Knowledge Regularized Negative Feature Tuning of Vision-Language Models for Out-of-Distribution Detection

Wenjie Zhu*
Hong Kong Polytechnic University
Hong Kong, China
Eastern Institute of Technology
Ningbo, China
22040319r@connect.polyu.hk

Yabin Zhang*
Stanford University
Stanford, USA
yabin@stanford.edu

Xin Jin
Ningbo Institute of Digital Twin,
Eastern Institute of Technology
Ningbo, China
jinxin@eitech.edu.cn

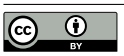
Wenjun Zeng†
Ningbo Institute of Digital Twin,
Eastern Institute of Technology
Ningbo, China
wzeng-vp@eitech.edu.cn

Lei Zhang†
Hong Kong Polytechnic University
Hong Kong, China
cslzhang@comp.polyu.edu.hk

Abstract

Out-of-distribution (OOD) detection is crucial for building reliable machine learning models. Although negative prompt tuning has enhanced the OOD detection capabilities of vision-language models, these tuned models often suffer from reduced generalization performance on unseen classes and styles. To address this challenge, we propose a novel method called Knowledge Regularized Negative Feature Tuning (KR-NFT), which integrates an innovative adaptation architecture termed Negative Feature Tuning (NFT) and a corresponding knowledge-regularization (KR) optimization strategy. Specifically, NFT applies distribution-aware transformations to pre-trained text features, effectively separating positive and negative features into distinct spaces. This separation maximizes the distinction between in-distribution (ID) and OOD images. Additionally, we introduce image-conditional learnable factors through a lightweight meta-network, enabling dynamic adaptation to individual images and mitigating sensitivity to class and style shifts. Compared to traditional negative prompt tuning, NFT demonstrates superior efficiency and scalability. To optimize this adaptation architecture, the KR optimization strategy is designed to enhance the discrimination between ID and OOD sets while mitigating pre-trained knowledge forgetting. This enhances OOD detection performance on trained ID classes while simultaneously improving OOD detection on unseen ID datasets. Notably, when trained with few-shot samples from ImageNet dataset, KR-NFT not only improves ID classification accuracy and OOD detection but also significantly reduces the FPR95 by 5.44% under an unexplored generalization setting with unseen ID categories. Codes can be found at https://github.com/ZhuWenjie98/KRNFT.

*Both authors contributed equally to this research.
†Corresponding author.



MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755120

CCS Concepts

• **Computing methodologies** → **Scene anomaly detection**.

Keywords

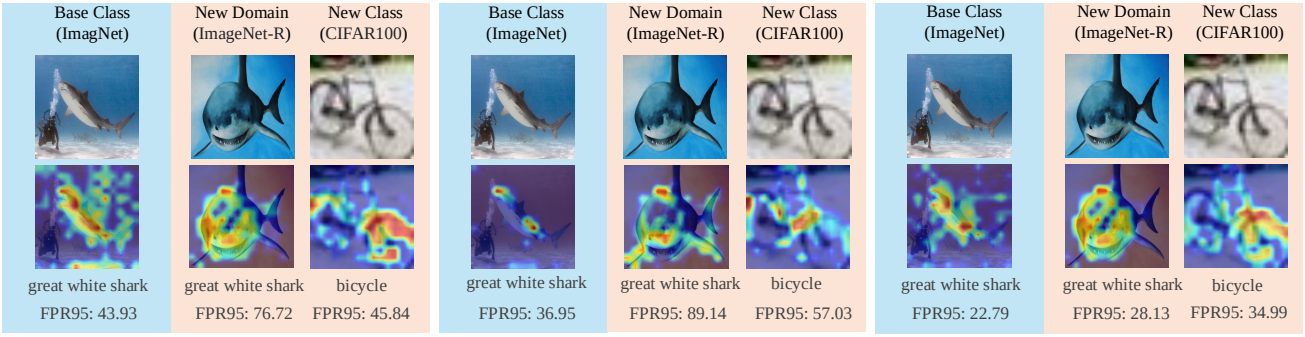
Out-of-Distribution Detection, Negative Feature Tuning, Knowledge Regularization

ACM Reference Format:
Wenjie Zhu, Yabin Zhang, Xin Jin, Wenjun Zeng, and Lei Zhang. 2025. Knowledge Regularized Negative Feature Tuning of Vision-Language Models for Out-of-Distribution Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755120

1 Introduction

In real-world applications, machine learning models often encounter inputs from unknown classes, known as out-of-distribution (OOD) data. Such OOD data can lead models to make erroneous predictions, posing significant safety risks, particularly in critical areas like autonomous driving [4] and medical diagnostics [2]. Consequently, the ability to detect these OOD samples is crucial in practice.

Traditional OOD detection methods in the image domain rely solely on visual information [14, 20, 21, 26, 31, 38, 40, 61, 62, 69]. Recently, with the rise of vision-language models (VLMs) [53], integrating textual information to improve OOD detection has gained increasing attention. Some initial studies have explored using pre-trained VLMs in a zero-shot manner [6, 30, 42], validating their strong capabilities. Recent efforts aim to further enhance their OOD detection abilities via model tuning, e.g., prompt tuning [3, 36, 45, 48, 81]. By enhancing the relevance of ID classes to the foreground object of ID images and reducing the impact of ID-irrelevant features, these tuned models enhance the separation for ID and OOD images. Although these tuned models bring certain improvements to the training data, their generalization performance to unseen classes and styles has been significantly reduced, as shown in Fig. 1. This indicates that current strategies fail to enhance the OOD detection capabilities of pre-trained VLMs comprehensively.



(a) CLIP **(b) LoCoOp** **(c) Ours(KR-NFT)**

Figure 1: GradCAM visualization of different methods on ID images. (a) In CLIP, the ID class shows high activation for both the foreground objects of ID images and the ID-irrelevant features. (b) In LoCoOp, although the activation of the ID class for ID-irrelevant features has decreased, there is also a reduction in the activation of the ID class for ID foreground objects when testing on unseen classes and styles. Consequently, the model's generalization performance for OOD detection has declined. (c) In our KR-NFT, the ID class shows a strong activation to ID foreground object when testing on unseen class and style, while exhibiting a low activation to ID-irrelevant features, demonstrating its strong comprehensive OOD detection capabilities.

We find that the reduced generalization performance of current tuning methods can be largely attributed to overfitting to training data and forgetting pre-trained knowledge. To address these issues, we propose a novel approach called Knowledge Regularized Negative Feature Tuning (KR-NFT). Specifically, unlike prompt tuning methods, which abandon the pre-trained text features and build new ones, our method decouples pre-trained knowledge and new knowledge in the text feature space, enabling the model to have stronger generalization capabilities for various types of OOD detection. This novel feature-tuning strategy is characterized by three critical properties that enhance generalization in OOD detection. *First*, the negative feature tuning directly introduces the distribution-aware learnable parameters on text features. By applying classification loss and OOD detection loss, the positive and negative text features can be optimized into different spaces. This creates a well-separated boundary to classify the ID and OOD images. *Second*, the image-conditional transformation dynamically generates input-conditional factors for each image by a lightweight meta-network. Integrating instance-adaptive characteristics into the feature tuning can drive it to learn invariant representations [10] and alleviate overfitting to class and style shifts. Furthermore, this image-conditional feature tuning shows high efficiency, as shown in Tab. 7. *Third*, the knowledge regularization optimization strategy minimizes the discrepancy between the pre-trained text features and tuned text features, alleviating knowledge forgetting. Together, these properties significantly reduce the forgetting of pre-trained knowledge and overfitting to training data, thereby enhancing the OOD detection capabilities of VLMs comprehensively.

