Harnessing Large Language and Vision-Language Models for Robust Out-of-Distribution Detection

Pei-Kang Lee   R11922116@NTU.EDU.TW
Dept. of Computer Science and Information Engineering, National Taiwan University

Jun-Cheng Chen   PULLPULL@CITI.SINICA.EDU.TW
Research Center for Information Technology Innovation, Academia Sinica

Ja-Ling Wu   WJL@CMLAB.CSIE.NTU.EDU.TW
Dept. of Computer Science and Information Engineering, National Taiwan University

Editors: Hung-yi Lee and Tongliang Liu

Abstract

Out-of-distribution (OOD) detection has seen significant advancements with zero-shot approaches by leveraging the powerful Vision-Language Models (VLMs) such as CLIP. However, prior research works have predominantly focused on enhancing Far-OOD performance, while potentially compromising Near-OOD efficacy, as observed from our pilot study. To address this issue, we propose a novel strategy to enhance zero-shot OOD detection performances for both Far-OOD and Near-OOD scenarios by innovatively harnessing Large Language Models (LLMs) and VLMs. Our approach first exploit an LLM to generate superclasses of the ID labels and their corresponding background descriptions followed by feature extraction using CLIP. We then isolate the core semantic features for ID data by subtracting background features from the superclass features. The refined representation facilitates the selection of more appropriate negative labels for OOD data from a comprehensive candidate label set of WordNet, thereby enhancing the performance of zero-shot OOD detection in both scenarios. Furthermore, we introduce novel few-shot prompt tuning and visual prompt tuning to adapt the proposed framework to better align with the target distribution. Experimental results demonstrate that the proposed approach consistently outperforms current state-of-the-art methods across multiple benchmarks, with an improvement of up to 2.9% in AUROC and a reduction of up to 12.6% in FPR95. Additionally, our method exhibits superior robustness against covariate shift across different domains, further highlighting its effectiveness in real-world scenarios.

Keywords: Out-of-distribution detection, VLM, LLM

1. Introduction

Out-of-distribution (OOD) detection is a critical and active research area that aims to identify whether input samples belong to the distribution of the training data used in machine learning (ML) models. This capability is crucial for preventing unpredictable outputs from ML models when faced with unseen inputs in real-world applications Ulmer et al. (2020).

Traditionally, OOD detection approaches have primarily relied on Convolutional Neural Networks (CNN), focusing on analyzing visual features Lee et al. (2018); Sastry and Oore (2020), or logit spaces Liu et al. (2020). However, these methods are usually limited by either their limited model capacities or their reliance on visual information alone. The emergence of powerful Vision-Language Models (VLMs) like Contrastive Language-Image Pretraining (CLIP) Radford et al. (2021), which is trained with 400 million image-text pairs, has opened new avenues for tackling OOD detection by utilizing their rich semantic information in both visual and label spaces.

Recent studies Esmaeilpour et al. (2022); Ming et al. (2022); Jiang et al. (2024) have explored using CLIP for zero-shot OOD detection, demonstrating promising results. For instance, one of the representative CLIP-based zero-shot OOD detectors, NegLabel Jiang et al. (2024), augments the class labels for ID data with the selected negative labels from the noun and adjective lexname categories of WordNet Fellbaum (1998) as the proxies for OOD classes to enhance the performance. However, these works mainly focus on studying Far-OOD scenarios where the distributions of in-distribution (ID) and OOD data are distant. Additionally, while zero-shot OOD detection methods using CLIP have shown great potential, they may struggle to capture the nuances and specific characteristics of downstream tasks Wei et al. (2023). Therefore, when the distributions of ID and OOD data are similar for the Near-OOD scenarios, it is still a challenging problem requiring more discriminative and detailed information.

To tackle this challenge, we propose a novel approach to enhance the power of CLIP for zero-shot OOD detection in both Near-OOD and Far-OOD scenarios by expanding the selection space of OOD candidate labels and refining the semantic space of ID labels. To achieve the expansion, we employ a comprehensive set of WordNet categories for candidate OOD label selection by performing a similar negative-mining algorithm as NegLabel without any candidate label filtering. As we find from our experimental results (c.f. Table 5) when NegLabel performs negative-mining for appropriate negative labels, it filters out specific candidate labels and leads to suboptimal performance. Next, to refine the semantic space of ID labels, we use Large Language Models (LLMs) to yield superclass labels and background descriptions to create a proxy for selecting negative labels. This process involves initially broadening the ID semantic space through superclass identification with LLM followed by a subtraction step of background information to refine the space. The refined ID space enables negative-mining to select more representative negative labels from the expanded selection space of OOD candidate labels. With our proposed hierarchical strategy, our zero-shot approach outperforms not only current state-of-the-art zero-shot but also CLIP-based training methods, demonstrating the effectiveness of combining VLMs with LLMs for OOD detection.

To further enhance the performance of our zero-shot OOD method, we adapt CLIP to the target distributions by our two-phase training method with few-shot prompt tuning (PT) and visual prompt tuning (VPT). Moreover, to address the practical challenge of collecting unlabeled OOD samples for training, we adopt the same approach as ID-like Bai et al. (2024) by leveraging less label-relevant portions of ID samples as OOD data to create a robust OOD proxy dataset while maintaining the inherent characteristics of the ID domain. Together with this method, it allows us to benefit from outlier exposure during few-shot training without relying on external OOD data. Our proposed few-shot method has been extensively evaluated on multiple benchmarks with state-of-the-art performances, including the ImageNet-1K OOD benchmark Huang et al. (2021) and the challenging OpenOOD V1.5 full-spectrum benchmark Zhang et al. (2023); Yang et al. (2022). Our method significantly outperforms existing approaches, achieving the average AUROC of 97.67% on the ImageNet-1K OOD benchmark and the average AUROC of 83.04% on the challenging OpenOOD V1.5 Near-OOD full-spectrum benchmark. In summary, our key contributions are as follows:

- Improved zero-shot OOD detection: We introduce a hierarchical strategy using LLMs to generate superclass labels and background descriptions, creating more representative negative labels to enhance the zero-shot OOD detection performance.

