`ID-like` Prompt Learning for Few-Shot Out-of-Distribution Detection

Yichen Bai [1*], Zongbo Han [1*], Bing Cao [1,2†], Xiaoheng Jiang [3], Qinghua Hu [1,2], Changqing Zhang [1,2†]
[1]College of Intelligence and Computing, Tianjin University
[2]Tianjin Key Lab of Machine Learning
[3]School of Computer and Artificial Intelligence, Zhengzhou University
{ycfate, zongbo, caobing, huqinghua, zhangchangqing}@tju.edu.cn, {jiangxiaoheng}@zzu.edu.cn

Abstract

Out-of-distribution (OOD) detection methods often exploit auxiliary outliers to train model identifying OOD samples, especially discovering challenging outliers from auxiliary outliers dataset to improve OOD detection. However, they may still face limitations in effectively distinguishing between the most challenging OOD samples that are much like in-distribution (ID) data, i.e., `ID-like` samples. To this end, we propose a novel OOD detection framework that discovers `ID-like` outliers using CLIP [32] from the vicinity space of the ID samples, thus helping to identify these most challenging OOD samples. Then a prompt learning framework is proposed that utilizes the identified `ID-like` outliers to further leverage the capabilities of CLIP for OOD detection. Benefiting from the powerful CLIP, we only need a small number of ID samples to learn the prompts of the model without exposing other auxiliary outlier datasets. By focusing on the most challenging `ID-like` OOD samples and elegantly exploiting the capabilities of CLIP, our method achieves superior few-shot learning performance on various real-world image datasets (e.g., in 4-shot OOD detection on the ImageNet-1k dataset, our method reduces the average FPR95 by 12.16% and improves the average AUROC by 2.76%, compared to state-of-the-art methods). Code is available at https://github.com/ycfate/ID-like

1. Introduction

When deploying machine learning models in practical settings, it is possible to come across OOD samples that were not encountered during training. The risk of incorrect decisions rises when it comes to these OOD inputs, which could pose serious safety issues, particularly in applications like autonomous driving and medical diagnosis. The system needs to identify OOD samples in addition to performing well on ID samples in order to produce trustworthy predictions. OOD detection is therefore quite critical for safely deploying machine learning models in reality.

Figure 1. Hard OOD samples typically contain more features correlated to ID samples, i.e., they behave more `ID-like`.

Existing methods [9, 18, 21] usually focus on detecting OOD examples only using ID data in training to predict lower confidence [8, 27] or higher energy [22] for OOD samples. However, due to the lack of OOD information, these models struggle to be effective in OOD detection. Therefore, some studies [10, 22] suggest using auxiliary outliers to regularize the models and identity OOD samples. Chen et al. [1] and Ming et al. [29] suggested that selecting more challenging outlier samples can help the model learn a better decision boundary between ID and OOD. However, these auxiliary outliers usually contain limited challenging outliers. Furthermore, most of these methods require additional outlier data, which makes them ineffective when outlier datasets are unavailable. Recently, Du et al. [4] proposed to synthesize virtual outlier data in the feature space of ID data to construct outliers during training without additional data. This method shows strong efficacy in distinguishing between ID and OOD. However, there are two main limitations: i) it assumes that ID data in the feature space conforms to a class conditional Gaussian distribution, which does not always hold in the complex real-world applications [34]; ii) it requires numerous ID samples to construct a more accurate distribution of ID data, while obtaining a large number of ID samples is often costly. Accordingly, in this work, we focus on flexibly constructing challenging outliers with few-shot

*Equal contribution. †Corresponding author.

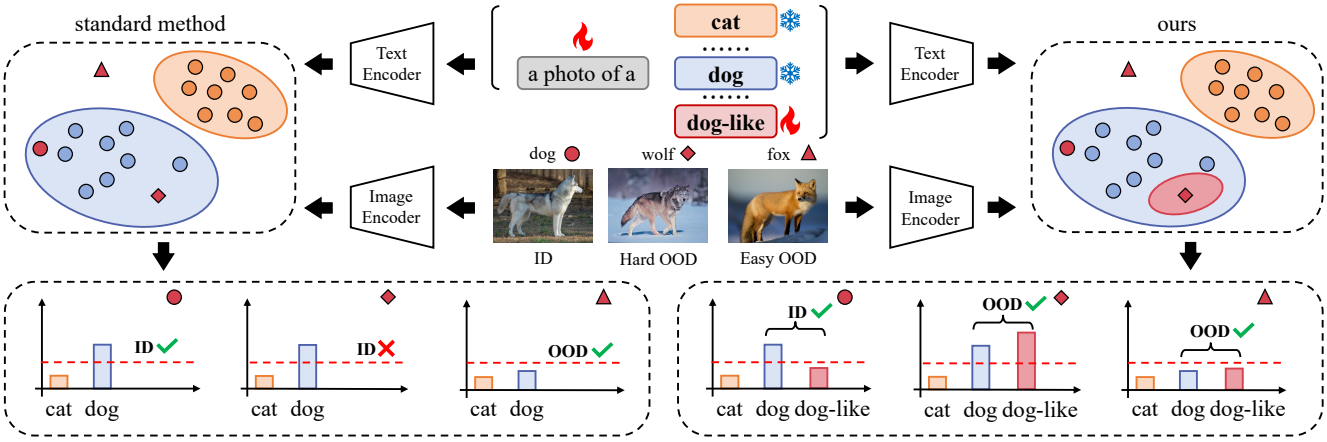


Figure 2. The standard method can only output the predicted probabilities of samples for each ID class. In contrast, our approach can automatically learn additional classes that are highly correlated but distinct from the ID classes, thereby effectively identifying challenging `ID-like` OOD samples. (Note that the dog-like prompt in the figure is learnable.)

