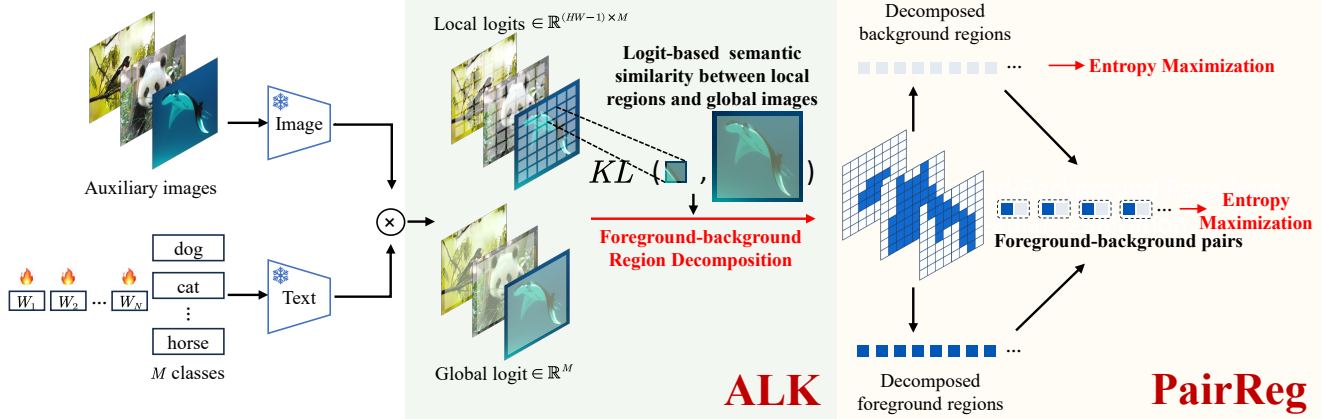


Figure 3. Overview of the proposed APT. APT first projects both local and global features from the CLIP visual feature map into the textual embedding space to obtain their classification logits. Then, APT decomposes foreground-background regions based on their semantic similarities to the global features, which are calculated by the KL divergence between local logits and the corresponding global logits. APT subsequently constructs foreground-background pairs for each foreground region and performs joint regularization on the background regions and paired instances via entropy maximization.

approach requires ground-truth labels of the input images, which can be too costly for large-scale auxiliary data with diverse unknown foreground objects. We propose an adaptive logit-based KL divergence method for foreground-background decomposition to eliminate this requirement.

Test-time OOD Detection. During inference, following prior studies [13, 36], we use GL-MCM [23] to obtain OOD scores by default since it has demonstrated superior performance over other methods, such as MCM [21]. It combines the maximum softmax probability derived from global and local image features. Formally, the scoring is defined as:

$$S_{\text{GL-MCM}}(x) = \max_m \frac{\exp(\text{sim}(\mathbf{f}, \mathbf{g}_m) / \tau)}{\sum_{m'=1}^M \exp(\text{sim}(\mathbf{f}, \mathbf{g}_{m'}) / \tau)} + \max_{m,i} \frac{\exp(\text{sim}(\mathbf{f}^i, \mathbf{g}_m) / \tau)}{\sum_{m'=1}^M \exp(\text{sim}(\mathbf{f}^i, \mathbf{g}_{m'}) / \tau)}, \quad (4)$$

where \mathbf{f} denotes the global image features of a test sample x , \mathbf{f}^i denotes the local image features of the i -th region of x , and τ is a temperature parameter fixed to one.

4. The Proposed Approach APT

We introduce the APT approach to effectively utilize the auxiliary data for empowering prompt tuning methods in VLM-based OOD detection. As illustrated in Fig. 3, APT consists of two novel components: 1) ALK: foreground-background region decomposition via adaptive logit-based KL divergence and 2) PairReg: foreground-background pairing for foreground OOD feature regularization.