- Novel few-shot OOD detection method: We propose a new two-phase few-shot learning framework that exploits our proposed PT and VPT to adapt CLIP to the target distributions.
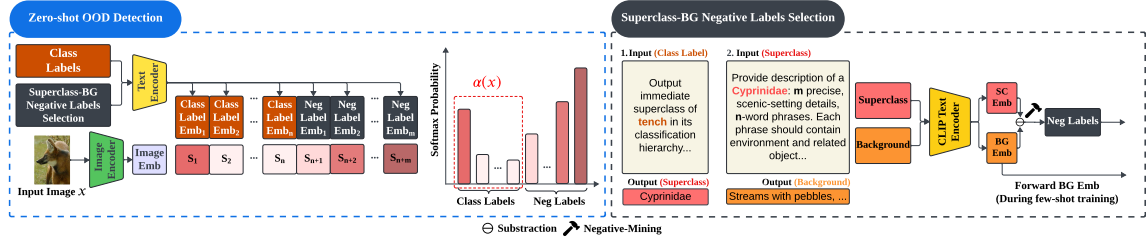
Figure 1: Illustration of the proposed zero-shot OOD detection and Superclass-BG negative label selection. We harness the capabilities of LLMs to select more representative negative labels. The full LLM prompt can be found in Appendix B.2.
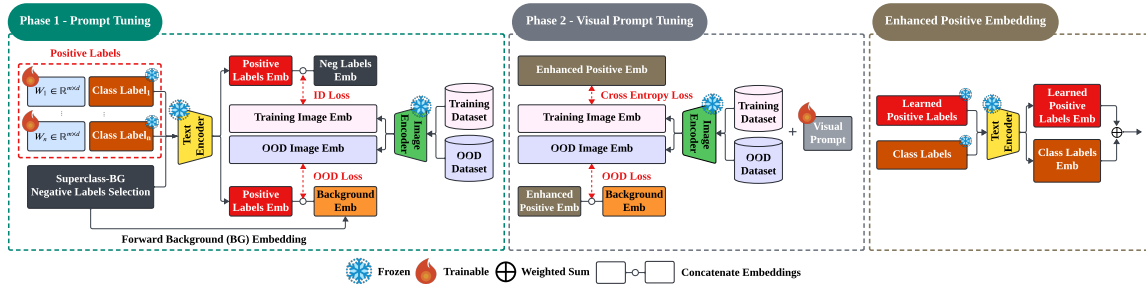

Figure 2: Illustration of the proposed few-shot learning framework. The training dataset consists of a few-shot sample from the ImageNet-1K training set, while the OOD dataset is generated using an ID-like Bai et al. (2024) approach. Here, the term *positive labels* refer to the class labels with learnable prompt.

Together with the ID-like auxiliary OOD data generation, the proposed approach achieves state-of-the-art OOD detection performance.

2. Related Work

While OOD detection has a rich history in computer vision, our work focuses on leveraging VLMs for OOD detection. For a comprehensive review of traditional OOD detection methods using CNNs and Vision Transformers Dosovitskiy et al. (2021), please refer to Appendix A. Here, we focus on recent advancements in CLIP-based OOD detection.

CLIP-based OOD Detection. VLMs, such as CLIP, possess formidable zero-shot capabilities and robustness against distribution shifts. Recognizing these strengths, an increasing number of researchers are focusing on harnessing the power of VLMs to enhance OOD detection Miyai et al. (2024). ZOC Esmaeilpour et al. (2022) trains an image descriptor to obtain descriptions of images, using these additional descriptions as supplementary OOD labels. MCM Ming et al. (2022) uses class labels as prototypes and exploits the uniformity property of contrastive learning to enhance the discrepancy between ID and OOD samples through softmax rescaling. Another research direction aims to teach CLIP the logic of contrary semantics, namely the concept of *no*. CLIPN Wang et al. (2023) fine-tunes using the CC12M Changpinyo et al. (2021) dataset, a *no* text encoder, and *no* prompts to teach CLIP the logic of *no*. Meanwhile, LSN Nie et al. (2024) leverages ID data to learn

this *no* logic. NegPrompt Li et al. (2024) learns a set of negative prompts that contain semantics opposite but related to the ID labels. Furthermore, employing additional OOD labels to make ID images gravitate towards ID labels and away from selected OOD labels is also a popular research direction. TOE Park et al. (2023b) demonstrates the utility of textual outliers, achieving promising results through training methods similar to image outlier exposure. ID-like Bai et al. (2024) uses random cropping to identify images most closely aligned and unrelated to the corresponding label, aiming to align textual prompts and learnable OOD labels with these images. NegLabel Jiang et al. (2024), on the other hand, employs negative-mining from a large word corpus to identify labels semantically distant from ID semantic space.

3. Preliminaries

Zero-shot Classification with CLIP. CLIP consists of a text encoder $\mathcal{T} : t \rightarrow \mathbb{R}^d$ and an image encoder $\mathcal{I} : x \rightarrow \mathbb{R}^d$, mapping text t and image x to d-dimensional feature vectors. For zero-shot classification with label set $C = \{c_1, ..., c_m\}$, the vectors t_j are constructed using prompts "a photo of a $\langle c_j \rangle$". The similarity score between image features $\mathcal{I}(x)$ and text features $\mathcal{T}(t_j)$ is computed as the cosine similarity of $\mathcal{T}(t_j)$ and $\mathcal{I}(x)$. The predicted probability for class c_j is computed using the softmax function over the similarity scores, with a temperature parameter τ.