ID samples to improve the identification of OOD samples.

In this paper, we first construct outliers highly correlated with ID data and introduce a novel `ID-like` prompts for OOD detection, thereby effectively identifying challenging OOD samples. We find that challenging OOD samples often behave highly correlated with ID data, exhibiting high visual or semantic similarity, e.g., the local feature of OOD being relevant to ID (as shown in Fig. 1). Since these `ID-like` features of OOD samples lead to erroneous predictions, a natural idea arises: extracting relevant features from ID samples to construct challenging OOD samples. To this end, we perform multiple samplings on vicinity space of ID samples. Among these samplings, those with lower similarity to the ID prompts are not classified as ID classes, even they contain the features correlated with ID classes. Therefore, these samples are naturally selected as challenging OOD samples. Differing from VOS [4] and NPOS [34], which synthesize virtual outliers in low-likelihood regions of the feature space, our method constructs outliers directly from the original images, enhancing the flexibility and interpretability.

Although we can construct challenging OOD samples, it is still challenging to effectively identify these OOD samples. As shown in the left part of Fig. 2, "*wolf*" represents a challenging OOD example of "*dog*" class. These images are similar to ID prompts, resulting in high classification probabilities and significant challenges in distinguishing between ID and OOD. We argue that relying solely on ID prompts is insufficient to address this issue. Therefore, we introduce additional prompts to enhance OOD identification. As shown in the right part of Fig. 2, we develop an additional prompt, termed "*dog-like*", which is similar to the prompt of "*dog*". If we can increase the similarity between the "*dog-like*" prompt and OOD samples that are highly correlated with "*dog*", the model would recognize dogs through the "*dog*" prompt and identify challenging OOD samples (including "*wolf*") through the "*dog-like*" prompt. Specifically, we align the additional prompts with these constructed challenging OOD, creating OOD prompts similar to ID prompts to effectively identify challenging OOD samples. Extensive experiments demonstrate that our method achieves superior few-shot OOD detection performance on a wide variety of real-world tasks. Compared to methods [4, 34] that require a large amount of data during training, our method significantly reduces the average FPR95 score from 38.24% to 24.08% and improves the average AUROC from 91.60% to 94.70% even using only one image for each class. We summarize our main contributions as follows:

- We propose a novel framework without additional training to automatically explore `ID-like` OOD samples in the vicinity space of ID samples by leveraging CLIP, which assists the model in effectively identifying challenging OOD samples correlated to the ID.
- By exploiting the capacity of a pre-trained visual-language model, an `ID-like` prompt learning method is proposed to identify the most challenging OOD samples, which behave `ID-like` yet are distinct.
- We validated our method on several real-world large-scale datasets, and the results show that our method achieved impressive performance, with an average AUROC of 96.66% in 4-shot OOD detection on ImageNet-1K. Additional ablation experiments are also conducted to demonstrate the effectiveness of the designed approach.

2. Related Work

OOD Detection with Pre-trained Vision-language Models. Hendrycks and Gimpel [9] established a baseline for OOD detection using the maximum softmax probability (MSP). Subsequent works have explored OOD detection via ODIN scores [12, 21] and Mahalanobis scores [18]. Fort et al. [6] first extended the OOD detection task to pre-trained vision-

language models. Esmaeilpour et al. [5] enhanced the OOD detection performance of pre-trained vision-language models by generating additional negative labels to construct negative prompts. Recently, Ming et al. [28] extended MSP to pre-trained vision-language models and explored the impact of softmax and temperature scaling on OOD detection. CLIPN [38] fine-tuned CLIP to enable it to output negative prompts to assess the probability of a concept not being present in the image.

Contrastive Vision-language Models. Compared to traditional multi-modal learning models, recent large-scale pre-trained vision-language models [14, 15, 32, 40] have achieved great progress in various downstream tasks. For instance, CLIP [32] and ALIGN [14] leverages contrastive loss, such as InfoNCE loss [35], to learn aligned representations of images and text. The representation distance of matching image-text pairs becomes closer while those of non-matching pairs are farther apart. Specifically, these methods employ a straightforward dual-stream architecture comprising an image encoder and a text encoder, which maps image and text features into a shared space for similarity computation. The performance of CLIP [32] and ALIGN [14] both benefits from a large number of text image-pairs data. The decision risk of multi-modal is also the focus of current research [7, 25, 41].

CLIP-based Prompt Learning. In Natural Language Processing (NLP), Petroni et al. [31] conceptualized prompting as akin to a fill-in-the-blanks task. The core idea is to induce a pre-trained language model to generate answers given cloze-style prompts. However, it relies heavily on a well-designed prompt. To avoid manually designing a large number of prompts, some studies [19, 20] introduce prompt tuning as a solution. Prompt tuning learns the prompt from downstream data in the continual input embedding space, which presents a parameter-efficient way of fine-tuning foundation models. Despite the widespread adoption of prompt learning within NLP, its exploration within the visual domain remains limited. Recently, CoOp [44] and CoCoOp [43] apply prompt tuning to CLIP [32], which tune prompts via minimizing the classification loss on the target task and effectively improves CLIP's performance on the corresponding downstream tasks. Plenty of studies [24, 26, 30, 33] leverage prompt learning based on CLIP to enhance performance across various downstream tasks.