4.1. Adaptive Logit-based KL Divergence

Existing methods rely on ground-truth labels of input images to decompose foreground-background regions, which

is infeasible for large-scale unlabeled auxiliary data. To address this issue, we propose ALK, which adaptively quantifies the semantic similarity between local image regions and their corresponding global image by measuring the KL divergence between their classification logits. It then uses this semantic similarity to decompose foreground-background regions. This is because the foreground regions of an image typically have higher semantic similarity to the corresponding global image than its background regions. More importantly, the regions with higher semantic similarity to the global image exhibit classification logits more closely aligned with those of the corresponding global image, primarily due to the CLIP’s superior local visual and textual feature alignment [39].

To be specific, for an input image $x \in X_{aux}$, we extract local feature \mathbf{f}^i from the CLIP visual feature map for each of its regions x_i . Let a set of all region indices be $\mathbf{I} = \{0, 1, 2, \dots, H \times W - 1\}$, where H and W denote the height and width of the feature map of x , for each region x_i , we can then adaptively calculate local logits $p(x_i)$ by computing the similarity between the local features \mathbf{f}^i and the text features of the ID classes [23, 39] as follows:

$$p(x_i) = \frac{\exp(\text{sim}(\mathbf{f}^i, \mathbf{g}_m) / \tau)}{\sum_{m'=1}^M \exp(\text{sim}(\mathbf{f}^i, \mathbf{g}_{m'}) / \tau)}. \quad (5)$$

The decomposition of foreground and background regions for each x from the auxiliary data can be formulated as:

$$J_x^{back} = \{i \in \mathbf{I} : \text{KL}(S(p(x)) \parallel S(p(x_i))) > \epsilon\}, \\ J_x^{fore} = \{i \in \mathbf{I} : \text{KL}(S(p(x)) \parallel S(p(x_i))) \leq \epsilon\}, \quad (6)$$

Where KL is the KL divergence function, $S(\cdot)$ is a Softmax operation, ϵ is a threshold hyperparameter – a percentage to

measure the ratio of foreground regions to all image regions – to decompose foreground-background regions, and $p(x)$ is the resulting global logits of x as in Eq. 1. Note that the region set $J_x = J_x^{back} \cup J_x^{fore}$ of x is continuously updated relative to $p(x)$ and $p(x_i)$ during training.

4.2. Foreground-background Pair Regularization

While directly performing regularization on foreground regions J_{fore} can improve hard OOD detection performance, it can overfit the seen OOD features, impairing OOD detection capability on OOD samples that lack clear foreground objects. To address this issue, we propose a foreground-background pairing-based regularization method, PairReg. Specifically, it constructs local foreground-background pairs for a joint regularization, enabling knowledge transfer from background to foreground representations during auxiliary feature optimization. By doing so, it enhances the detection of OOD samples with foreground objects, *e.g.*, hard OOD samples, while at the same time preserving its discriminative OOD detection capability on background/scene-focused OOD samples.

Background-foreground Pairing. Formally, given an input image $x \in X_{aux}$, for each i -th foreground region $x_i \in J_x^{fore}$ with classification logits $p(x_i)$, we randomly sample a j -th background region from the image \hat{x} , having $\hat{x}_j \in J_{\hat{x}}^{back}$ with classification logits $p(\hat{x}_j)$ (x and \hat{x} come from the same training batch). After that, the foreground-background pair is constructed via element-wise addition of their classification logits:

$$p_{pair}(x_i) = p(x_i) + p(\hat{x}_j). \quad (7)$$

We construct the foreground-background pair for each foreground region and regularize these pairs instead of directly regularizing the foreground regions.