Prompt Tuning and Visual Prompt Tuning Prompt tuning (PT) methods like CoOp Zhou et al. (2022b) initialize learnable prompt tensors $t = [V]_1...[V]_L[\text{CLASS}]$, where $[V]_l$ are learnable vectors. These prompts are optimized using few-shot examples to improve the performance on target distributions. The concept of the visual prompt tuning (VPT) Jia et al. (2022), akin to that used in CoOp, involves the insertion of fixed-length learnable vectors between the image patch embeddings and the class token.

Out-of-Distribution Detection. OOD detection aims to construct a binary classifier $G(x)$:

$$G(x) = \begin{cases} \text{ID}, & \text{if } \alpha(x) \geq \gamma \\ \text{OOD}, & \text{if } \alpha(x) < \gamma \end{cases} \tag{1}$$

where $\alpha(x)$ is a score function and γ is a threshold.

4. Methodology

In this section, we describe the detailed of the proposed zero-shot and few-shot training OOD detection as follows.

4.1. Zero-Shot OOD Detection

As shown in Figure 1, we perform similar CLIP-based zero-shot classification for zero-shot OOD detection. Moreover, we augment the class labels for ID data with negative labels for OOD classes from a comprehensive set of WordNet categories using our proposed LLM-based Superclass-**B**ack**G**round (Superclass-BG) negative label selection algorithm, as detailed below (i.e., we do not apply any candidate label filtering during the process). Additionally, the selected negative labels are partitioned into K distinct groups following the grouping strategy in NegLabel Jiang et al. (2024).

LLM-based Superclass-BG Negative Label Selection. Let $C = \{c_1, c_2, ..., c_m\}$ denote the set of class labels, $SC = \{sc_1, sc_2, ..., sc_n\}$ the set of superclass labels where SC is generated from C through LLMs, $n \leq m$, and $N = \{n_1, n_2, ..., n_p\}$ the set of candidate labels. Define $S(u, v)$ as the measure of cosine similarity between the embedding u and v. The negative-mining algorithm introduced in NegLabel calculates $S(\mathcal{T}(n_i), \mathcal{T}(c_j))$ for all $n_i \in N$ and $c_j \in C$, where $\mathcal{T}$ denotes the CLIP text encoder. Defining $\text{Score}(n_i)$ as the q-quantile of $\{S(\mathcal{T}(n_i), \mathcal{T}(c_j))|c_j \in C\}$. This approach of using a quantile, rather than a minimum distance or average similarity, provides robustness against outlier ID labels by considering the overall distribution of similarities. The negative labels are determined by selecting the labels corresponding to the lowest p percentage of $\text{Score}(n_i)$ values, where p is a predefined threshold. Our proposed method extends this approach by leveraging superclass labels generated by LLMs. We calculate $\text{Score}'(n_i)$ as the q-quantile of $\{S(\mathcal{T}(n_i), \mathcal{T}(sc_j))|sc_j \in SC\}$. This superclass-based approach offers an improved representation of high similarity across a broader range of concepts, as each sc_j encapsulates multiple related classes, mitigating the influence of semantically similar class clusters on the quantile calculation. We leverage the semantic comprehension capability of CLIP Goh et al. (2021) to further refine the semantic representation. For each superclass label sc_j, we also generate a background description $BG(sc_j)$ using LLMs. We then compute adjusted superclass label embeddings as $\mathcal{T}'(sc_j) = \mathcal{T}(sc_j) - \mathcal{T}(BG(sc_j))$. This subtraction aims to isolate the core semantic content of each superclass label by removing general background features. The rationale behind this additional step is twofold. First, it addresses potential inaccuracies in LLM-generated superclass labels, which may not perfectly align with optimal taxonomies. Second, it focuses the embeddings on distinguishing features of the superclass labels, potentially reducing semantic overlap.

4.2. Two-Phase Training for Few-shot OOD Detection

Our method is structured into two distinct phases. In the initial phase, we focus exclusively on prompt tuning for class labels while maintaining all other model components fixed. The second phase builds upon the results of the first. We fix the tuned prompts obtained from the previous stage and proceed with VPT. The overview is illustrated in Figure 2. To derive the OOD datasets from few-shot training samples, we employ an ID-like Bai et al. (2024) approach. Multiple cropped images are initially extracted from each training sample via random cropping. Subsequently, these cropped images are bifurcated based on their cosine similarity to the corresponding class label. Images exhibiting lower similarity are allocated to the OOD dataset. This approach enables the creation of a robust OOD proxy dataset while maintaining the inherent characteristics of the ID domain.

Phase-1 Prompt Tuning. To facilitate subsequent discussions, we introduce the term *positive labels*, denoted as P, to refer to the class labels with learnable prompts while all other components remain fixed. Let $\mathcal{D}_{\text{train}}$ be the few-shot training dataset, $\mathcal{D}_{\text{ID}}$ be the ID dataset and $\mathcal{D}_{\text{OOD}}$ be the OOD dataset. The dissimilarity measure between embedding u and v is given by $D(u, v) = 1 - S(u, v)$. Let C be the set of class labels. Let x be the input image from the few-shot training dataset, y be the corresponding ground-truth label and τ be the temperature. During the prompt tuning phase, the optimization process incorporates two distinct loss functions: the ID loss and the OOD loss. The ID loss is formulated in Equation 2 and 3.

$$\mathcal{L}_{\text{ID}} = \max_k \mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{train}}} \left[\mathcal{F}_k(x, y) \right], \tag{2}$$

where

$$\mathcal{F}_k(x,y) = -\log \frac{e^{(S_y + \lambda_1 D_y)/\tau}}{\sum_{i=1}^{|P|} e^{(S_i^p + \lambda_1 D_i^p)/\tau} + \sum_{j=1}^{|N_k|} e^{(\lambda_2(S_j^n + D_j^n))/\tau}}. \tag{3}$$