We conduct thorough analyses on the negative feature tuning structure and its corresponding knowledge regularization strategy. With extensive experiments, our KR-NFT not only achieves state-of-the-art performance on the trained ID classes but also presents strong generalization capabilities to unseen classes and styles. Especially, with models tuned on the ImageNet dataset, our method outperforms the closest competitor [81] by 5.44% FPR95 on unseen classes and by 16.77% FPR95 on unseen styles on average. Besides its outstanding performance, our method is characterized by its fast training and testing speeds, as analyzed in Tab. 7. Moreover, KR-NFT is highly scalable and can be seamlessly combined with and enhance other methods, as shown in Tab. 4. We summarize our contribution as follows:

- We first identify a fundamental limitation of existing VLMs adaptation methods for OOD detection: while these tuned models bring improvements on the training data, they often suffer from reduced generalization performance on unseen classes and styles.
- To tackle this issue, we propose a novel KR-NFT method by integrating an innovative NFT adaptation architecture and a corresponding KR optimization strategy. The NFT presents high efficiency and scalability, characterized by the distribution-aware transformation design and instance-adaptive image-conditional modulation. The KR balances the new task learning against pre-trained knowledge forgetting, enhancing the generalization capabilities of OOD detection on both trained and unseen ID datasets.
- Validated with extensive experiments, our approach not only achieves new state-of-the-art results on the training data but also enhances the OOD detection capabilities of VLMs on unseen classes and styles. Especially, with models tuned on the ImageNet dataset, our method outperforms the closest competitor [81] by 5.44% FPR95 on unseen classes and by 16.77% FPR95 on unseen styles on average.

2 Related Work

Traditional OOD Detection. Traditional OOD detection methods can be roughly divided into three categories: classification-based [13, 16, 20, 22, 27, 29, 35, 38–40, 43, 49, 50, 57, 59, 60, 63, 70, 74, 76, 79], density-based [1, 28, 35, 52, 56, 69, 86], and distance-based [35, 44, 64, 77]. Classification-based methods can be further divided into post-hoc [13, 20, 29, 35, 38–40, 50, 57, 59, 60, 79], training-based [27, 63, 70, 76], and outlier exposure [16, 22, 43, 49, 74] approaches. For example, Liang *et al.* [38] enhanced the softmax score [20] with temperature scaling and small input perturbations. The confidence-estimating branch was designed in [11] for interpretable prediction. Hendrycks *et al.* [22] introduced real outliers to facilitate OOD detection. Density-based methods mainly rely on probabilistic models

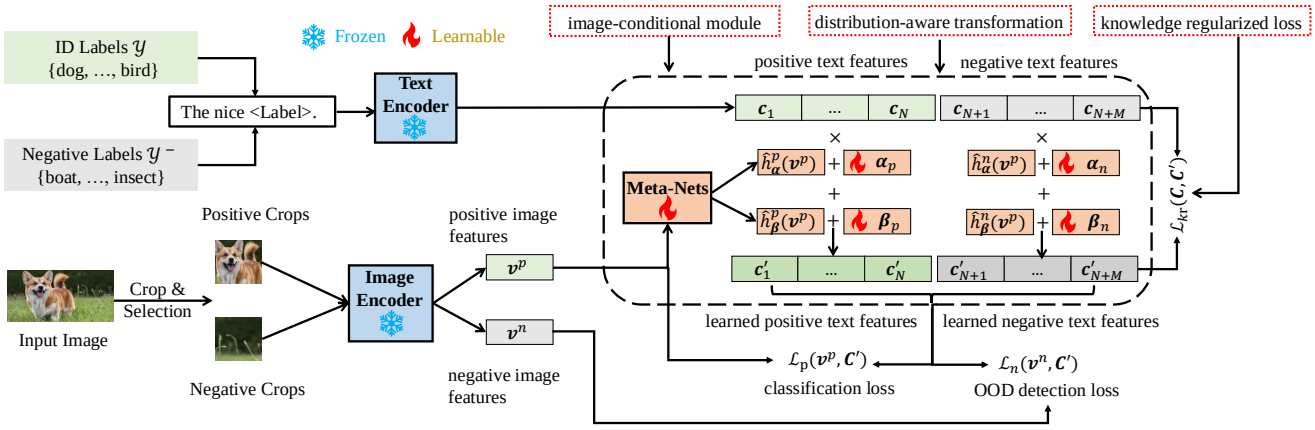


Figure 2: The overall framework of our KR-NFT, where we introduce a knowledge regularized negative feature tuning with three critical properties, i.e., image-conditional module, distribution-aware transformation, and knowledge regularized loss. These components are designed to enhance the generalization performance to unseen classes and styles in OOD detection.

and predict low-density regions as OOD [56]. Conversely, distance-based methods differentiate OOD samples by calculating the distance between test data and ID, employing metrics such as the Mahalanobis distance [35] or nearest-neighbor distance [61].

OOD Detection with VLMs. Enhancing OOD detection by incorporating language knowledge has gained increasing attention, which can be roughly divided into zero-shot [6, 30, 42, 80] and fine-tuning methods [3, 36, 45, 68, 81]. MCM [42] pioneered the zero-shot setting by revisiting the softmax score [20] with pre-trained VLMs. Then, negative labels were introduced in [6, 30] to further boost the performance. Recently, fine-tuning pre-trained VLMs to enhance OOD detection capabilities has been extensively studied [48, 68], with prompt tuning emerging as the most popular technique [3, 45, 81]. For instance, learning negative prompts corresponding to OOD samples has been explored in various studies [3, 48, 68, 81]. However, we have observed that while these model-tuning methods achieve certain improvements in the training data, their generalization performance on unseen classes and styles is significantly reduced. This suggests that these approaches fail to enhance the OOD detection capabilities of VLMs comprehensively.

CLIP Adaptation Methods. The common CLIP Adaptation Methods include full fine-tuning, prompt tuning [83, 85], feature tuning[17, 37, 75, 82], and parameter tuning [78]. Among these, TaskRes [75] and SSF [37] are the methods most similar to our approach. TaskRes[75] enhances classification performance on training data by adding a learnable residual vector to the text feature for each class, demonstrating its strengths in prior knowledge preservation and flexible task learning. SSF[37] employs feature scaling and shifting parameters to boost a model's category recognition capabilities. However, the learning objectives of these methods are centered on classification, which limits their ability to effectively detect OOD samples. Consequently, nvestigating the utilization of these adaptation methods to bolster the robustness of CLIP against OOD samples emerges as a critical area.