Joint Regularization with the Pairs. For background regions J_{back} , we adopt the same regularization strategy as background features from ID data, as they are both background features, which helps consistently improve detection performance on both standard and hard OOD scenarios. We also perform regularization with the obtained pairs to regularize the foreground OOD feature learning, which further improves the OOD detection in challenging scenarios. The overall APT loss is then defined as follows:

$$\mathcal{L}_{APT} = \mathbb{E}_{x \sim X_{aux}} [\alpha \ell_{OOD}(p(J_x^{back})) + \beta \ell_{OOD}(p_{pair}(J_x^{fore}))], \quad (8)$$

where $\ell_{OOD}(\cdot)$ denotes the negative entropy of the given probability vector [22], and α and β are hyperparameters.

Overall Objective. The final training objective as follows:

$$\mathcal{L} = \mathcal{L}_{ID} + \mathcal{L}_{APT}, \quad (9)$$

Method	ID:ImageNet-10 OOD:ImageNet-20		ID:ImageNet-20 OOD:ImageNet-10		ID:ImageNet-1k OOD:ImageNet-1k-OOD	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
<i>Zero-shot methods</i>						
MCM [21]	5.00	98.71	12.91	98.09	93.90	53.82
GL-MCM [23]	10.10	98.04	9.00	98.62	91.35	60.32
<i>Prompt tuning-based methods</i>						
1-shot						
LoCoOp [22]	16.00	96.59	18.50	95.78	95.79	54.40
IDLike [2]	14.36	96.92	16.85	96.27	94.88	54.70
LSN [24]	20.26	96.31	21.74	94.95	96.53	53.14
GalLoP [13]	12.77	97.52	14.30	96.29	94.67	56.20
SCT [36]	13.50	97.03	15.00	96.62	94.06	56.70
Ours	7.80	97.95	9.60	98.18	89.12	65.62
16-shot						
LoCoOp [22]	8.60	97.92	10.60	98.56	93.86	57.29
IDLike [2]	9.43	97.32	10.28	97.96	93.07	57.59
LSN [24]	14.93	96.77	15.86	97.54	94.42	54.29
GalLoP [13]	7.65	97.86	9.22	98.12	93.03	56.80
SCT [36]	6.80	98.21	8.10	98.80	93.43	57.72
Ours	2.97	99.27	4.99	98.95	87.64	65.86

Table 1. Comparison results on hard OOD benchmarks.

where \mathcal{L}_{APT} (defined in Eq. 8) denotes our proposed regularization loss applied to the auxiliary data, and \mathcal{L}_{ID} denotes a regularization loss on ID data. Our proposed APT provides a general framework through the objective in Eq. 9, enabling current prompt tuning methods based solely on ID data to be easily plugged in via the use of their loss to implement \mathcal{L}_{ID} , *e.g.*, using Eq. 2 and Eq. 3 to instantiate \mathcal{L}_{ID} to derive APT-enabled LoCoOp and SCT respectively.

5. Experiments

5.1. Datasets

ID Datasets. We adopt ImageNet-1K [4] as an ID dataset following [22, 36]. For a more comprehensive evaluation, we also use ImageNet-10 [23] and ImageNet-20 [23], both of which are subsets of ImageNet-1k, as ID datasets, following hard OOD detection protocols [23, 36].

OOD Datasets. Following [23, 36], we use ImageNet-20 as the OOD test set for ImageNet-10, and use ImageNet-10 as the OOD test set for ImageNet-20. For ImageNet-1k, we follow the same protocols as [11, 22, 36] by using standard OOD benchmarks that include subsets of iNaturalist [31], SUN [34], Places [38], and Textures [3]. Notably, these benchmarks for ImageNet-1k lack hard OOD samples that have semantically similar foreground objects to ID objects. The ImageNet-10/20 benchmarks are hard OOD samples to each other, but they are limited in scale. Therefore, we propose to use ImageNet-1k-OOD [32] as an OOD test set to evaluate the large-scale hard OOD detection capability against the ID data in ImageNet-1k. ImageNet-1k-OOD contains 50,000 images for 1,000 classes from ImageNet-21k [28], having no overlap but semantically similar to the ID classes in ImageNet-1k.