3. Method

3.1. Preliminaries

Zero-shot classification with CLIP. CLIP consists of a text encoder $\mathcal{T} : t \to \mathbb{R}^d$ and an image encoder $\mathcal{I} : x \to \mathbb{R}^d$, which are used to obtain the feature vectors of text t and image x, respectively. When performing a classification task, assuming the known label set $\mathcal{Y} = \{y_1, y_2, ..., y_K\}$, we can construct a collection of concept vectors $\mathcal{T}(t_k), k \in \{1, 2, ..., K\}$, where t_k is the text prompt "a photo of a $\langle y_k \rangle$" for a label y_k. We denote the features of text and images as $h = \mathcal{T}(t)$ and $z = \mathcal{I}(x)$, respectively. We first obtain the similarity of image features relative to all text features $s_k(x) = sim(h_k, z) = sim(\mathcal{T}(t_k), \mathcal{I}(x))$, where $sim(\cdot, \cdot)$ denotes the cosine similarity. The predicted probability p_k corresponding to y_k on x can be expressed as

$$p_k(x; \mathcal{Y}, \mathcal{T}, \mathcal{I}) = \frac{e^{s_k(x)/\tau}}{\Sigma_{k=1}^{K} e^{s_k(x)/\tau}}, \tag{1}$$

where τ is the temperature of the softmax function.

Prompt Learning. To further improve the performance of CLIP on few-shot classification, CoOp [44] constructs a learnable tensor on the embedding layer of the text. Specifically, CoOp initializes the learnable tensor of prompt as $t = [V]_1[V]_2...[V]_L[CLASS]$, where L is the token length, $[V]_l (l \in \{1, 2, ..., L\})$ is a learnable vector with the same dimension as the word embedding. Then a loss function e.g., cross-entropy loss, can be constructed to optimize the learnable prompt according to classification probability of few-shot examples.

OOD Detection. The OOD detection usually constructs an OOD detector denoted as $F(x)$, i.e., a binary classifier

$$F(x) = \begin{cases} ID, & S(x) \geq \gamma \\ OOD, & S(x) < \gamma, \end{cases} \tag{2}$$

where $S(x)$ is a score function in OOD detection task, and γ is a threshold to decide whether the samples belong to ID or OOD. For example, Hendrycks and Gimpel [9] and Liu et al. [22] use the maximum classification probability of softmax and energy as the score function $S(x)$, respectively.

3.2. ID-like Prompt Learning

In this paper, we introduce a novel model for few-shot OOD detection, which employs cropping and the CLIP model to create challenging outliers to improve the OOD detection ability. Additionally, we employ prompt learning to acquire `ID-like` OOD prompts. As shown in Fig. 3, our framework consists of two main components: **(1) Constructing outliers from ID samples:** The training set with N samples is represented as $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$. To suffiiently explore vicinal space of training samples, we perform multiple random cropping on each ID sample x_i to obtain the set $X_i^{crop} = \{x_{i,1}^{crop}, x_{i,2}^{crop}, ..., x_{i,M}^{crop}\}$, where M is the number of random cropping iterations. Concurrently, we create corresponding class description text t_k using predefined templates, such as "a photo of a $\langle y_k \rangle$", where $y_k \in \mathcal{Y}$ represents the corresponding class name. Subsequently, leveraging the pre-trained CLIP model, we calculate the cosine similarity between the samples in set X_i^{crop} and the descriptions t_k. Based on the strength of cosine similarity, we then