In Equation 3, negative labels selected by Superclass-BG are partitioned into K groups. Let N_k denote the kth negative label group. Define the similarity between the image and the ground-truth positive label as $S_y = S(\mathcal{I}(x), \mathcal{T}(P_y))$ and dissimilarity between ground-truth positive label and ground-truth class label as $D_y = -D(\mathcal{T}(C_y), \mathcal{T}(P_y))$. Similarly, we also define the similarity between the image and ith positive label as $S_i^p = S(\mathcal{I}(x), \mathcal{T}(P_i))$ and the corresponding positive dissimilarity as $D_i^p = D(\mathcal{T}(P_y), \mathcal{T}(P_i)) - D(\mathcal{T}(C_y), \mathcal{T}(P_i))$. Additionally, we denote the similarity between the image and j-th negative label in group k as $S_j^n = S(\mathcal{I}(x), \mathcal{T}(N_{k_j}))$ and the corresponding negative dissimilarity as $D_j^n = -D(\mathcal{T}(P_y), \mathcal{T}(N_{k_j}))$. Finally, λ_1 and λ_2 are weighting hyperparameters. The ID loss aims to enhance the alignment between positive label embeddings and image embeddings while preserving the robust generalization capabilities inherent in the pre-trained CLIP model. Additionally, it aims to maximize the distinction between the image embeddings and the most challenging group of negative labels, thereby improving the model's discriminative power in the worst-case scenario. Through the utilization of S_i^p and D_i^p, we refine the semantic relationships between positive labels, encouraging intra-class cohesion while preserving some aspects of the original inter-class structure. Concurrently, D_j^n further amplifies the distance between positive and negative labels. Conversely, the OOD loss, formalized in Equation 4 and 5, is designed to minimize the overall similarity between OOD images x, sourced from $\mathcal{D}_{\text{OOD}}$, and the positive labels, while explicitly exposing the model to background description generated through the LLMs.

$$\mathcal{L}_{\text{OOD}} = \mathbb{E}_{(x) \sim \mathcal{D}_{\text{OOD}}} \big[\mathcal{G}(x) \big], \tag{4}$$

where

$$\mathcal{G}(x) = \frac{e^{\sum_{i=1}^{|P|} (S(\mathcal{I}(x), \mathcal{T}(P_i))/\tau)}}{\sum_{i=1}^{|P|} e^{(S(\mathcal{I}(x), \mathcal{T}(P_i))/\tau)} + \sum_{j=1}^{|SC|} e^{(S(\mathcal{I}(x), \mathcal{T}(BG(sc_j)))/\tau)}}. \tag{5}$$

To construct our final training objective, we employ gradient-aware scaling before balancing the two loss components. Specifically, we scale the OOD loss to match the gradient magnitude of the ID loss, ensuring comparable influence on model updates.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{scaled\_OOD}} + \lambda_3 \cdot \mathcal{L}_{\text{ID}}, \tag{6}$$

where λ_3 is a balancing hyperparameter.

Phase-2 Visual Prompt Tuning. In the first phase, we exclusively employ PT of the label space to enhance the ability of positive labels to distinguish ID and OOD samples while preserving the visual architecture of the model. This approach prompts an inquiry into potential OOD detection improvements achievable through visual space tuning. Post-acquisition of the refined positive labels via PT, we implement VPT to align the training images with these labels using cross entropy and OOD losses. The process initiates with a enhanced positive embedding derived from a weighted sum of the learned positive labels and class labels. This methodology leverages class labels to preserve

the pretrained knowledge of CLIP, mitigating overfitting risks in few-shot training scenarios. Let x be the input image with a learnable visual prompt from a few-shot training dataset and y be the corresponding ground truth label, and P' denote the enhanced positive embeddings. The cross entropy loss is formulated as

$$\mathcal{L} = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{train}}} \left[-\log \frac{e^{S(\mathcal{I}(x),P'_y)/\tau}}{\sum\limits_{i=1}^{|P'|} e^{S(\mathcal{I}(x),P'_i)/\tau}} \right]. \tag{7}$$

The design of the OOD loss for VPT aligns with the principles of PT, both aiming to minimize the similarity between OOD images and learned positive labels. The OOD loss for VPT is formulated as

$$\mathcal{L}_{\text{VPT\_OOD}} = \mathbb{E}_{(x)\sim\mathcal{D}_{\text{OOD}}} \big[\mathcal{H}(x)\big], \tag{8}$$

where

$$\mathcal{H}(x) = \frac{\max\limits_{i,0\leq i\leq|P'|} e^{(S(\mathcal{I}(x),P'_i)/\tau)}}{\sum\limits_{i=1}^{|P'|} e^{(S(\mathcal{I}(x),P'_i)/\tau)} + \sum\limits_{j=1}^{|SC|} e^{(S(\mathcal{I}(x),\mathcal{T}(BG(sc_j)))/\tau)}}. \tag{9}$$

In this phase, we deal with a single learnable visual prompt while the positive labels are now fixed. With only one learnable prompt, optimizing against all positive labels simultaneously might be too constrained and potentially lead to conflicts. Given the current visual representation, the max operation allows the model to focus on the most critical case - the positive label most similar to the OOD sample. To formulate our final training objective for VPT, we employ a methodology analogous to that used in the prompt tuning phase.

$$\mathcal{L}_{\text{VPT\_total}} = \mathcal{L} + \lambda_4 \cdot \mathcal{L}_{\text{scaled\_VPT\_OOD}}, \tag{10}$$

where λ_4 is a balancing hyperparameter.