3 Method

3.1 Preliminaries

OOD Detection. Define $\mathcal{X}$ as the image domain and $\mathcal{Y} = \{y_1, \ldots, y_N\}$ as the class label domain, with examples $\mathcal{Y} = \{cat, dog, \ldots, bird\}$ and N denoting the total number of classes. We have $\boldsymbol{x}_{in} \in \mathcal{X}$ as the ID random variable and $\boldsymbol{x}_{ood} \in \mathcal{X}$ as the OOD random variable, with their respective distributions $\mathcal{P}_{\boldsymbol{x}_{in}}$ and $\mathcal{P}_{\boldsymbol{x}_{ood}}$. Typically, a test image $\boldsymbol{x}$ is expected to follow the ID and to belong to one ID class, $\boldsymbol{x} \in \mathcal{P}_{\boldsymbol{x}_{in}}$ and $y \in \mathcal{Y}$ where y is the label of $\boldsymbol{x}$. However, in practice, AI systems may encounter samples that do not match any known class, $\boldsymbol{x} \in \mathcal{P}_{\boldsymbol{x}_{ood}}$ and $y \notin \mathcal{Y}$, resulting in potential misclassifications and security issues [47, 58]. To mitigate this, OOD detection distinguishes between ID and OOD samples using a scoring function S:

$$G_\gamma(\boldsymbol{x}) = \begin{cases} \text{ID} & S(\boldsymbol{x}) \geq \gamma \\ \text{OOD} & S(\boldsymbol{x}) < \gamma \end{cases} \tag{1}$$

where G_γ is the OOD detector with threshold γ, determining ID if $S(\boldsymbol{x}) \geq \gamma$.

OOD detection with VLMs. Detecting OOD images by employing textual information has garnered increasing attention recently. For a test image $\boldsymbol{x}$ and the target label set $\mathcal{Y}$, we derive the image feature $\boldsymbol{v} = f_{img}(\boldsymbol{x}) \in \mathcal{R}^D$ and the textual features $\boldsymbol{C} = f_{txt}(\rho(\mathcal{Y})) \in \mathcal{R}^{N \times D}$ using pre-trained encoders, where D is the feature dimension. Here, $f_{img}(\cdot)$ and $f_{txt}(\cdot)$ represent the image and text encoders, respectively. The function $\rho(\cdot)$ generates text prompts, typically formatted as 'a photo of a <label>', where <label> is replaced with specific class names such as 'cat' or 'dog'. Both $\boldsymbol{v}$ and $\boldsymbol{C}$ are processed via L_2 normalization across the dimension D.

To explore additional knowledge from a broader text space, NegLabel [30] introduces negative labels $\mathcal{Y}^- = \{y_{N+1}, \ldots, y_{N+M}\}$ by identifying words that are far from the ID labels $\mathcal{Y}$ in extensive text corpora, leading to the following score function:

$$S_{NegLabel}(\boldsymbol{v}) = \frac{\sum_{i=1}^{N} e^{\cos(\boldsymbol{v}, \boldsymbol{c}_i)}}{\sum_{i=1}^{N} e^{\cos(\boldsymbol{v}, \boldsymbol{c}_i)} + \sum_{j=N+1}^{N+M} e^{\cos(\boldsymbol{v}, \boldsymbol{c}_j)}}. \tag{2}$$

Prompt Learning and Task Residual Learning. CoOp [85] applies prompt learning to CLIP by adding some learnable context tokens in the input template. The prompt can be formulated as $\rho_p(y_i) = [V]_1[V]_2 \cdots [V]_L[y_i]$ and each $[V]_l$ is a vector with the same dimension as word embeddings of CLIP. To improve generalization performance, CoCoOp [84] introduces a lightweight MetaNet that generates each image an input-conditional token $h(x)$, then adds this token to each context token $\boldsymbol{V}_i(x) = \boldsymbol{V}_i(x) + h(x)$. Different from prompt learning, which introduces learnable context tokens on input, TaskRes [75] introduces a portion of prior-independent parameters as a residual on the pre-trained classifier, by decoupling the old knowledge and the new target knowledge on the text features, TaskRes[75] enhanced retention of existing knowledge and allows for more flexible investigation of task-specific knowledge. The new text features $\boldsymbol{c}'$ can be formulated as: $\boldsymbol{c}' = \boldsymbol{c} + \mu\boldsymbol{\beta}$, μ is a scaling factor and $\boldsymbol{\beta}$ is a set of learnable parameters.

3.2 Motivation

Previous studies [3, 36, 45, 79] have explored applying prompt learning to enhance CLIP's capabilities on OOD detection. Though it can enhance the separation for ID and OOD images and bring certain improvements for OOD detection, these methods have a common issue: the tuned model's OOD detection performance on ID images from unseen classes and styles drops significantly, as illustrated in the Tab. 2 and Tab. 3. This motivates us to the following research question:

How to enhance the generalization performance of VLMs in OOD detection?

To this end, we conduct experiments to investigate the correlation between the ID images and the positive text features. First, we find that CLIP has a strong alignment capability for ID images and ID labels, but it is easily affected by ID-irrelevant features, as shown in Fig. 1a. Then, we observe that the drop in OOD detection for unseen ID classes in existing prompt learning methods is closely related to the reduction in the gradient weights of class activation mapping for ID foreground objects, suggesting a forgetting of CLIP's pre-trained knowledge, as shown in Fig. 1b. Therefore, a new framework is desired to balance the preservation of CLIP pre-trained knowledge and the learning of new tasks.

3.3 Negative Feature Tuning.

Existing works[3, 36, 48, 68] focus on designing negative prompts to empower CLIP with OOD detection capability. However, as illustrated in [48], employing a shared negative prompt fails to capture the diversity of negative features, and learning class-specific negative prompts will bring huge computational cost, as shown in the Tab. 7. To address these issues, we consider introducing learnable factors directly on negative text features. We first formulate the transformation process of the text feature $\boldsymbol{C}$ in prompt tuning as:

$$\boldsymbol{c}_i' = f_{txt}(\rho_p(y_i)) = T(\boldsymbol{c}_i), \tag{3}$$

where the prompt $\rho_p(y_i) = [V]_1[V]_2 \cdots [V]_L[y_i]$ and each $[V]_l$ is a vector with the same dimension. The tuned text feature $\boldsymbol{c}_i'$ is a transformed output of the vanilla text feature $\boldsymbol{c}_i$ generated by applying a general feature transformation function $T(\cdot)$. In other words, we could potentially achieve similar effects to prompt tuning by directly transforming pre-trained text features. To verify this, we experimented with various implementations of transformation functions, including shifting by a constant, element-wise scaling/shifting, and MLP-based transformation, as analyzed in the Supplementary Material. We find that the simple element-wise scaling and shifting transformation achieves a good balance between capability and complexity. The learnable independent parameters are denoted as $\boldsymbol{\alpha}_{\text{neg}} = \{\omega_1^{\text{neg}}, \omega_2^{\text{neg}}, \ldots, \omega_n^{\text{neg}}\}$ and $\boldsymbol{\beta}_{\text{neg}} = \{b_1^{\text{neg}}, \omega_2^{\text{neg}}, \ldots, \omega_n^{\text{neg}}\}$, where $\boldsymbol{\alpha}_{\text{neg}} \in \mathcal{R}^D$ and $\boldsymbol{\beta}_{\text{neg}} \in \mathcal{R}^D$ are the element-wise scaling and shifting parameters, respectively. The negative feature tuning can ne denoted as:

$$\boldsymbol{c}_i' = T(\boldsymbol{c}_i) = L_2(\boldsymbol{\alpha}_{\text{neg}}\boldsymbol{c}_i + \boldsymbol{\beta}_{\text{neg}}), \tag{4}$$

$L_2(\cdot)$ indicates the L_2 normalization along the feature dimension.