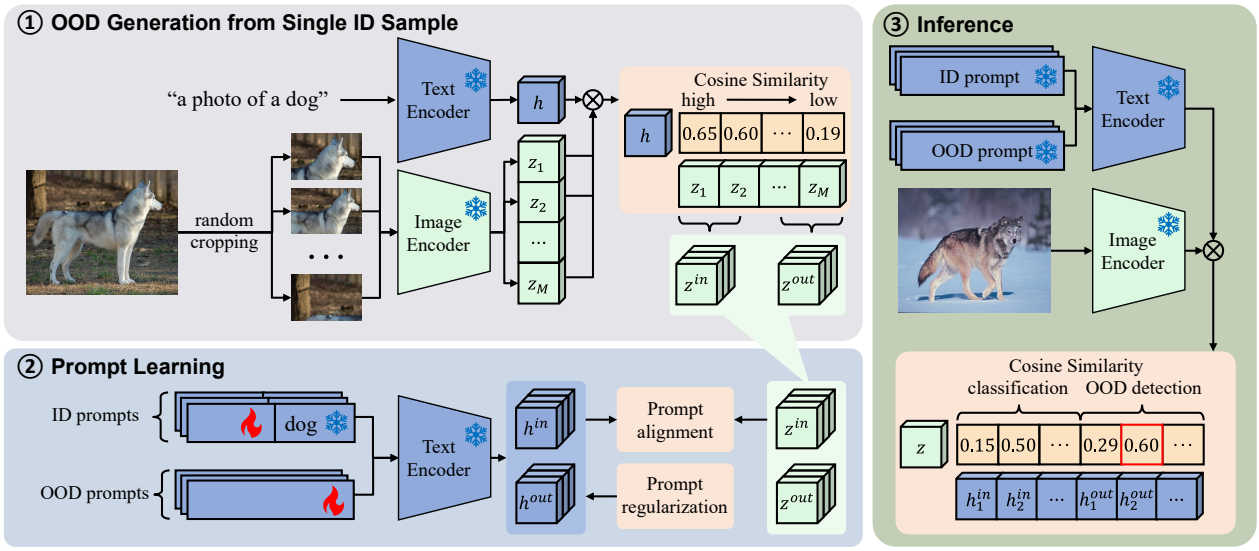


Figure 3. Overview of our method. We conduct multiple random cropping on ID sample and filter them based on their cosine similarity with established ID zero-shot prompts, thereby generating both ID and OOD data. Subsequently, prompt learning is employed to acquire prompts corresponding to the ID and `ID-like` OOD samples. The obtained prompts can effectively identify OOD samples in the inference stage.

respectively extract ID and OOD samples from the highest and lowest similarity segments, defining them as $X_i^{in} = \{x_{i,1}^{in}, x_{i,2}^{in}, ..., x_{i,Q}^{in}\}$ and $X_i^{out} = \{x_{i,1}^{out}, x_{i,2}^{out}, ..., x_{i,Q}^{out}\}$, where Q is a user-defined hyperparameter. In the end, we obtain the $D^{in} = \{(x_{1,1}^{in}, y_1), (x_{1,2}^{in}, y_1), ..., (x_{N,Q}^{in}, y_N)\}$ and $D^{out} = \{x_{1,1}^{out}, x_{1,2}^{out}, ..., x_{N,Q}^{out}\}$, constructed from all the ID samples. **(2) Prompt learning:** We initialize a learnable prompt for each class, forming the ID prompts set $T^{in} = \{t_1^{in}, t_2^{in}, ..., t_K^{in}\}$, and initialize an additional set of OOD prompts, $T^{out} = \{t_1^{out}, t_2^{out}, ..., t_C^{out}\}$, where C is the number of OOD prompts. Given the limited scope covered by individual descriptions, we introduce multiple OOD descriptions to enhance the coverage. Similar to CoOp [44], we initialize embeddings for these text descriptions randomly and then optimize them using a loss function proposed in the Sec. 3.3.

3.3. Loss Functions

During training, we can obtain ID and OOD data, denoted as D^{in} and D^{out}, based on the algorithm mentioned in the previous section. We optimize prompts through a loss function that consists of three terms.

In-distribution loss. To ensure classification performance on the in-distribution data, we utilize a standard cross-entropy loss function, which measures the divergence between the predicted label probabilities and ground truth labels for ID samples. Formally, the ID cross-entropy loss $\mathcal{L}_{in}$ is defined as:

$$\mathcal{L}_{in} = \mathbb{E}_{(x,y)\sim D^{in}}[-\log \frac{e^{s_*/\tau}}{\Sigma_{k=1}^{K} e^{s_k^{in}/\tau} + \Sigma_{c=1}^{C} e^{s_c^{out}/\tau}}], \quad (3)$$

where $s_* = sim(\mathcal{T}(t_*), \mathcal{I}(x))$, $s_k^{in} = sim(\mathcal{T}(t_k^{in}), \mathcal{I}(x))$, $s_c^{out} = sim(\mathcal{T}(t_c^{out}), \mathcal{I}(x))$, t_* represents the features of textual description of ground-truth label y_* corresponding to x, $t_k^{in} \in T^{in}$ and $t_c^{out} \in T^{out}$.

Out-of-distribution loss. To align OOD prompts with outliers, we introduce the OOD loss. It is important to note that in an ideal scenario, each category would have an ID prompt and an OOD prompt. However, to conserve computational resources and enhance training efficiency, we have fixed the number of OOD prompts at 100. Consequently, when there are insufficient OOD prompts to establish a one-to-one correspondence with the ID categories, we maximize the holistic similarity between the OOD prompts and outliers. To accomplish this, we propose the following loss $\mathcal{L}_{out}$:

$$\mathcal{L}_{out} = \mathbb{E}_{x\sim D^{out}}[-\log \frac{\Sigma_{c=1}^{C} e^{s_c^{out}/\tau}}{\Sigma_{k=1}^{K} e^{s_k^{in}/\tau} + \Sigma_{c=1}^{C} e^{s_c^{out}/\tau}}]. \quad (4)$$

Additionally, we observed that implementing $\mathcal{L}_{out}$ during training in the following form is more conducive to optimizing prompts:

$$\mathcal{L}_{out} = \mathbb{E}_{x\sim D^{out}}[\log \frac{\Sigma_{k=1}^{K} e^{s_k^{in}/\tau}}{\Sigma_{k=1}^{K} e^{s_k^{in}/\tau} + \Sigma_{c=1}^{C} e^{s_c^{out}/\tau}}]. \quad (5)$$

Although their optimization goals are similar, the former tends to maximize the similarity between OOD prompts and outliers, while the latter tends to minimize the similarity between ID prompts and outliers, resulting in slight differences during training.