Auxiliary Datasets. Current CLIP-based OOD detection lacks a dedicated auxiliary dataset for training. To

Method	iNaturalist		SUN		Places365		Textures		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
<i>Zero-shot methods</i>										
MCM [21]	31.86	94.17	37.28	92.55	42.94	90.09	58.37	85.83	42.61	90.66
GL-MCM [23]	15.16	96.71	29.16	93.41	37.07	90.37	58.85	83.11	35.06	90.90
<i>prompt tuning-based methods</i>										
1-shot										
LoCoOp [22]	23.53	94.89	24.15	94.55	32.84	91.51	50.67	87.01	32.79	91.99
IDLike [2]	12.07	97.65	40.55	91.07	47.94	88.31	38.34	89.67	34.72	91.67
LSN [24]	59.28	87.20	40.15	91.47	46.11	88.74	60.34	83.92	51.47	87.84
GalLoP [13]	19.63	95.21	27.66	92.82	33.49	90.61	45.52	87.44	31.58	91.52
SCT [36]	19.16	95.70	23.52	94.58	32.81	91.23	48.87	86.66	31.09	92.04
Ours	13.94	96.89	18.93	95.80	28.04	92.96	50.96	86.49	27.97	93.03
16-shot										
LoCoOp [22]	17.58	96.30	22.82	95.20	32.21	92.03	45.27	88.86	29.47	93.10
IDLike [2]	9.71	98.05	38.93	90.54	47.06	88.06	32.82	91.89	32.12	92.14
LSN [24]	36.17	92.66	34.27	93.53	41.47	90.52	46.43	89.38	39.58	91.53
GalLoP [13]	13.70	97.10	24.90	94.00	32.50	91.30	38.40	90.40	27.30	93.20
SCT [36]	13.94	95.86	20.55	95.33	29.86	92.24	41.51	89.06	26.47	93.37
Ours	9.70	97.79	20.12	95.52	28.54	92.84	45.78	88.43	26.03	93.64

Table 2. Comparison results on standard OOD benchmarks with ImageNet-1k as ID dataset.

bridge this gap, we introduce an auxiliary dataset, namely ImageNet-Extra 2.0, based on a curated extension of ImageNet-Extra [32], which contains 500 non-overlapping classes from ImageNet-21k [28] relative to ImageNet-1k and ImageNet-1k-OOD [32]. We also manually remove 23 classes from ImageNet-Extra [32] to eliminate class overlaps with the aforementioned standard OOD benchmarks. ImageNet-Extra 2.0 is used for auxiliary OOD feature learning for both standard and hard OOD detection settings.

5.2. Experimental Setup

Implementation Details. Following [22, 36], we use CLIP ViT-B/16 as the backbone and train for 25 epochs. The batch size is set to 96, with 32 ID samples and 64 auxiliary samples. For the threshold ϵ in ALK, we use a percentage $\epsilon = 0.4$ (*i.e.*, 40% is the foreground region). In the APT loss, we use $\alpha = 0.6$ and $\beta = 0.2$ by default. For other hyperparameters in the ID loss (\mathcal{L}_{ID} in Eq. 9), we use the same implementation as the original papers. Following this, we implement the plug-in of APT in three SotA prompt tuning-based methods, LoCoOp [22], IDLike [2], and SCT [36], denoted by LoCoOp+APT, IDLike+APT, and SCT+APT. For the few-shot training, following the prior studies [22, 36], we report the results of 1-shot and 16-shot ID data, respectively. The average results over three runs are reported for comparison.

Evaluation Metrics. Following [2, 36], we use the following two common metrics for OOD detection: 1) the false positive rate of OOD images when the true positive rate of ID images is at 95% (FPR95), and 2) the area under the receiver operating characteristic curve (AUROC).

Comparison Baselines. We compare APT with several SotA methods from two categories, including post-

hoc methods MCM [21] and GL-MCM [23], and prompt tuning-based methods LoCoOp [22], IDLike [2], LSN [24], GalLoP [13] and SCT [36].