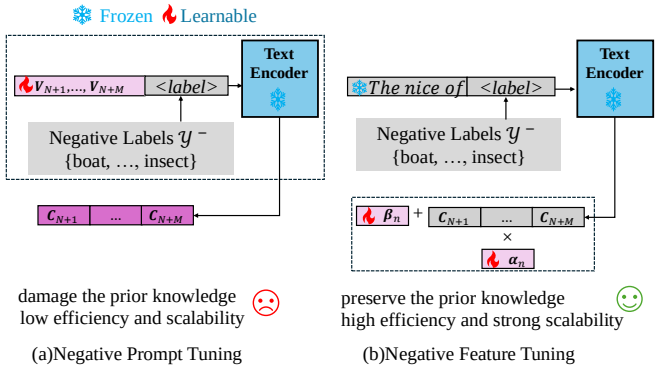


Figure 3: Illustration of (a) Negative Prompt Tuning and (b) our Negative Feature Tuning. Our method introduces independent learnable factors directly on the pretrained features and does not require gradient backpropagation through the heavy text encoder. This demonstrates its advantage in preserving pre-trained knowledge and enhancing efficiency.

Image-Conditional Module. To enhance the generalization performance of learned prompts, conditioning the text prompts [84] on image features is a potential solution. We find that such conditional prompts indeed enhance the generalization performance; however, the testing speed is significantly reduced since it requires recalculating text features for each instance, as shown in Tab. 7. To balance the computational efficiency and generalization, we propose an image-conditional feature transformation module. Specifically, we can dynamically generate image conditional tokens for each test sample using neural networks, this can be formulated as:

$$\boldsymbol{c}_i' = L_2(h_{\boldsymbol{\alpha}}^n(\boldsymbol{v})\boldsymbol{c}_i + h_{\boldsymbol{\beta}}^n(\boldsymbol{v})), \tag{5}$$

The lightweight meta-net is denoted as $\widehat{h}_{\boldsymbol{\alpha}}^n(*)$ that takes image feature $\boldsymbol{v}$ as input and outputs the residual parameters. Then, these conditional residual parameters will be added to the learnable parameters $\boldsymbol{\alpha}$ to get a new image-conditional learnable parameters $h_{\boldsymbol{\alpha}}^n(\boldsymbol{v})$, which can formulated as $h_{\boldsymbol{\alpha}}^n(\boldsymbol{v}) = \boldsymbol{\alpha} + \widehat{h}_{\boldsymbol{\alpha}}^n(\boldsymbol{v})$. And the $h_{\boldsymbol{\beta}}(\boldsymbol{v})$ is similarly defined. The proposed image-conditional module can largely reduce the model's sensitivity to shifts in class and style, thereby enhancing its generalization capability, as validated in Tab

6. Furthermore, it demonstrates strong advantages in computational efficiency, as presented in Tab. 7.

Distribution-aware Transformations. Unlike traditional classification tasks, which focus on a single ID, OOD detection tasks often introduce additional negative proxies to represent the negative distribution [3, 30]. Considering these varying distribution characteristics, we design a distribution-aware transformation functions for positive and negative text features, which can be formulated as:

$$\boldsymbol{c}_i' = \begin{cases} L_2(h^p_{\boldsymbol{\alpha}}(\boldsymbol{v})\boldsymbol{c}_i + h^p_{\boldsymbol{\beta}}(\boldsymbol{v})), & \text{if } i \le N \\ L_2(h^n_{\boldsymbol{\alpha}}(\boldsymbol{v})\boldsymbol{c}_i + h^n_{\boldsymbol{\beta}}(\boldsymbol{v})), & \text{otherwise} \end{cases} \tag{6}$$

where h^p_* and h^n_* are the transformation functions for positive and negative text features, respectively. As analyzed in Figure 4d, such a distribution-aware setting significantly outperforms its distribution-agnostic counterpart.

3.4 Knowledge Regularized Optimization.

To enhance the separation capability between ID and OOD images through text features and avoid the text features overfitting on the training classes, we design a strategy named Knowledge Regularized Optimization. This strategy consists of three losses: traditional classification loss, ood detection loss, and pre-trained knowledge regularization loss.

Outlier Samples Generation. We validate our method under the few-shot setting, where S ID samples per class and a total of $K = SN$ ID instances are available for training, $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_K, y_K)\}$. We follow [3] to generate the negative training samples via image cropping and selection. Specifically, we apply multiple random cropping to each ID sample $\boldsymbol{x}_k$ and get the cropping set $X_k^{crop} = \{\boldsymbol{x}_{k,1}^{crop}, ..., \boldsymbol{x}_{k,P}^{crop}\}$, where P is the number of random cropping. These croppings include positive samples that contain the target object, as well as negative samples do not contain any objects of interest (e.g., background only). We distinguish these two types of croppings by measuring their cosine similarity to the text feature of the corresponding label (e.g., 'a photo of a $<y_k>$'). We define croppings with the highest and lowest cosine similarities as $X_k^p = \{\boldsymbol{x}_{k,1}^p, \ldots, \boldsymbol{x}_{k,Q}^p\}$ and $X_k^n = \{\boldsymbol{x}_{k,1}^n, \ldots, \boldsymbol{x}_{k,Q}^n\}$, where Q is a hyperparameter that determines the number of selected croppings. Finally, we construct the training data as $\mathcal{D}_p = \{(\boldsymbol{x}_{1,1}^p, y_1^p), (\boldsymbol{x}_{1,2}^p, y_1^p), \ldots, (\boldsymbol{x}_{K,Q}^p, y_K)\}$ and $\mathcal{D}_n = \{(\boldsymbol{x}_{1,1}^n, \boldsymbol{x}_{1,2}^n, \ldots, \boldsymbol{x}_{K,Q}^n)\}$. Our method is also compatible with other strategies that introduce negative training samples (e.g., exploring negative local features in LoCoOp [45]), as validated in the Supplementary Material.

Traditional Classification Loss. Given the constructed model and the prepared training data, we optimize our method by maximizing the distinction between ID and OOD data. Specifically, we first employ the cross-entropy loss to enhance the model's ability to recognize ID classes:

$$\mathcal{L}_p(\boldsymbol{x}^p, y^p) = -\log \frac{e^{\cos(\boldsymbol{v}^p, \boldsymbol{c}'_{y^p})}}{\sum_{i=1}^{N} e^{\cos(\boldsymbol{v}^p, \boldsymbol{c}_i')} + \sum_{j=N+1}^{N+M} e^{\cos(\boldsymbol{v}^p, \boldsymbol{c}_j')}}, \tag{7}$$

where $(\boldsymbol{x}^p, y^p) \in \mathcal{D}_{\text{p}}$, and $\boldsymbol{v}^p = f_{img}(\boldsymbol{x}^p)$. Here, we reuse y^p as the label indices of $\boldsymbol{x}^p$ with a slight abuse of notation. The scaling temperature is also omitted for readability.