Diversity regularization. Since all OOD prompts are randomly initialized and optimized under the same objective

shown in Eq. 4 , there arises a risk of excessive similarity between OOD prompts. Similar OOD prompts may lead to a reduction in the number of detectable OOD classes. To mitigate this issue and ensure the diversity of OOD prompts, we introduce an additional loss $\mathcal{L}_{div}$ that explicitly maximizes the dissimilarity between prompts:

$$\mathcal{L}_{div} = \frac{\Sigma_{c=1}^{C-1}\Sigma_{j=c+1}^{C} sim(h_c^{out}, h_j^{out})}{C(C-1)/2}, \tag{6}$$

where $h_c^{out} = \mathcal{T}(t_c^{out}), h_j^{out} = \mathcal{T}(t_j^{out})$. $t_c^{out}, t_j^{out} \in T^{out}$ denote the c-th and j-th prompt in the OOD prompts. $sim(\cdot,\cdot)$ denotes the cosine similarity.

The overall loss function with balanced hyperparameter λ_{out} and λ_{div} is:

$$\mathcal{L} = \mathcal{L}_{in} + \lambda_{out}\mathcal{L}_{out} + \lambda_{div}\mathcal{L}_{div}. \tag{7}$$

Inference. When performing the classification task, we utilize the same classification method as CLIP, relying solely on the ID prompts for classification. For OOD detection, we define the scoring function as:

$$S(x) = \frac{\Sigma_{k=1}^{K} e^{s_k^{in}/\tau}}{\Sigma_{k=1}^{K} e^{s_k^{in}/\tau} + \Sigma_{c=1}^{C} e^{s_c^{out}/\tau}}. \tag{8}$$

4. Experiments

4.1. Experimental Setup

Datasets. Different from previous OOD detection tasks, we mainly aim to achieve OOD detection in the open-world setting, so we do not choose some toy (e.g., low-resolution) datasets, such as CIFAR [16] and MNIST [17]. In our work, we follow the settings of MOS [13] and MCM [28], which use ImageNet-1k [3] as ID data and a subset of iNaturalist [11], PLACES [42] and TEXTURE [2] as OOD data. SUN [39] is tested independently as a specific OOD dataset. Following MOS [13], these OOD data are randomly selected from the categories that do not overlap with ImageNet-1k [3]. Furthermore, some of the ablation experiments are conducted using ImageNet-100 as the ID data. This dataset follows the configuration of MCM [28] which selects 100 classes from ImageNet-1k as the ID data.

Pre-trained Model. In our experiments, we employ CLIP-B/16 as the pre-trained model for OOD prompt learning. Concretely, we choose CLIP-B/16, which consists of a ViT-B/16 Transformer as the image encoder and a self-attention Transformer as the text encoder. CLIP is one of the most popular pre-trained models, which learns from large-scale image-text datasets to create a shared embedding space where images and their associated text descriptions are represented coherently. By using contrastive learning, CLIP ensures similar image-text pairs closer together and dissimilar pairs farther apart, allowing it to understand the semantic relationships between visuals and language. In our experiment, we keep all the network parameters of CLIP fixed, including both the image encoder and the text encoder. We only update the embedding layer on the text input side, following the approach of prompts learning.

Implementation Details. For few-shot training, it is necessary to randomly select a certain number of samples from each class in the complete training data to form the training set. For example, we randomly choose one (one-shot) or four samples (four-shot) from each class in ImageNet-1k. When constructing ID and OOD data, we conduct M (256 in our experiment) random crops on each sample, and choose the top Q (32 in our experiment) and bottom Q samples based on the similarity to the manually prompts. For ID prompts, there is only one learnable prompt per class, and class name information is retained. For OOD prompts, we set their total number to C (100 in our experiment), and class name information is not retained. We set λ_1 to 0.3, λ_2 to 0.2, and use AdamW [23] as the optimizer. Other hyperparameters settings are as follows: training epoch = 3, learning rate = 0.005, batch size = 1, and token length L = 16.

Competing Methods. We compare our method to several OOD detection works, including fully supervised, zero-shot, and few-shot approaches. For fully supervised methods, we follow the same setting as NPOS [34], and compare with MSP [9], Fort/MSP [6], Energy score [22], ODIN score [21], VOS [4], NPOS [34], and CLIPN [38]. For zero-shot methods, we select MCM [28] for comparison. For few-shot methods, we compare with CoOp [44] and LoCoOp [30]. For fairness, all methods are trained using the same pre-trained model (CLIP/ViT-B/16), and we reproduce some results from NPOS [34] and LoCoOp [30].

Evaluation Metrics. We adopt the following evaluation metrics that are commonly used in OOD detection: (1) the false positive rate of OOD examples when the true positive rate of in-distribution examples is at 95% (FPR95); (2) the area under the receiver operating characteristic curve (AUROC); (3) ID classification accuracy (ID ACC).

4.2. Results

Table 1 shows our main comparison results, which demonstrate that using our method can achieve better OOD detection performance, outperforming most comparisons. More importantly, our method still has good results in 1-shot setting even compared to those methods that require full data. Specifically, in the 4-shot setting, we obtain 26.08% in terms of FPR95 and 94.36% in terms of AUROC on average, implying a reduction of 12.16% and an improvement of 2.76%, respectively compared to the best-performing method under the same settings. Fig. 4 shows a comparison between our method and MCM on the iNaturalist dataset. Our approach demonstrates superior performance with a significantly larger discrepancy between ID and OOD. This sug-

Table 1. OOD detection performance for ImageNet-1k [3] as ID. ViT-B/16 is an image encoder for CLIP-B/16, ViT-B$^+$/16 uses the text encoder of CLIP-B/16 for initialization, CLIP-B$^+$/16 uses an additional text encoder for training.