5.3. Main Results

Hard OOD Detection Performance. In Table 1, we present the comparison of our method (SCT+APT) with SotA OOD detectors on three hard OOD detection benchmarks: ImageNet-10, ImageNet-20, and ImageNet-1K. It can be observed that existing prompt tuning-based methods are ineffective on these benchmarks, suffering significant performance degradation compared to zero-shot baselines. This can be attributed to the lack of diversity in the extracted OOD features from the few-shot ID data. In contrast, our method achieves SotA performance across both evaluation metrics in all three hard OOD benchmarks, outperforming the competing methods by a large margin, especially in FPR95. As the only method to consistently outperform zero-shot baselines by substantial margins, APT successfully leverages the auxiliary data to overcome the OOD feature diversity issue inherent to prompt tuning methods in hard OOD scenarios. Notably, our method shows particularly remarkable gains in the 1-shot setting, where the diversity issue is more severe.

Standard OOD Detection Performance. In Table 2, we present the comparison results on the four standard OOD benchmarks with ImageNet-1k as ID data. Our method (SCT+APT) achieves consistently better overall performance in both FPR95 and AUROC metrics. Similar to the results in Table 1, our method also achieves particularly larger improvement in the 1-shot setting. This showcases the effectiveness of APT in leveraging auxiliary data to enhance the OOD detection of VLMs under varying scenarios.

Method	iNaturalist		SUN		Places365		Textures		Average		Hard OOD Data	
	Standard OOD Data		Standard OOD Data		Standard OOD Data		Standard OOD Data		Standard OOD Data		ImageNet-1k-OOD	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
1-shot												
LoCoOp [22]	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 [36]	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 [22]	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 [36]	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

Table 3. OOD detection performance results of plugging APT in three SotA prompt tuning methods, with ImageNet-1k as ID dataset.

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

Table 4. Ablation study results on ImageNet-1K. Standard OODs denote the average results on four standard OOD benchmarks.

Enabling Existing SotA Methods. Table 3 evaluates the effectiveness of our APT when plugging in three SotA prompt tuning methods (LoCoOp, IDLike, and SCT) for utilizing auxiliary data. The results show that APT can consistently enhance these three SotA OOD detectors across the standard and hard OOD scenarios, highlighting its universal effectiveness in leveraging auxiliary data for distinguishing OOD samples with different prompt tuning methods. This consistent improvement also justifies that one main limitation of current SotA is the lack of diverse OOD features, and APT is an effective plug-in for tackling this issue. Note that APT works less effectively on the OOD test set Texture due to its texture images having a huge semantic difference with natural images, making it difficult to learn this OOD information from the natural images in the auxiliary dataset ImageNet-Extra.

5.4. Discussion

Ablation Study. The effectiveness of regularization applied to background regions (*Back*) and foreground-background pairs (*Pair*) is presented in Table 4, using SCT [36] as the baseline. The results show that 1) *Back*: regularizing diverse background features extracted from auxiliary data

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

Table 5. OE methods on ImageNet-1K under 1-shot setting. Standard OODs present the average results on four standard OODs.

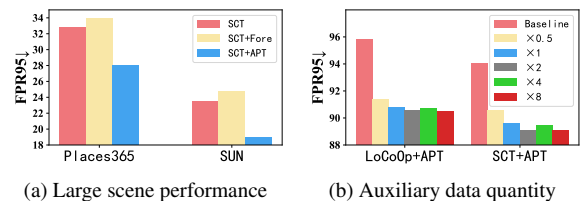


Figure 4. Performance on ImageNet-1K under 1-shot setting.

can consistently improve OOD detection performance on both standard and hard OOD scenarios, 2) *Pair*: performing regularization on the foreground-background pairs significantly improves hard OOD detection performance with comparable performance in standard OOD detection, 3) combining these two regularization strategies contributes to the overall superior performance of the full model of APT.