OOD Detection Loss. In addition to enhancing the recognition of ID classes, Eq. (7) also pushes positive samples away from the negative labels. Similarly, we push negative samples away from the ID labels. Given that there is no one-to-one correspondence between negative samples and negative labels, we maximize the similarity between negative samples and the aggregate of negative labels. For optimization considerations, we adopt an equivalent objective by minimizing the similarity between negative samples and the aggregate of ID labels:

$$\mathcal{L}_n(\boldsymbol{x}^n) = \log \frac{\sum_{i=1}^{N} e^{\cos(\boldsymbol{v}^n, \boldsymbol{c}_i')}}{\sum_{i=1}^{N} e^{\cos(\boldsymbol{v}^n, \boldsymbol{c}_i')} + \sum_{j=N+1}^{N+M} e^{\cos(\boldsymbol{v}^n, \boldsymbol{c}_j')}}, \tag{8}$$

where $\boldsymbol{x}^n \in \mathcal{D}_{\text{n}}$, and $\boldsymbol{v}^n = f_{img}(\boldsymbol{x}^n)$.

Knowledge Regularization Loss. Although the adopted simple transformation function significantly preserves pre-trained knowledge, its implicitly structured design leaves room for further improvement. To further enhance the retention of pre-trained knowledge, we are inspired by the knowledge distillation framework [25] to introduce an explicit consistency constraint. Specifically, we implement an objective that ensures the transformed text features remain consistent with the original ones in their outputs. As shown in Fig. 4c, we investigated different locations for applying this consistency regularization, including on features, logits, or predicted probabilities, to determine the most effective approach. The best performance is achieved by maximizing the cosine similarity between pre-trained text features $\boldsymbol{c}_i$ and the learned $\boldsymbol{c}_i'$:

$$\mathcal{L}_{\text{kr}} = \frac{1}{N+M} \sum_{i=1}^{N+M} (1 - \boldsymbol{c}_i \boldsymbol{c}_i'). \tag{9}$$

This consistency regularization objective can be understood as applying an additional constraint on the adopted transformation function. This constraint explicitly ensures that the transformations applied are subtle, in line with residual learning principles [18]. By limiting these transformations to minimal changes, this approach effectively preserve pre-trained knowledge, achieving a balance between acquiring new task capabilities and preserving the integrity of pre-trained knowledge.

Learning Objectives. We construct our model based on NegLabel [30], which identifies OOD samples by comparing the similarity of image features to ID labels and mined negative labels, as shown in Eq. (2). The overall learning objective is formulated as follows:

$$\mathcal{L} = \mathcal{L}_p + \lambda_1 \mathcal{L}_n + \lambda_2 \mathcal{L}_{kr}, \tag{10}$$

where λ_1 and λ_2 are balancing parameters. The overall framework is illustrated in Fig. 2.

3.5 Inference.

Existing fine-tuning methods for VLMs typically conduct evaluation on the training classes only [3, 45, 81]. However, we found that these tuned models exhibit reduced generalization performance on unseen classes, as shown in Fig. 1. This suggests that current tuning methods tend to overfit the training data and do not comprehensively enhance the VLMs' OOD detection capabilities. To comprehensively evaluate the OOD detection ability of tuned VLMs,

Methods	OOD datasets									
	INaturalist		SUN		Places		Textures		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Zero-shot methods)										
MCM [42]	94.59	32.20	92.25	38.80	90.31	46.20	86.12	58.50	90.82	43.93
EOE [6]	97.52	12.29	95.73	20.40	92.95	30.16	85.64	57.63	92.96	30.09
NegLabel [30]	99.49	1.91	95.49	20.53	91.64	35.59	90.22	43.56	94.21	25.40
Tuning-based methods										
MSP [20]	87.44	58.36	79.73	73.72	79.67	74.41	79.69	71.93	81.63	69.61
ZOC [15]	86.09	87.30	81.20	81.51	83.39	73.06	76.46	98.90	81.79	85.19
CLIPN [68]	95.27	23.94	93.93	26.17	92.28	33.45	90.93	40.83	93.10	31.10
LSN [48]	95.83	21.56	94.35	26.32	91.25	34.48	90.42	38.54	92.26	30.22
LoCoOp [45]	93.93	29.45	90.32	41.13	90.54	44.15	93.24	33.06	92.01	36.95
ID-like [3]	98.19	8.98	91.64	42.03	90.57	44.00	**94.32**	**25.27**	93.68	30.07
NegPrompt [36]	90.49	37.79	92.25	32.11	91.16	35.52	88.38	43.93	90.57	37.34
SCT [36]	95.86	13.94	95.33	20.55	92.24	29.86	89.06	41.51	93.27	26.47
LAPT [81]	99.63	1.16	96.01	19.12	92.01	33.01	91.06	40.32	94.68	23.40
KR-NFT (Ours)	99.62	**0.82**	96.15	17.83	92.64	36.12	91.96	36.38	95.09	22.79
KR-NFT (λ_2=0)	**99.67**	1.33	**96.28**	**17.46**	**93.68**	**28.17**	93.26	29.34	**95.82**	**19.08**

Table 1: OOD detection results for ID of ImageNet-1k and four OOD datasets using a VITB/16 encoder.

we conduct inference not only on the trained ID classes but also directly evaluate the model, which is tuned on Base dataset, on unseen New classes. Specifically, given a VLM typically tuned with samples $\boldsymbol{x}_{in}$ from the base ID distribution $\mathcal{P}_{\boldsymbol{x}_{in}}$, we first assess its OOD detection performance using samples $\boldsymbol{x}_{in}/\boldsymbol{x}_{ood}$ from $\mathcal{P}_{\boldsymbol{x}_{in}}/\mathcal{P}_{\boldsymbol{x}_{ood}}$, following the common pipeline [26, 81]. Additionally, we evaluate the same model with samples $\boldsymbol{x}_{in}^{new}/\boldsymbol{x}_{ood}^{new}$ from unseen class distributions $\mathcal{P}_{\boldsymbol{x}_{in}^{new}}/\mathcal{P}_{\boldsymbol{x}_{ood}^{new}}$, where the label space $\mathcal{Y}^{new}$ of $\mathcal{P}_{\boldsymbol{x}_{in}^{new}}$ differs from the Base classes $\mathcal{Y}$. In addition to examining new distributions with class shifts, we also investigated another new distribution shift featured by image styles, which retains the same label space $\mathcal{Y}$ as the Base classes and presents new image styles. We follow NegLabel to calculate the score in the testing stage, where we replace the pre-trained text feature $\boldsymbol{c}_i$ with the learned $\boldsymbol{c}_i'$ in Eq. (2).

4 Experiments

4.1 Experiments Setup

Datasets and Benchmarks. Following [26], We select ImageNet-1K [8] as the ID dataset and use iNaturalist [65], SUN [71], Places [83], and Textures [7] as OOD test datasets. To evaluate the generalization capability to unseen classes, we directly apply the model tuned on ImageNet to CIFAR10 [34], CIFAR100 [34], and four fine-grained datasets [5, 33, 51, 66]. For CIFAR datasets, we follow the OpenOOD setting [72] and use MNIST [9], SVHN [46], Places [83], and Textures [7] as OOD test datasets. For fine-grained datasets, we follow [6] to randomly select half of the classes as ID and use the remaining half as OOD. We also evaluate the generalization to unseen image styles, where we apply the model tuned on ImageNet to covariate-shifted ID datasets of ImageNet-R [19], ImageNet-V2 [54], ImageNet-A [24], and ImageNet-Sketch [67].