Few-shot Inference Similar to zero-shot OOD setting, The negative labels, mined through the proposed Superclass-BG approach, are also partitioned into K distinct groups during inference. We compute softmax probabilities for each group, incorporating both enhanced positive embeddings and the respective negative label group. The OOD detection score function aggregates probabilities assigned to enhanced positive embeddings across all groups. Formally, for input image x, enhanced positive label embeddings P', negative label groups N_k, and total group count K, the score function $\alpha(x)$ is defined as:

$$\alpha(x) = \sum_{i=1}^{K} \frac{\sum\limits_{i=1}^{|P'|} e^{(S(\mathcal{I}(x),P'_i)/\tau)}}{\sum\limits_{i=1}^{|P'|} e^{(S(\mathcal{I}(x),P'_j)/\tau)} + \sum\limits_{j=1}^{|N_k|} e^{(S(\mathcal{I}(x),\mathcal{T}(N_{k_j})/\tau)}}. \tag{11}$$

Table 1: Comparisons of the proposed method and competitive baselines on the ImageNet-1K as the ID dataset. The best result in each column is in bold, and the second-best is underlined. The values for Superclass and Superclass-BG are derived from the average of 3 independent generations using identical prompts. All methods are based on CLIP-B/16, which employs a ViT-B/16 as the image encoder and a masked self-attention Transformer as the text encoder. All values are percentages. Performance metrics for baseline methods are cited from NegLabel and their respective original publications. The # column indicates the required number of samples per class for training, where F denotes full fine-tuning. The methods with * indicate the requirement of additional training. The shaded part represents our method.

Method	#	iNaturalist		SUN		Places		Texture		Average	
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
					zero-shot						
Mahalanobis Lee et al. (2018)	0	99.33	55.89	99.41	59.94	98.54	65.96	98.46	64.23	98.94	61.50
Energy Liu et al. (2020)	0	81.08	85.09	79.02	84.24	75.08	83.38	93.65	65.56	82.21	79.57
ZOC Esmaeilpour et al. (2022)	*	87.30	86.09	81.51	81.20	73.06	83.39	98.90	76.46	85.19	81.79
MCM Ming et al. (2022)	0	20.91	94.61	37.59	92.57	44.69	89.77	57.77	86.11	42.74	90.77
CLIPN Wang et al. (2023)	*	23.94	95.27	26.17	93.93	33.45	92.28	40.83	90.93	31.10	93.10
NegLabel Jiang et al. (2024)	0	1.91	99.49	20.53	95.49	35.59	91.64	35.59	90.22	25.40	94.21
Ours (Superlcass)	0	1.26	99.68	20.74	94.85	35.80	91.60	47.95	88.21	26.44	93.58
Ours (Superclass-BG)	0	1.15	99.71	13.36	96.64	26.62	93.70	44.26	89.57	21.34	94.90
					requires training						
NPOS Tao et al. (2023)	F	16.58	96.19	43.77	90.44	45.27	89.44	46.12	88.80	37.93	91.22
LSN Nie et al. (2024)	64	21.56	95.83	23.62	94.35	34.48	91.25	38.54	90.42	30.22	92.96
ID-like Bai et al. (2024)	4	8.98	98.19	42.03	91.64	44.00	90.57	25.27	94.32	30.07	93.68
LoCoOp Miyai et al. (2023)	16	16.05	96.86	23.44	95.07	32.87	91.98	42.28	90.19	28.66	93.52
NegPrompt Li et al. (2024)	16	6.32	98.73	22.89	95.55	27.60	93.34	35.21	91.60	23.01	94.81
Ours (Train)	4	**0.65**	**99.83**	**12.07**	96.91	23.21	94.47	28.10	93.80	16.01	96.25
Ours (Train)	16	0.68	99.81	13.15	**96.94**	**16.03**	**96.55**	**11.74**	**97.38**	**10.4**	**97.67**

5. Experiments

Datasets and Benchmarks. We utilize the ImageNet-1K OOD benchmark Huang et al. (2021) to compare our method with existing zero-shot and few-shot training-based OOD detection approaches. The ImageNet-1K OOD benchmark employs ImageNet-1K as the ID dataset and uses iNaturalist Van Horn et al. (2018), SUN Xiao et al. (2010), Places Zhou et al. (2018), and Texture Cimpoi et al. (2014) as OOD datasets, which have no class overlap with the ID dataset. For the more challenging OOD detection, we follow the settings of MCM Ming et al. (2022), which adopt ImageNet-10 and ImageNet-20 alternately as ID and OOD datasets. Beyond the aforementioned datasets, we evaluate our proposed method on OpenOOD V1.5 ImageNet-1K full-spectrum benchmark Zhang et al. (2023); Vaze et al. (2022). These benchmarks introduce more challenging Near-OOD datasets, such as SSB-hard Vaze et al. (2022) and NINCO Bitterwolf et al. (2023), and also test robustness against covariate shifts. **Implementation Details.** We employ CLIP-B/16 as our backbone architecture, which uses a ViT-B/16 as the image encoder and a masked self-attention Transformer as the text encoder. For few-shot training, we randomly sample 16 images per class. We create 256 random crops per image to generate OOD datasets and select the bottom two crops based on label similarity. Claude 3.5 Sonnet Anthropic (2024) generates 10 three-word background descriptions using a temperature of 0. Following NegLabel Jiang et al. (2024), we use the lowest 15% similarity scores in nouns and adjectives as negative labels. Prompt tuning is performed for 200 epochs, with batch size 256 and learning rate 0.025 (SGD). VPT is trained for 5 epochs, with

Table 2: Comparison of the OpenOOD V1.5 full-spectrum benchmark Zhang et al. (2023); Yang et al. (2022) between competitive baseline, including and the current state-of-the-art, NNGuide Park et al. (2023a), LSA Lu et al. (2023) and CoCoOp Zhou et al. (2022a). Near-OOD is the average of SSB-hard and NINCO, while Far-OOD is the average of iNaturalist, Texture, and OpenImage-O Wang et al. (2022).

Method	#	Near-OOD		Far-OOD	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑
zero-shot					
MCM	0	94.74	58.11	77.47	82.56
NegLabel	0	76.97	70.11	28.80	93.52
Ours (Superclass-BG)	0	66.52	77.98	32.53	93.03
requires training					
NNGuide	F	74.28	71.85	34.28	92.24
CoOp	16	95.82	56.36	80.31	82.41
CoCoOp	16	96.23	53.11	79.59	74.48
LoCoOp	16	90.91	59.34	54.33	84.02
LSA	16	70.56	78.22	48.06	86.85
Ours (Train)	16	**51.35**	**83.04**	**17.36**	**95.73**

Table 3: Zero-shot performance comparison of OOD detection methods on challenging OOD datasets ImageNet-10 and ImageNet-20.