Method	Backbone	OOD Dataset							
		iNaturalist		Places		Texture		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Full/Sub Data Fine-tune									
MSP [9]	CLIP-B/16	40.89	88.63	67.90	80.14	64.96	78.16	57.92	82.31
Energy [22]	CLIP-B/16	29.75	94.68	56.40	85.60	51.35	88.00	45.83	89.43
ODIN [21]	CLIP-B/16	30.22	94.65	55.06	85.54	51.67	87.85	45.65	89.35
Fort/MSP [6]	ViT-B/16	54.05	87.43	72.98	78.03	68.85	79.06	65.29	81.51
VOS [4]	ViT-B/16	31.65	94.53	41.62	90.23	56.67	86.74	43.31	90.50
NPOS [34]	ViT-B$^+$/16	**16.58**	**96.19**	45.27	89.44	46.12	88.80	35.99	91.48
CLIPN [38]	CLIP-B$^+$/16	23.94	95.27	**33.45**	**92.28**	**40.83**	**90.93**	**32.74**	**92.83**
Zero-shot									
MCM [28]	CLIP-B/16	30.91	94.61	44.69	89.77	57.77	86.11	44.46	90.16
One-shot									
CoOp [44]	CLIP-B/16	43.38	91.26	46.68	89.09	50.64	87.83	46.90	89.39
LoCoOp [30]	CLIP-B/16	38.49	92.49	**39.23**	91.07	49.25	89.13	42.32	90.90
Ours	CLIP-B/16	**14.57**	**97.35**	41.74	**91.15**	**26.77**	**94.38**	**27.69**	**94.29**
Four-shot									
CoOp [44]	CLIP-B/16	35.36	92.60	45.38	89.15	43.74	89.68	41.49	90.48
LoCoOp [30]	CLIP-B/16	29.45	93.93	**41.13**	90.32	44.15	90.54	38.24	91.60
Ours	CLIP-B/16	**8.98**	**98.19**	44.00	**90.57**	**25.27**	**94.32**	**26.08**	**94.36**

Table 2. ID accuracy on ImageNet-1k [3].

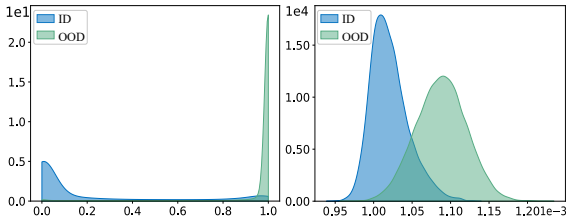
Method	ID acc	Full Data	Zero-shot	One-shot
VOS [4]	79.64	✓		
NPOS [34]	79.42	✓		
MCM [28]	67.01		✓	
CoOp [44]	66.23			✓
LoCoOp [30]	66.88			✓
Ours	68.28			✓

Table 3. OOD detection performance for ImageNet-1k as ID, SUN [39] as OOD.

Method	SUN			
	One-shot		Four-shot	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CoOp [44]	38.53	91.95	37.06	92.27
LoCoOp [30]	33.27	93.67	33.06	93.24
Ours	44.02	91.08	42.03	91.64

gests that MCM is more sensitive to threshold when distinguishing between ID and OOD, whereas our method allows a more intuitive distinction between ID and OOD. Furthermore, as shown in Table 2, our method outperforms other few-shot methods, achieving superior classification results on ID data at 68.28%.

Discussion on the SUN dataset. We also conduct an evaluation on the SUN dataset [39] as OOD data, and the results are shown in Table 3. The results indicate that our method performs not well on the SUN dataset. To investigate the reasons, we conduct a detailed examination of the SUN dataset. Afterward, we find that some samples belong actually ID classes (as shown in Fig. 7), but they are labeled as OOD. To investigate whether the observed case is a prevalent phenomenon in the SUN dataset, we conduct a more detailed analysis. We randomly select 400 samples from the SUN dataset and observe whether they belong to the ID category. We find that among these samples, 145 belong to the ID category. This investigation implies the following fact: the SUN dataset may require more detailed annotation and filtering to be suitable as OOD data for testing the OOD detection performance (Places [42] might also have similar issues, but due to the space limitation, we leave this in future).



(a) Ours (b) MCM

Figure 4. Density of the obtained ID and OOD score with the proposed method (left) and MCM [28] (right).

Discussion of performance differences. We briefly ana-

Table 4. Ablation study by constructing different outlier training data. The experimental results show that the proposed method of constructing outlier has achieved significant improvements.