Implementation Details. We conduct the experiments under the four-shot learning setting (S=4) with a ViT-B/16 visual encoder. For each training sample, we generate P=256 random-sized crops, supplemented with color augmentation methods such as color jitter and grayscale. From these crops, we select Q=32 crops to construct the positive training set $\mathcal{D}_p$ and the negative training set $\mathcal{D}_n$, respectively. We utilize 10,000 negative text labels following [30]. In the optimization process, we use the AdamW optimizer [32] to train the model for three epochs at a learning rate of 1e-5. We use λ_1=0.3 and λ_2=100 in all experiments. We initialize $\boldsymbol{\alpha}$ as all ones, $\boldsymbol{\beta}$ as all zeros, and constrain the initial output of $\widehat{h}_*(\cdot)$ to be zeros to avoid disturbing pre-trained knowledge at the beginning of training.

4.2 Main Results

Results on Training Dataset. As shown in the Tab. 1, our KR-NFT outperforms other competitors on average. We report the traditional methods [14, 20, 26, 38, 40, 61, 62, 69] by fine-tuning CLIP visual encoders with the ImageNet training data as described in [30], and reference the results of other methods [3, 6, 15, 36, 45, 48, 68, 81] directly from their respective papers. Although setting $\lambda_2 = 0$ yields superior results on the Base ImageNet dataset, configuring λ_2 to a moderate value optimally balances performance across both the Base and New test sets, as analyzed in Fig. 4b.

Generalization Results to Unseen Classes. A good OOD detection model should not only perform well on training data but also demonstrate strong generalization on unseen classes and styles. To this end, we assess the generalization performance of our model by applying it, once tuned on the ImageNet dataset, to the ID datasets of CIFAR10, CIFAR100, and four fine-grained datasets. As shown in the Tab. 2, while existing model tuning methods, such as LoCoOp and LAPT, achieve certain improvements in training classes, their generalization to new classes significantly declines. In contrast, our method not only achieves substantial improvements on training classes but also enhances performance on unseen classes, presenting a comprehensively enhanced OOD detection capability.

Generalization Results to Unseen Styles. In addition to evaluating the generalization performance on unseen classes, we also validate our model on unseen image styles. As shown in Tab. 3, our method demonstrates strong generalization across various natural style shifts compared to other few-shot training methods, confirming its robustness.

Methods	Unseen New Classes							
	CIFAR10		CIFAR100		Fine-grained		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Zero-shot methods								
MCM [42]	96.27	14.28	79.92	54.54	78.21	68.72	84.80	45.84
NegLabel [30]	96.69	12.15	80.21	54.05	89.99	41.28	88.96	35.82
Tuning-based methods								
LoCoOp [45]	93.61	26.36	73.53	57.03	81.32	70.24	82.82	52.21
ID-like [3]	85.80	39.98	69.54	67.11	61.10	82.20	72.15	63.09
NegPrompt [36]	92.79	29.21	74.82	63.44	78.83	68.32	82.14	54.66
SCT [73]	95.17	18.28	77.55	62.99	83.11	65.29	85.28	48.85
LAPT [81]	95.21	17.26	78.94	57.11	89.23	46.91	87.39	40.43
KR-NFT (Ours)	**96.82**	**11.41**	**80.95**	**53.24**	**89.99**	**40.33**	**89.25**	**34.99**

Table 2: OOD detection results on unseen classes, where results of [3, 36, 45, 81] and ours are reported with models tuned on the ImageNet dataset. Pre-trained CLIP model is utilized in [30, 42]. Detailed results are provided in the Supplementary Material.

Methods	Unseen New Styles									
	ImageNet-S		ImageNet-A		ImageNet-R		ImageNet-V2		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Zero-shot methods										
MCM [42]	82.26	70.13	72.16	80.88	76.68	76.72	87.43	55.61	79.68	70.84
NegLabel [30]	93.59	27.42	87.94	43.23	94.54	20.63	93.08	29.77	92.29	30.26
Tuning-based methods										
LoCoOp [45]	68.32	73.35	72.66	77.64	50.36	89.14	74.33	62.48	66.42	76.65
ID-like [3]	74.29	75.92	80.23	67.24	83.03	61.09	88.11	49.57	81.41	63.46
NegPrompt [36]	77.57	65.18	69.82	76.57	80.70	56.01	74.06	63.42	75.53	65.30
SCT [73]	85.11	53.30	80.30	63.13	85.23	48.81	90.66	36.47	85.33	50.43
LAPT [81]	84.80	51.87	87.18	51.47	87.75	47.42	93.30	**28.84**	88.26	44.90
KR-NFT (Ours)	**94.10**	**24.81**	**89.13**	**38.46**	**94.66**	**20.12**	**93.61**	29.13	**92.88**	**28.13**

Table 3: OOD detection results on unseen styles, where results of [3, 36, 45, 81] and ours are reported with models tuned on the ImageNet dataset. Pre-trained CLIP model is utilized in [30, 42]. Detailed results are provided in the Supplementary Material.

Compatibility. As shown in Tab. 4, our method is highly compatible with existing prompt-tuning-based approaches, bringing consistent improvement to OOD detection capabilities.

Evaluation on OpenOOD benchmark. We evaluated our method on the OpenOOD benchmark, which includes far-OOD and near-OOD datasets. As shown in the table 5, our KR-NFT performs better than Neglabel on both far-OOD and near-OOD datasets. This demonstrates the robustness of KR-NFT, making it applicable across various OOD scenarios.

Methods	Unseen New Classes		Unseen New Styles	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓
LoCoOp [45]	82.82	52.21	66.42	76.65
LoCoOp + KR-NFT [45]	**85.30**	**49.37**	**78.98**	**70.06**
SCT [73]	85.28	48.85	85.33	50.43
SCT + KR-NFT [73]	**86.01**	**45.53**	**85.96**	**49.58**
LAPT [81]	87.39	40.43	88.26	44.90
LAPT + KR-NFT [36]	**88.57**	**36.35**	**89.68**	**41.08**

Table 4: OOD detection performance on compatibility experiments. We report the average OOD detection results on unseen classes and style.

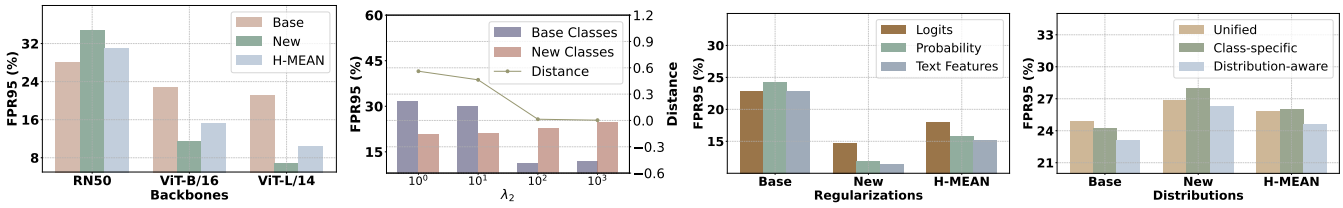
4.3 Ablation and Analyses

4.3.1 Ablation Study. Our analyses are mainly conducted under the cross-class generalization setting, where we tune a pre-trained VLM on ImageNet and validate its performance on both ImageNet (Base) and the unseen CIFAR10 (New). We introduce H-MEAN, a harmonic mean of results from both settings, to assess the comprehensive OOD detection capability.