ID Dataset	OOD Dataset	Method	FPR95↓	AUROC↑
		MCM	6.69	98.45
ImageNet-10	**ImageNet-20**	NegLabel	8.69	98.39
		Superclass-BG	**6.26**	**98.70**
		MCM	9.46	98.33
ImageNet-20	**ImageNet-10**	NegLabel	8.51	98.27
		Superclass-BG	**6.56**	**98.46**

batch size 32 and learning rate 0.2 (SGD). ID loss hyperparameters are $\lambda_1 = \lambda_2 = 0.25$. ID and OOD loss balancing weights are $\lambda_3 = \lambda_4 = 0.3$ Enhanced positive embeddings incorporate equal contributions (0.5) for class labels and learned positive labels. All experiments are conducted on a single NVIDIA 4090 GPU. Further training details are fully described in Appendix B.1.

Evaluation Metrics. We employ two primary metrics: (1) FPR95, which is the probability that an OOD example is misclassified as ID when the true positive rate is as high as 95%; (2) Area Under the Receiver Operating Characteristic (AUROC).

5.1. Results and Discussions

Main Results. Table 1 presents a comprehensive comparison of our proposed method against existing OOD detection approaches. On the ImageNet OOD benchmark, our method achieves an average improvement of 2.8% in AUROC and a reduction of 12.6% in FPR95 compared to the current
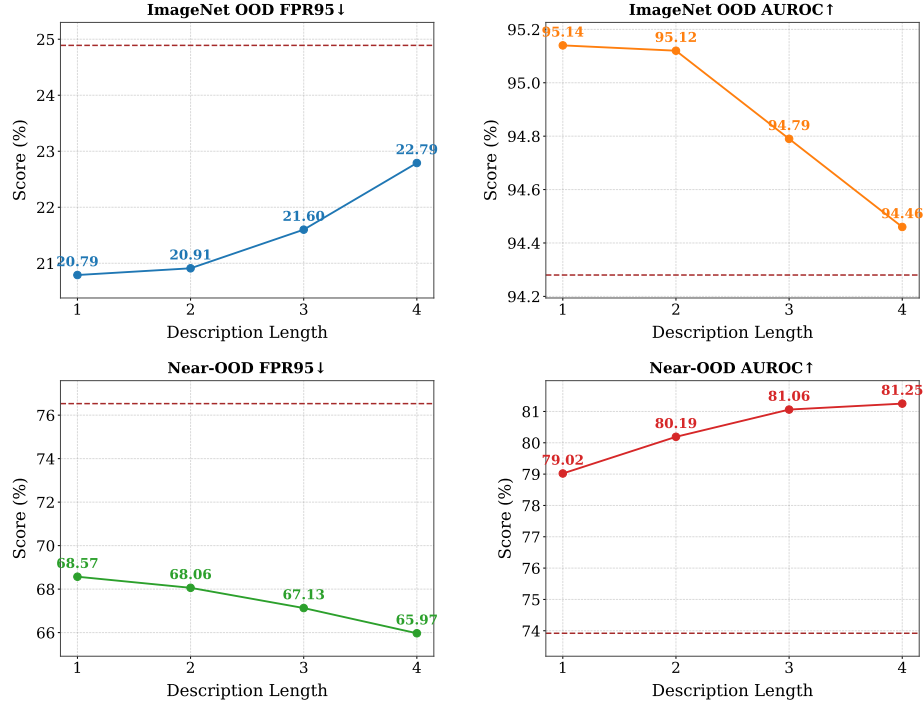
Figure 3: Superclass-BG zero-shot performance metrics as a function of description length. Each subplot shows the trend of a specific OOD detection metric for description lengths ranging from 1 to 4. Data points represent average scores computed from 5-15 descriptions. The brown dashed line indicates the performances of NegLabel Jiang et al. (2024).

state-of-the-art. For the more challenging OpenOOD V1.5 full-spectrum benchmark, the results are shown in Table 2. Our method exhibits remarkable robustness to covariate shifts. In the Far-OOD detection scenarios, we observe a 2.2% increase in AUROC and a 9.4% decrease in FPR95. The improvement is even more pronounced in Near-OOD scenarios, with an 4.8% increase in AUROC and a 19.2% decrease in FPR95. Notably, our zero-shot Superclass-BG approach surpasses current training-based state-of-the-art methods on the ImageNet-1K OOD benchmark. It also performs superior on the challenging ImageNet-10 and ImageNet-20 benchmarks, as evidenced in Table 3.

Influence of the Quantity of Background Description on Zero-Shot Performance. Figure 3 reveals that shorter descriptions enhance performance on the ImageNet-1K OOD benchmark, while longer descriptions prove more effective for Near-OOD detection. We hypothesize that shorter descriptions, limited in capturing nuanced background features, may be oversimplified by omitting general characteristics. This approach benefits the simpler ImageNet-1K OOD benchmark but potentially removes possible indicators for distinguishing challenging Near-OOD samples. Conversely, removing broader background concepts might reduce overlap with straightforward OOD samples, contributing to improved performance on the ImageNet-1K OOD benchmark. On the other hand, longer descriptions provide a more comprehensive representation of both specific and general attributes, aiding in Near-OOD sample differentiation where similarities with ID data are shared. However, longer descriptions may retain general characteristics that overlap with straightforward

Table 4: Superclass-BG zero-shot OOD detection performance comparison using various LLMs on ImageNet OOD and Near-OOD datasets. The results are the average of 3 independent generations using identical prompts.