APT vs. Existing Outlier Exposure Methods. Table 5 presents the comparison of our APT with two SotA outlier exposure methods, OE [9] and EnergyOE [16], for utilizing the auxiliary data within SCT. The results show that existing SotA outlier exposure methods cannot be directly applied to boost CLIP-based OOD detection. In contrast, our APT achieves significant improvement.

Performance on Scene-based OODs. Fig. 4a reveals the ineffectiveness of directly applying regularization on foreground representations (SCT+Fore) on two OOD benchmarks Places365 and SUN where OOD samples do not have

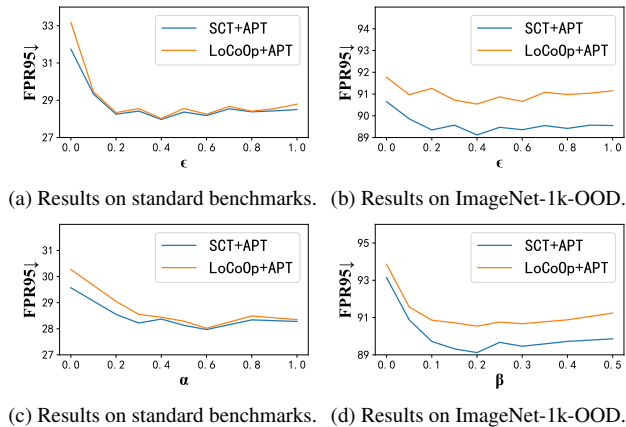


Figure 5. Hyperparameter analysis under 1-shot setting.

clear foreground objects. In contrast, enforcing regularization on our proposed foreground-background pairs can effectively address this issue, demonstrating that transferring knowledge from background features to foreground features during foreground OOD representation optimization helps overcome overfitting to the foreground OOD features.

Impact of Auxiliary Data Quantity. Fig. 4b presents the impact of varying auxiliary data quantities on OOD capability. We evaluate different auxiliary data batch sizes (Baseline, $\times 0.5$, $\times 1$, $\times 2$, $\times 4$, and $\times 8$). The “Baseline” refers to the original LoCoOp or SCT without auxiliary data, where \times denotes the ratio of the auxiliary data size to the ID data size in a batch. The ID data size in each batch is fixed at 32 for all experiments. The results indicate that as the number of auxiliary data increases, the OOD detection performance also improves and tends to gradually stabilize. The underlying reason is that the utilization of auxiliary data is relative to the expressiveness of ID data. When the ID feature information is fixed, the auxiliary data cannot infinitely improve the performance through more OOD features.

Hyperparameter Analysis. We conduct a comprehensive analysis on ϵ in Eq. 6, which is an important hyperparameter in foreground-background decomposition. Experiments are performed using ϵ values ranging from 0 (original image regions without pairing) to 1 (fully paired all image regions). Fig. 5a presents the average results across four standard OOD benchmarks, while Fig. 5b presents results on the hard OOD benchmark. Both experiences degraded performance at $\epsilon = 0$, *i.e.*, directly regularizing all original image regions. The performance improves significantly as the number of paired regions increases, demonstrating the advantage of foreground-background pairing. Notably, even when most image regions are paired, its performance still outperforms directly regularizing all original regions. This demonstrates the robustness of our image pairing method in benefiting from diverse OOD features in the auxiliary data. In general, determining ϵ is easy, as the performance of APT