Methods	FPR95 ↓		AUROC ↑	
	NearOOD	FarOOD	NearOOD	FarOOD
GEN [41]	–	–	78.97	90.98
AugMix [23] + ReAct [59]	–	–	79.94	93.70
RMDS [55]	–	–	80.09	92.60
AugMix [23] + ASH [12]	–	–	82.16	96.05
MCM [42]	79.02	68.54	60.11	84.77
NegLabel [30]	68.18	25.40	76.92	93.30
Ours	67.24	19.08	77.73	95.82

Table 5: OOD detection results on the OpenOOD benchmark, where ImageNet-1k is adopted as ID dataset. Full results are available in the Supplementary Materials.

Component Ablation. As illustrated in Tab. 6, adapting text features with the transformation function indeed improved performance on the Base classes, but significantly compromised generalization to unseen New classes. Introducing the image-conditional transformation mitigated the performance decline in New classes, but there still remains a gap compared to the pre-trained CLIP, suggesting a forgetting of pre-trained knowledge. The inclusion of an knowledge regularization objective $\mathcal{L}_{kr}$ successfully balanced learning for the new task with the preservation of pre-trained knowledge, achieving results that surpass the pre-trained CLIP



(a) Backbones **(b) Values of λ_2** **(c) Knowledge Regularization** **(d) Distributions of NFT**

Figure 4: Analyses on (a) hidden dimensions, (b) values of λ_2, (c) knowledge regularization strategies, and (d) distributions.

in both Base and New settings. Lastly, incorporating distribution-aware information further enhanced model performance, leading to the best results.

Components				FRR95 ↓		
FT	CoFT	KR	NFT	Base	New	H-MEAN
NegLabel with Pre-trained CLIP				25.40	12.15	16.45
✓	✗	✗	✗	25.02	26.08	25.54
✓	✓	✗	✗	25.21	20.43	22.57
✓	✓	✓	✗	25.09	11.86	16.11
✓	✓	✓	✓	22.79	11.41	15.20

Table 6: Ablation results on the proposed components, where 'FT' indicates the feature tuning in Eq. (4), 'CoFT' denotes the image-conditional feature tuning variant in Eq. (5), 'KR' is the pre-trained feature knowledge regularization loss in Eq. (9), and 'NFT' indicates the negative feature tuning in Eq. (6).

Different Backbones. Our KR-NFT is compatible with different VLM backbones. As shown in the Fig. 4a, KR-NFT achieves a lower FPR95 on larger backbones across various OOD detection scenarios, indicating that larger backbones provide greater performance improvement. Furthermore, compared to Neglabel, KR-NFT demonstrates better performance across different backbones, highlighting the strong robustness of KR-NFT with various architectures.

The sensitivity to different λ_2. As shown in Fig. 4b, with the increase of λ_2, the performance on the Base dataset tends to decrease, as this regularization limits learning on the Base classes. However, a higher value of λ_2 is beneficial for the preservation of pre-trained knowledge, by reducing the distance between the pre-trained text features and new learned ones, as validated by the improved performance on the New classes dataset. We find that a moderate value of $\lambda_2 = 100$ achieves good results on both datasets, balancing the preservation of pre-trained knowledge and acquisition of new information.

Knowledge Regularization Implementations. Besides applying regularization on text features C' as in Eq. 9, we also explore the knowledge regularization objective on the logits (e.g., $C'v$) and probabilities (e.g., Softmax($C'v$)), which are detailed in the Supplementary Material. As shown in Fig. 4c, feature regularization yields the best results.

Distribution-aware Transformation. As shown in Fig. 4d, using distribution-aware transformation as in Eq. 6 outperforms using one unified and class-specific transformation, which is expected as binary transformations can better capture the characteristics of the distributions of ID and OOD images.

4.3.2 The Effectiveness of KR-NFT. To understand the effectiveness of KR-NFT, we calculated the cosine similarity between the ID/OOD images and the positive/negative text features with features of pre-trained CLIP and our KR-NFT, the detailed results can be found in the supplementary materials.

4.3.3 Computation Efficiency. As shown in Tab. 7, although we introduce certain training parameters, our KR-NFT enjoys fast training and small TFLOPs, because we do not need gradient backpropagation through a heavy text encoder like other prompt tuning-based methods. We also observed that although ConNegPT improves generalization on the New dataset compared to NegPT, it significantly slows down the test speed, because it requires re-forwarding the text encoder for each test image. In contrast, our method maintains a fast test speed and outperforms in performance, presenting advantages in both effectiveness and efficiency.

Methods	Train	Test	TFLOPs	Param.	Base	New
LoCoOp [45]	0.33h	10.9ms	38.25	8K	42.32	17.35
ID-like [3]	3.3h	7.4ms	39.12	16K	31.78	54.99
NegPT*	1.60h	21.5ms	193.76	2K	24.21	27.33
ConNegPT*	1.65h	936.1ms	194.53	3K	27.27	25.42
KR-NFT (Ours)	0.25h	21.6ms	0.11	6K	22.79	11.41

Table 7: 'Train' and 'Test' measures the training time and testing time, respectively. 'TFLOPs' are calculated during training with gradient back-propagation, and 'Param.' presents the number of learnable parameters. 'Base' and 'New' indicate the FPR95 on the ImageNet and CIFAR10 datasets, respectively. *We compare our KR-NFT with negative prompt tuning (NegPT) and conditional negative prompt tuning (ConNegPT). Results are reported with a GeForce RTX 3090 GPU.

5 Conclusion

In this paper, we reveal a limitation in existing negative prompt tuning methods for OOD detection: these models suffer from reduced generalization performance on unseen classes and styles. To address this issue, we propose a novel framework named Knowledge Regularized Negative Feature Tuning (KR-NFT). NFT optimizes positive and negative features into distinct spaces, maximizing the separation between ID and OOD images. The image-conditional learnable factors explore instance-adaptive knowledge, reducing overfitting to the training class and style. KR minimizes the discrepancy between pre-trained text features and tuned ones, alleviating knowledge forgetting. Extensive experiments demonstrate the effectiveness of our KR-NFT and its compatibility with other negative prompt-tuning methods.

References

[1] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. 2019. Latent space autoregression for novelty detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 481–490.

[2] Syed Muhammad Anwar, Muhammad Majid, Adnan Qayyum, Muhammad Awais, Majdi Alnowami, and Muhammad Khurram Khan. 2018. Medical image analysis using convolutional neural networks: a review. *Journal of medical systems* 42 (2018), 1–13.

[3] Yichen Bai, Zongbo Han, Bing Cao, Xiaoheng Jiang, Qinghua Hu, and Changqing Zhang. 2024. ID-like Prompt Learning for Few-Shot Out-of-Distribution Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17480–17489.

[4] Daniel Bogdoll, Maximilian Nitsche, and J Marius Zöllner. 2022. Anomaly detection in autonomous driving: A survey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4488–4499.