LLMs	ImageNet OOD		Near-OOD	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
Claude 3.5 Sonnet	21.34	94.90	67.22	81.04
GPT-4-Turbo	21.39	94.93	**66.60**	**81.16**
GPT-4o	22.07	94.62	67.44	81.05
GPT-4o-mini	**20.18**	**95.13**	68.74	80.27

OOD samples, presenting a trade-off in effectiveness across different OOD detection scenarios.

Various Large Language Models. The results presented in Table 4 show that utilizing smaller LLMs such as GPT-4o-mini OpenAI (2024) yields performance similar to employing fewer words for background description generation. This phenomenon may be attributed to the limited capacity of these models to articulate nuanced features. Compared to GPT-4-Turbo, Claude 3.5 Sonnet demonstrates comparable performance while offering a more affordable solution with a faster generation speed. Consequently, we have opted to employ Claude 3.5 Sonnet for our experiments.

Computational Complexity. Training takes 194 minutes for phase 1 and 6 minutes for phase 2 on average, using a single NVIDIA 4090 GPU. Following the score function proposed by NegLabel, the computational complexity of our proposed method incurs about $\mathcal{O}(2Md)$ FLOPs extra computational burden per image during inference, where M denotes the number of negative labels and d represents the dimension of the embedding feature.

5.2. Ablation Study

The Effectiveness of Prompt Tuning and Visual Prompt Tuning. In Table 5, we find that few-shot prompt tuning yields performance improvements on both ImageNet OOD and Near-OOD datasets, regardless of whether Superclass-BG or filtering is employed. Notably, the enhancement is particularly pronounced for Near-OOD, with AUROC improvement of at least 6.6%. Furthermore, the results show that the application of VPT in the second phase leads to additional performance gains.

Pilot study of Candidate Label Filtering. When using ImageNet as the ID dataset, words from animal and food categories are excluded in NegLabel due to their prevalence classes; however, from the zero-shot and few-shot training results presented in Table 5, we observe that filtered methods consistently outperform their unfiltered counterparts on the ImageNet OOD benchmark. Notably, the performance disparity is substantially more pronounced on Near-OOD datasets. We hypothesize that the slight decrease in performance on ImageNet OOD, observed when reincorporating candidate labels belonging to the major lexname category of the ID labels into the pool of selectable negative labels, may be attributed to the potential selection of negative labels semantically proximate to the major category. Conversely, the significant performance improvement on Near-OOD datasets can be explained by the inherent semantic proximity of Near-OOD samples to the ID semantic space. Consequently, these semantically related negative labels effectively serve as discriminative

Table 5: Performance comparisons of various configurations in the proposed methods with ImageNet-1K as ID dataset. S-BG denotes Superclass-BG. Results are reported in terms of AU-ROC, with higher values indicating better performance.

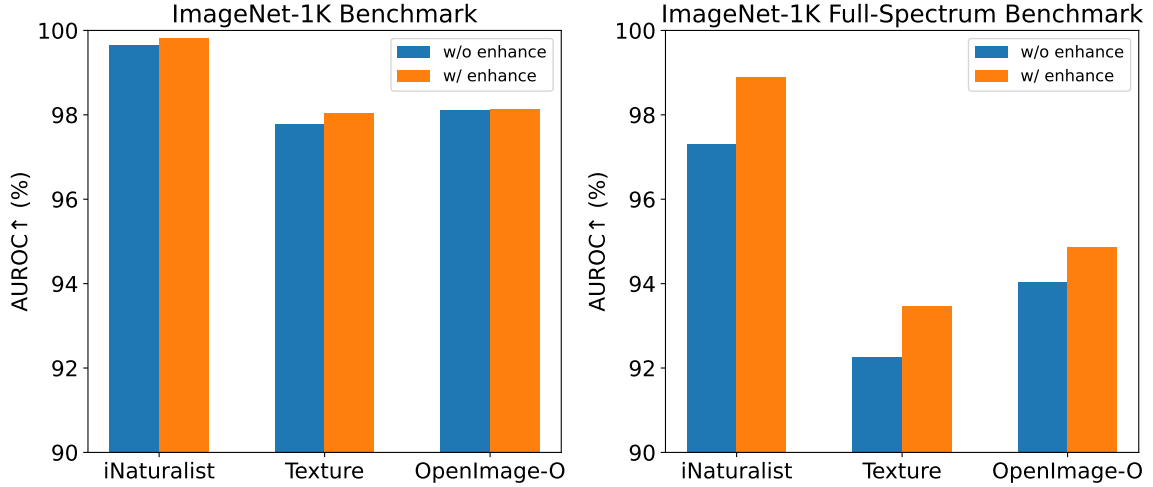| S-BG	Filter	PT	VPT	ImageNet OOD	Near-OOD
zero-shot					
×	×			93.39	80.79
✓	×			94.90	81.04
×	✓			94.28	73.92
✓	✓			95.51	72.52
few-shot					
×	×	✓	×	95.49	87.47
×	×	×	✓	94.60	76.06
✓	×	✓	×	96.63	87.96
×	✓	✓	×	96.39	83.24
✓	✓	✓	×	97.33	80.34
×	×	✓	✓	96.59	<u>91.25</u>
✓	×	✓	✓	<u>97.67</u>	**91.98**
×	✓	✓	✓	97.25	87.43
✓	✓	✓	✓	**97.77**	82.49

Figure 4: Performance comparison on the OpenOOD V1.5 benchmark Zhang et al. (2023); Yang et al. (2022). The evaluation contrasts the effectiveness of positive labels enhanced by original class labels against the baseline without enhancement.

features for identifying Near-OOD samples. Additionally, in the absence of filtering, our proposed Superclass-BG method, which employs a more refined ID semantic space to select negative labels, further enhances performance on Near-OOD datasets.

The Effectiveness of Enhancing Positive Labels with Class Labels. As shown in Figure 4, the

OpenOOD V1.5 ImageNet-1K benchmark, which does not account for covariate shift, demonstrates that positive labels enhanced by class labels consistently outperform their non-enhanced counterparts across various datasets, albeit with marginal differences. However, we observe a strong amplification of this performance gap when examining the full-spectrum benchmark, which incorporates covariate shift robustness. This amplification underscores the efficacy of enhancing positive labels, adapted to the target distribution, with class labels that possess strong generalization capabilities.