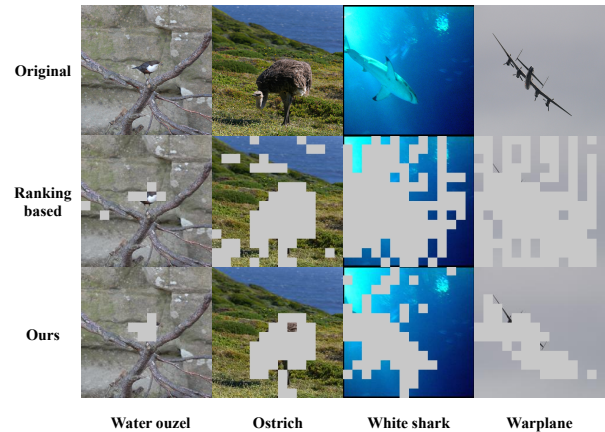


Figure 6. Visualization of extracted foreground regions (masked).

is generally stable. Furthermore, we also present the sensitivity of parameter α and β in Eq. 8 in Fig. 5c and Fig. 5d, which are also relatively stable within a range of values.

Quantitative Analysis. Fig. 6 presents a visual comparison of the extracted foreground regions from our proposed ALK and the ranking-based approach [22]. The results show that ALK more effectively identifies foreground regions, particularly in scenarios where background regions exhibit strong spurious semantic correlation with foreground regions (*e.g.*, fish in marine or planes against sky). This is because the ranking-based approach treats regions containing ground-truth labels within the top-k predicted classes as foreground. But the ground-truth class may wrongly appear in the top-k predictions of the background regions when these background regions (*e.g.*, seawater) have a relatively high semantic similarity with foreground objects (*e.g.*, fish), compared to other ID objects (*e.g.*, bird). In contrast, ALK quantifies the semantic alignment between local regions and global images based on prediction logits, effectively avoiding interference from spurious correlations and eliminating the requirement of ground-truth labels.

6. Conclusion

To address the problem of the lack of diversity in the extracted OOD features from few-shot ID data, we propose Auxiliary Prompt Tuning (APT), a novel framework that can be plugged into current prompt tuning-based methods to enable the utilization of diverse OOD feature from auxiliary data for enhanced OOD detection. APT first uses adaptive logit-based KL divergence to decompose foreground-background regions, then constructs foreground-background pairs to improve foreground OOD feature regularization. Comprehensive experiments across standard and hard OOD benchmarks show that APT achieves state-of-the-art performance, showing significant improvements in challenging scenarios.

Acknowledgments

In this research, the participation of W. Miao, Z. Wang, J. Zheng, and X. Bai was supported by National Natural Science Foundation of China (No. 62372029 and No. 62276016), while the participation of G. Pang was supported by A*STAR under its MTC YIRG Grant (M24N8c0103), the Ministry of Education, Singapore under its Tier-1 Academic Research Fund (24-SIS-SMU-008), and the Lee Kong Chian Fellowship (T050273).