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*. Springer, 446–461.

[6] Chentao Cao, Zhun Zhong, Zhanke Zhou, Yang Liu, Tongliang Liu, and Bo Han. 2024. Envisioning Outlier Exposure by Large Language Models for Out-of-Distribution Detection. *arXiv preprint arXiv:2406.00806* (2024).

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3606–3613.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[9] Li Deng. 2012. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine* 29, 6 (2012), 141–142.

[10] Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor G Turrisi da Costa, Cees GM Snoek, Georgios Tzimiropoulos, and Brais Martinez. 2023. Bayesian prompt learning for image-language model generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15237–15246.

[11] Terrance DeVries and Graham W Taylor. 2018. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865* (2018).

[12] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. 2022. Extremely simple activation shaping for out-of-distribution detection. *arXiv preprint arXiv:2209.09858* (2022).

[13] Xin Dong, Junfeng Guo, Ang Li, Wei-Te Ting, Cong Liu, and HT Kung. 2022. Neural mean discrepancy for efficient out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19217–19227.

[14] Xuefeng Du, Xin Wang, Gabriel Gozum, and Yixuan Li. 2022. Unknown-aware object detection: Learning what you don't know from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13678–13688.

[15] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. 2022. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 6568–6576.

[16] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. 2021. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems* 34 (2021), 7068–7081.

[17] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision* 132, 2 (2024), 581–595.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[19] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. 2021. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8340–8349.

[20] Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136* (2016).

[21] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100* (2020).

[22] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. 2018. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606* (2018).

[23] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. 2019. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781* (2019).

[24] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15262–15271.

[25] Geoffrey Hinton. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531* (2015).

[26] Rui Huang, Andrew Geng, and Yixuan Li. 2021. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems* 34 (2021), 677–689.

[27] Rui Huang and Yixuan Li. 2021. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8710–8719.

[28] Dihong Jiang, Sun Sun, and Yaoliang Yu. 2021. Revisiting flow generative models for out-of-distribution detection. In *International Conference on Learning Representations*.

[29] Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. 2023. Detecting out-of-distribution data through in-distribution class prior. In *International Conference on Machine Learning*. PMLR, 15067–15088.

[30] Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. 2024. Negative label guided ood detection with pretrained vision-language models. *arXiv preprint arXiv:2403.20078* (2024).

[31] Di Jin, Shuyang Gao, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tür. 2022. Towards textual out-of-domain detection without in-domain labels. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 1386–1395.

[32] Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[33] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*. 554–561.

[34] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).

[35] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems* 31 (2018).

[36] Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. 2024. Learning transferable negative prompts for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17584–17594.

[37] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. 2022. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems* 35 (2022), 109–123.

[38] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690* (2017).

[39] Ziqian Lin, Sreya Dutta Roy, and Yixuan Li. 2021. Mood: Multi-level out-of-distribution detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 15313–15323.

[40] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems* 33 (2020), 21464–21475.

[41] Xixi Liu, Yaroslava Lochman, and Christopher Zach. 2023. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 23946–23955.

[42] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. 2022. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems* 35 (2022), 35087–35102.

[43] Yifei Ming, Ying Fan, and Yixuan Li. 2022. Poem: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning*. PMLR, 15650–15665.

[44] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. 2022. How to exploit hyperspherical embeddings for out-of-distribution detection? *arXiv preprint arXiv:2203.04450* (2022).

[45] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. 2024. Locoop: Few-shot out-of-distribution detection via prompt learning. *Advances in Neural Information Processing Systems* 36 (2024).

[46] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, Vol. 2011. Granada, 4.

[47] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 427–436.

[48] Jun Nie, Yonggang Zhang, Zhen Fang, Tongliang Liu, Bo Han, and Xinmei Tian. 2024. Out-of-Distribution Detection with Negative Prompts. In *The Twelfth International Conference on Learning Representations*.

[49] Aristotelis-Angelos Papadopoulos, Mohammad Reza Rajati, Nazim Shaikh, and Jiamian Wang. 2021. Outlier exposure with confidence control for out-of-distribution detection. *Neurocomputing* 441 (2021), 138–150.

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

[51] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 3498–3505.

[52] Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto. 2018. Generative probabilistic novelty detection with adversarial autoencoders. *Advances in neural information processing systems* 31 (2018).

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[54] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet?. In *International conference on machine learning*. PMLR, 5389–5400.

[55] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. 2021. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022* (2021).

[56] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. 2019. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems* 32 (2019).

[57] Chandramouli Shama Sastry and Sageev Oore. 2020. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning*. PMLR, 8491–8501.

[58] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. 2012. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence* 35, 7 (2012), 1757–1772.

[59] Yiyou Sun, Chuan Guo, and Yixuan Li. 2021. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems* 34 (2021), 144–157.

[60] Yiyou Sun and Yixuan Li. 2022. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*. Springer, 691–708.

[61] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. 2022. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*. PMLR, 20827–20840.

[62] Leitian Tao, Xuefeng Du, Xiaojin Zhu, and Yixuan Li. 2023. Non-parametric outlier synthesis. *arXiv preprint arXiv:2303.02966* (2023).

[63] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. 2019. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in neural information processing systems* 32 (2019).

[64] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. 2020. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*. PMLR, 9690–9700.

[65] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8769–8778.

[66] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. (2011).

[67] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. 2019. Learning Robust Global Representations by Penalizing Local Predictive Power. In *Advances in Neural Information Processing Systems*. 10506–10518.

[68] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. 2023. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1802–1812.

[69] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. 2022. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4921–4930.

[70] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. 2022. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*. PMLR, 23631–23644.

[71] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 3485–3492.

[72] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. 2022. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems* 35 (2022), 32598–32611.

[73] Geng Yu, Jianing Zhu, Jiangchao Yao, and Bo Han. 2024. Self-Calibrated Tuning of Vision-Language Models for Out-of-Distribution Detection. *Advances in Neural Information Processing Systems* 37 (2024), 56322–56348.

[74] Qing Yu and Kiyoharu Aizawa. 2019. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9518–9526.

[75] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. 2023. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10899–10909.

[76] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6023–6032.

[77] Alireza Zaeemzadeh, Niccolo Bisagno, Zeno Sambugaro, Nicola Conci, Nazanin Rahnavard, and Mubarak Shah. 2021. Out-of-distribution detection using union of 1-dimensional subspaces. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 9452–9461.

[78] Maxime Zanella and Ismail Ben Ayed. 2024. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1593–1603.

[79] Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Shi Han, Dongmei Zhang, et al. 2022. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *The Eleventh International Conference on Learning Representations*.

[80] Yabin Zhang and Lei Zhang. 2024. Adaneg: Adaptive negative proxy guided ood detection with vision-language models. *Advances in Neural Information Processing Systems* 37 (2024), 38744–38768.

[81] Yabin Zhang, Wenjie Zhu, Chenhang He, and Lei Zhang. 2024. Lapt: Label-driven automated prompt tuning for ood detection with vision-language models. *arXiv preprint arXiv:2407.08966* (2024).

[82] Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. 2024. Dual memory networks: A versatile adaptation approach for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 28718–28728.

[83] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2017), 1452–1464.

[84] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16816–16825.

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

[86] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. 2018. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*.