6. Conclusion

In this work, we present a novel zero-shot OOD detection approach that harnesses the power of LLMs and VLMs. Our method introduces a hierarchical strategy using superclass labels and background descriptions generated by LLMs to select more representative negative labels from a comprehensive set of unfiltered candidate labels. Combining this approach with prompt tuning and visual prompt tuning, the proposed few-shot method further pushes the boundaries of OOD detection, consistently outperforming current state-of-the-art methods across multiple benchmarks, including challenging full-spectrum benchmark. Moreover, our method demonstrates remarkable robustness to covariate shifts, addressing a critical challenge in real-world scenarios.

Limitations Our proposed method leverages the capabilities of LLMs and VLMs, which may introduce certain considerations. The semantic understanding provided by these models, while powerful, can potentially reflect biases present in their training data. Additionally, the effectiveness of candidate label filtering in our approach presents an area for extended research. While initial results are promising, comprehensive studies across various datasets and OOD detection tasks would provide valuable insights into optimal filtering strategies. These aspects offer opportunities for future work to further enhance the adaptability of the model and performance across diverse applications.

Societal Impact Advancements OOD detection have the potential to significantly enhance the reliability and safety of AI systems across critical domains such as healthcare, autonomous vehicles, and finance. Improved OOD detection could lead to more accurate diagnostic tools, safer autonomous systems, and robust fraud detection mechanisms. However, as these technologies evolve, it is crucial to address ethical considerations, including privacy, fairness, and transparency. Responsible development and deployment of OOD detection systems will be essential to maximize societal benefits while mitigating potential risks.

7. Acknowledgement

References

Anthropic. Claude 3.5 Sonnet. https://www.anthropic.com, 2024. (Jun 20 version) [Large language model].

Yichen Bai, Zongbo Han, Changqing Zhang, Bing Cao, Xiaoheng Jiang, and Qinghua Hu. ID-like Prompt Learning for Few-Shot Out-of-Distribution Detection. *arXiv preprint arXiv:2311.15243*, 2024.

Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? Fixing ImageNet Out-of-Distribution Detection Evaluation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 105–140, 2023.

Soravit Changpinyo, Piyush Kumar Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training to Recognize Long-Tail Visual Concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3557–3567, 2021.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing Textures in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3606–3613, 2014.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-Shot Out-of-Distribution Detection Based on the Pre-trained Model CLIP. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, pages 6568–6576, 2022.

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. doi: 10.23915/distill.00030.

Rui Huang, Andrew Geng, and Yixuan Li. On the Importance of Gradients for Detecting Distributional Shifts in the Wild. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 677–689, 2021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual Prompt Tuning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 709–727, 2022.

Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative Label Guided OOD Detection with Pretrained Vision-Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR, 2024.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7167–7177, 2018.

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 (CVPR)*. IEEE, 2024.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based Out-of-distribution Detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:21464–21475, 2020.

Fan Lu, Kai Zhu, Kecheng Zheng, Wei Zhai, and Yang Cao. Likelihood-aware semantic alignment for full-spectrum out-of-distribution detection. In *arXiv preprint arXiv:2312.01732*, 2023.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into Out-of-Distribution Detection with Vision-Language Representations. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:35087–35102, 2022.

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 (NeurIPS)*, 36:76298–76310, 2023.

Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Yueqian Lin, Qing Yu, Go Irie, Shafiq Joty, Yixuan Li, Hai Li, Ziwei Liu, Toshihiko Yamasaki, and Kiyoharu Aizawa. Generalized out-of-distribution detection and beyond in vision language model era: A survey. *arXiv preprint arXiv:2407.21794*, 2024.

Jun Nie, Yonggang Zhang, Zhen Fang, Tongliang Liu, Bo Han, and Xinmei Tian. Out-of-Distribution Detection with Negative Prompts. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR, 2024.

OpenAI. GPT-4o. https://chat.openai.com/chat, 2024. (July 18 version) [Large language model].

Jaewoo Park, Yoon Gyo Jung, and Andrew Beng Jin Teoh. Nearest neighbor guidance for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1686–1695, 2023a.

Sangha Park, Jisoo Mok, Dahuin Jung, Saehyung Lee, and Sungroh Yoon. On the powerfulness of textual outlier exposure for visual ood detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:51675–51687, 2023b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 8748–8763, 2021.

Chandramouli Shama Sastry and Sageev Oore. Detecting Out-of-Distribution Examples with Gram Matrices. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119, pages 8491–8501, 2020.

Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Dennis Ulmer, Lotta Meijerink, and Giovanni Cinà. Trust issues: Uncertainty estimation does not enable reliable ood detection on medical tabular data. In Emily Alsentzer, Matthew B. A. McDermott, Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy, and Stephanie L. Hyland, editors, *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 136 of *Proceedings of Machine Learning Research*, pages 341–354, 2020.

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8769–8778, 2018.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-Set Recognition: A Good Closed-Set Classifier is All You Need. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-Of-Distribution with Virtual-logit Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Hualiang Wang, Yi Li, Huifeng Yao, and X. Li. CLIPN for Zero-Shot OOD Detection: Teaching CLIP to Say No. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1802–1812, 2023.

Yixuan Wei, Han Hu, Zhenda Xie, Ze Liu, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Improving CLIP Fine-tuning Performance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5416–5426, 2023.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN Database: Large-Scale Scene Recognition from Abbey to Zoo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492, 2010.

Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-Spectrum Out-of-Distribution Detection. *arXiv preprint arXiv:2204.05306*, 2022.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection. *arXiv preprint arXiv:2306.09301*, 2023.

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, 2018.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16795–16804, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision (IJCV)*, 130(9):2337–2348, 2022b.