References

- [1] Xiao Bai, Pengcheng Zhang, Xiaohan Yu, Jin Zheng, Edwin R Hancock, Jun Zhou, and Lin Gu. Learning from human attention for attribute-assisted visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3
- [2] Yichen Bai, Zongbo Han, Bing Cao, Xiaoheng Jiang, Qinghua Hu, and Changqing Zhang. Id-like prompt learning for few-shot out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17480–17489, 2024. 1, 3, 5, 6, 7
- [3] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 5
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 3, 5
- [5] Choubo Ding and Guansong Pang. Improving out-of-distribution detection with disentangled foreground and background features. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8923–8931, 2024. 2
- [6] Choubo Ding and Guansong Pang. Zero-shot out-of-distribution detection with outlier label exposure. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024. 3
- [7] Ruohuan Fang, Guansong Pang, and Xiao Bai. Simple image-level classification improves open-vocabulary object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1716–1725, 2024.
- [8] Ruohuan Fang, Guansong Pang, Wenjun Miao, Xiao Bai, Jin Zheng, and Xin Ning. Unsupervised recognition of unknown objects for open-world object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. 3
- [9] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018. 1, 2, 3, 7
- [10] Claus Hofmann, Simon Schmid, Bernhard Lehner, Daniel Klotz, and Sepp Hochreiter. Energy-based hopfield boosting for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 37:131859–131919, 2025. 3
- [11] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021. 1, 5
- [12] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 1
- [13] Marc Lafon, Elias Ramzi, Clément Rambour, Nicolas Audebert, and Nicolas Thome. Gallop: Learning global and local prompts for vision-language models. In *European Conference on Computer Vision*, pages 264–282. Springer, 2024. 2, 4, 5, 6
- [14] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1):17816, 2017. 1
- [15] Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. Learning transferable negative prompts for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17584–17594, 2024. 2
- [16] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020. 1, 2, 3, 7
- [17] Yuyuan Liu, Choubo Ding, Yu Tian, Guansong Pang, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. Residual pattern learning for pixel-wise out-of-distribution detection in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1151–1161, 2023. 3
- [18] Wenjun Miao, Guansong Pang, Tianqi Li, Xiao Bai, and Jin Zheng. Out-of-distribution detection in long-tailed recognition with calibrated outlier class learning. In *Proceedings of the 38th Annual AAAI Conference on Artificial Intelligence*, 2024. 2
- [19] Wenjun Miao, Guansong Pang, Jin Zheng, and Xiao Bai. Long-tailed out-of-distribution detection via normalized outlier distribution adaptation. *Advances in Neural Information Processing Systems*, 37:132106–132132, 2024. 1
- [20] Wenjun Miao, Guansong Pang, Trong-Tung Nguyen, Ruohuan Fang, Jin Zheng, and Xiao Bai. Opencil: Benchmarking out-of-distribution detection in class incremental learning. *Pattern Recognition*, 171:112163, 2026. 1
- [21] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyun Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35:35087–35102, 2022. 2, 4, 5, 6
- [22] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. *Advances in Neural Information Processing Systems*, 36:76298–76310, 2023. 1, 2, 3, 5, 6, 7, 8
- [23] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Zero-shot in-distribution detection in multi-object settings using vision-language foundation models. *arXiv preprint arXiv:2304.04521*, 2023. 2, 4, 5, 6
- [24] Jun Nie, Yonggang Zhang, Zhen Fang, Tongliang Liu, Bo Han, and Xinmei Tian. Out-of-distribution detection with

- negative prompts. In *The twelfth international conference on learning representations*, 2024. 2, 5, 6
- [25] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12173–12182, 2020. 3
- [26] 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 *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1
- [27] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lih Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 3
- [28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 1, 3, 5, 6
- [29] Sina Sharifi, Taha Entesari, Bardia Safaei, Vishal M Patel, and Mahyar Fazlyab. Gradient-regularized out-of-distribution detection. In *European Conference on Computer Vision*, pages 459–478. Springer, 2024. 3
- [30] Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and Gustavo Carneiro. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In *European Conference on Computer Vision*, pages 246–263. Springer, 2022. 2
- [31] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 5
- [32] Haotao Wang, Aston Zhang, Yi Zhu, Shuai Zheng, Mu Li, Alex J Smola, and Zhangyang Wang. Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition. In *International Conference on Machine Learning*, pages 23446–23458. PMLR, 2022. 1, 3, 5, 6
- [33] Hualiang Wang, Yi Li, Hui Feng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812, 2023. 2
- [34] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 2, 5
- [35] Jing Kang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8301–8309, 2021. 3
- [36] Geng Yu, Jianing Zhu, Jiangchao Yao, and Bo Han. Self-calibrated tuning of vision-language models for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 37:56322–56348, 2024. 1, 2, 3, 4, 5, 6, 7
- [37] Haotian Zheng, Qizhou Wang, Zhen Fang, Xiaobo Xia, Feng Liu, Tongliang Liu, and Bo Han. Out-of-distribution detection learning with unreliable out-of-distribution sources. *Advances in neural information processing systems*, 36:72110–72123, 2023. 3
- [38] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 2, 5
- [39] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 2, 4
- [40] Kaiyang Zhou, Jing Kang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3