Outlier (train)	OOD Dataset									
	iNaturalist		Places		Texture		SUN		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Related outlier										
iNaturalist [11]	1.57	99.62	41.57	90.11	81.29	69.62	40.05	89.26	41.12	87.15
Places [42]	30.68	95.17	10.64	97.91	75.57	77.19	**22.62**	**95.14**	34.88	91.35
Texture [2]	15.24	97.27	22.07	95.60	33.51	93.03	41.45	90.40	28.07	94.07
SUN [39]	31.14	94.72	24.04	94.70	88.21	71.02	11.06	97.65	38.61	89.52
Unrelated outlier										
CUB [37]	60.88	89.63	59.40	87.34	48.63	89.00	78.39	78.62	61.83	86.15
Gaussian Noise	52.18	91.11	80.53	75.13	33.95	91.40	61.94	85.09	57.15	85.68
Our outlier										
One-shot	10.82	97.84	26.60	95.07	20.41	96.16	41.94	91.43	24.94	95.12
Four-shot	**1.62**	**99.60**	**20.81**	**96.05**	**13.55**	**97.30**	29.53	93.70	**16.38**	**96.66**

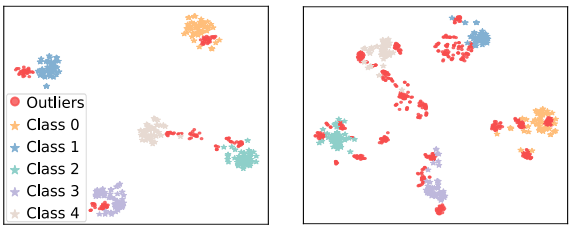
lyze the performance differences of our method across different OOD datasets. For example, our method exhibits significant performance improvement on iNaturalist [11] and Texture [2] datasets. The possible reason is that the cropped samples are more likely to contain image textures, plants, and animals in the background, making them correlated with ID classes. iNaturalist consists of various types of plants and animals, and Texture consists of natural textures. Therefore, our approach exhibits a greater improvement on these two datasets. In contrast, SUN [39] and Places [42] primarily consist of scene-based data, typically lacking specific objects (e.g., containing multiple objects). Therefore, our approach shows limited performance improvement on these two datasets.

4.3. Ablation Study

The effectiveness of our outliers. To show the effectiveness of the outliers constructed, we conduct the following experiments. Specifically, we using different additional outliers in training to investigate the improvement of our constructed outliers. Furthermore, we categorize the auxiliary outliers into "Related outlier", "Unrelated outlier" and "Our outlier". Concretely, "Related outlier" are selected from the challenging OOD samples (those with high MSP scores [9]). "Unrelated outlier" are selected from OOD datasets that are unrelated to the ID data. The results are shown in Table 4. Partial experimental results are in gray because it is unfair to compare them, since the outliers used during training and the OOD in testing come from the same distribution. Firstly, it is observed that models trained with our generated outliers outperform those trained with other outliers. This indicates the effectiveness of the outliers generated by our method. Secondly, we observe that the performance of models trained with "Related outlier" is generally better than those trained with "Unrelated outlier". This supports that outliers related to the ID can indeed help the model in learning a better decision boundary between the ID and OOD. The overall result strongly validates our `ID-like` outliers are quite effective and reasonable.

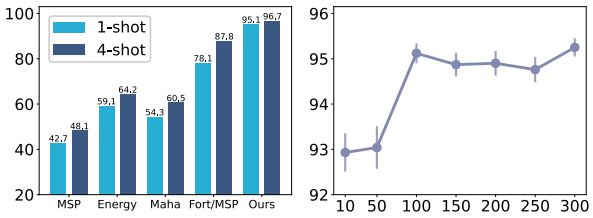
Furthermore, we utilize t-SNE [36] for visualization to illustrate the correlation between the outliers generated by our method and the ID samples. We employ a number of ID samples along with the outliers constructed based on the few-shot setting for visualization. For example, under the 1-shot setting, we use only one sample form each ID class to construct outliers, while the ID samples consist of a large number of samples form different ID classes. As shown in Fig. 5, the results show that even the generated outliers are from a quite small number of ID samples, they can also be correlated with the majority of ID samples. Moreover, with the increase of ID samples used in constructing outliers, both the number and the diversity of the `ID-like` outliers also increase.

The effectiveness of prompt learning. To show the advantages of prompt learning under few-shot setting, we train various models with our generated outliers (ImageNet-100 as ID), including fine-tuning the full model [9, 22], fine-tuning the last layer [6], training free [18], and prompt learning (ours). As shown in Fig. 6 (a), the results show that fine-tuning the full model performs worse in both 1-shot and 4-shot settings. The main reason is that these methods typically require abundant data for fine-tuning the model. Moreover, the better performance of all methods under 4-shot over 1-shot settings also validates this. Fort/MSP [6], which fine-tunes only the last layer of the model, performs better than fine-tuning the full model. This is because it preserves the majority of the model's prior knowledge, thereby reducing the dependence on the quantity of training data. However, it only utilizes the image encoder and does not leverage the pre-trained model's prior knowledge in text, thereby limiting the



(a) 1-shot (b) 4-shot

Figure 5. The obtained representations visualization of `ID-like` OOD samples and ID samples under 1-shot and 4-shot settings. The representation of the obtained `ID-like` OOD sample is close to the ID sample.



(a) AUROC of different methods under the few-shot setting (b)AUROC of different number of OOD prompts

Figure 6. Left: Performance in terms of AUROC of different methods trained on our constructed `ID-like` OOD dataset. Right: Performance in terms of AUROC at different number of OOD prompts during training.

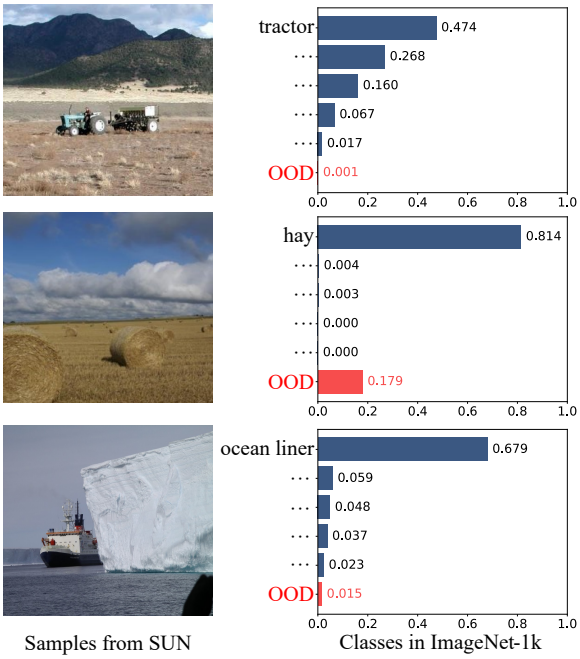


Figure 7. Left: Samples from the SUN [39] dataset that may be semantically identical to classes in ImageNet-1K [3], yet these samples are still considered as OOD during evaluation, which may result in a performance reduction. Right: Ours prediction results, including OOD scores.

performance of CLIP. Lee et al. [18] does not fine-tune the model, but it only utilizes the image encoder without fully leveraging CLIP's prior knowledge, limiting its performance. In contrast, prompt learning often utilizes both the image encoder and text encoder, leveraging the full prior knowledge of the pre-trained model. Therefore, for the limited amount of ID data, the performance is significant superior compared to existing methods.

The effectiveness of different quantities of `ID-like` prompts (OOD prompts). We test the impact of different quantities of OOD prompts for OOD detection performance. We set different values for C (10, 50, 100, 150, 200, 250, 300) and train the model using ImageNet-100 as ID data. The results are shown in Fig. 6 (b). The results demonstrate that as the number of OOD prompts increases, the OOD detection performance is also improved and tends to be stable. The underlying reason is that the expressive capacity is highly related to the number of prompts. Therefore, when ID data is complex, more prompts are required to characterize OOD samples related to the ID samples.

5. Conclusion

In this work, we propose a novel few-shot prompt learning method for out-of-distribution detection using pre-trained visual-language models. Our method introduces ID-like prompts and constructs outliers highly correlated with ID data from the training samples. By aligning ID-like prompts with the constructed outliers, we explore ID-like regions within the text feature space that are highly correlated with ID but do not belong to the ID. Our method elegantly addresses the key limitations in previous OOD detection methods, i.e., the challenge of constructing challenging outliers without auxiliary outliers and with a limited number of ID samples. Additionally, the introduction of the ID-like prompt provides a more effective way for the model to identify challenging OOD data. In challenging real-world OOD detection tasks, our method outperforms existing approaches. We conducted various ablation experiments to demonstrate the effectiveness of our approach. We hope that our work could inspire more future research on few-shot OOD detection based on prompts learning. We also hope to discover more interpretable ways to construct hard OOD in the future work.

6. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (Grant No.62376193, No.62106171, No.62172371, No.U21B2037, No.61976151). The authors gratefully acknowledge the support of MindSpore and CAAI-CANN Open Fund, developed on Openl Community. Thanks for the support provided by OpenI Community (https://openi.pcl.ac.cn). The authors also appreciate the suggestions from CVPR anonymous peer reviewers.

References

[1] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, 2021.

[2] 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*, 2014.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009.

[4] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. VOS: learning what you don't know by virtual outlier synthesis. In *International Conference on Learning Representations*, 2022.

[5] 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*, 2022.

[6] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2021.

[7] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2021.

[8] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[9] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

[10] Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

[11] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[12] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized ODIN: detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[13] 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*, 2021.

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021.

[15] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3d-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10984–10994, 2023.

[16] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

[18] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 2018.

[19] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Empirical Methods in Natural Language Processing*, 2021.

[20] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Annual Meeting of the Association for Computational Linguistics*, 2021.

[21] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

[22] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2020.

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[24] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[25] Huan Ma, Qingyang Zhang, Changqing Zhang, Bingzhe Wu, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. Calibrating multimodal learning. In *International Conference on Machine Learning*, 2023.

[26] Huan Ma, Yan Zhu, Changqing Zhang, Peilin Zhao, Baoyuan Wu, Long-Kai Huang, Qinghua Hu, and Bingzhe Wu. Invariant test-time adaptation for vision-language model generalization. In *arXiv*, 2024.

[27] Alexander Meinke and Matthias Hein. Towards neural networks that provably know when they don't know. In *International Conference on Learning Representations*, 2020.

[28] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems*, 2022.

[29] Yifei Ming, Ying Fan, and Yixuan Li. POEM: out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning*, 2022.

[30] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt

learning. In *Advances in Neural Information Processing Systems*, 2023.

[31] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. Language models as knowledge bases? In *Empirical Methods in Natural Language Processing*, 2019.

[32] 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 *International Conference on Machine Learning*, 2021.

[33] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, 2022.

[34] Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *International Conference on Learning Representations*, 2023.

[35] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[36] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[37] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[38] Hualiang Wang, Yi Li, Huifeng 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*, 2023.

[39] 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*, 2010.

[40] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022.

[41] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International Conference on Machine Learning*, 2023.

[42] Bolei Zhou, Àgata 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 (TPAMI)*, 40(6):1452–1464, 2018.

[43] 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*, 2022.

[44] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.