A Novel Fine-Tuned CLIP-OOD Detection Method with Double Loss Constraint Through Optimal Transport Semantic Alignment

Heng-yang Lu[1*†], Xin Guo[1*], Shuai Feng[2], Wenyu Jiang[3], Yuntao Du[3, 4], Chang Xia[5], Chenyou Fan[6]

[1]Engineering Research Center of Intelligent Technology for Healthcare of Ministry of Education, School of Artificial Intelligence and Computer Science, Jiangnan University,
[2]College of Smart Agriculture (College of Artificial Intelligence), Nanjing Agricultural University,
[3]State Key Laboratory for Novel Software Technology, Nanjing University,
[4]C-FAIR&School of Software, Shandong University,
[5]Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University,
[6]School of Artificial Intelligence, South China Normal University
luhengyang@jiangnan.edu.cn, guoxin@stu.jiangnan.edu.cn, shuaifeng@njau.edu.cn, lygjwy@smail.nju.edu.cn, yuntaodu@sdu.edu.cn, 23104@ahu.edu.cn, fanchenyou@scnu.edu.cn

Abstract

Detecting Out-Of-Distribution (OOD) samples in image classification is crucial for model reliability. With the rise of Vision-Language Models (VLMs), CLIP-OOD has become a research hotspot. However, we observe the Low Focus Attention phenomenon from the image encoders of CLIP, which means the attention of image encoders often spreads to non-in-distribution regions. This phenomenon comes from the semantic misalignment and inter-class feature confusion. To address these issues, we propose a novel fine-tuned OOD detection method with the Double loss constraint based on Optimal Transport (DOT-OOD). DOT-OOD integrates the Double Loss Constraint (DLC) module and Optimal Transport (OT) module. The DLC module comprises the Aligned Image-Text Concept Matching Loss and the Negative Sample Repulsion Loss, which respectively (1) focus on the core semantics of ID images and achieve cross-modal semantic alignment, (2) expand inter-class distances and enhance discriminative. While the OT module is introduced to obtain enhanced image feature representations. Extensive experimental results show that in the 16-shot scenario of the ImageNet-1k benchmark, DOT-OOD reduces the FPR95 by over 10% and improves the AUROC from 94.48% to 96.57% compared with SOTAs.

Code — https://github.com/jncsnlp/DOT-OOD

Introduction

When a well-trained image classification model is deployed in practice, it often encounters Out-Of-Distribution (OOD) samples that differ from the distribution of training data (i.e., categories not seen during model training, while the training involves In-Distribution (ID) samples) (Tao et al. 2022; Du et al. 2022; Xiang, Zhang, and Chen 2024). Misclassifying OOD samples as ID can lead to serious consequences, especially in fields such as autonomous driving and medical diagnosis (Dong et al. 2022; Karimi and Gholipour 2022; Li et al. 2024a). Therefore, accurately detecting OOD samples is crucial for the practical application of AI models. In recent years, the rapid development of visual-language pre-trained models (VLMs) such as CLIP (Radford et al. 2021) has inspired researchers to explore multi-modal OOD detection methods (Ming et al. 2022; Bai et al. 2024; Zeng et al. 2025). Their unique advantage of establishing cross-modal associations between visual and textual semantics provides a strong foundation for distinguishing OOD and ID samples.

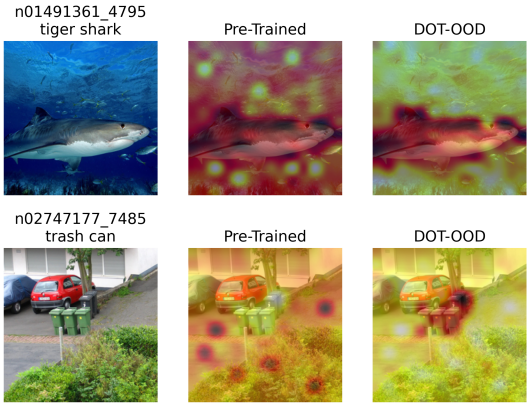

Figure 1: Attention heatmap (Grad-CAM (Selvaraju et al. 2017)) generated by the CLIP-B/16 model on ID images. The more reddish the color in the map, the higher the model's attention to the corresponding region.

Ideally, a well-trained VLM, such as CLIP, should anchor its attention strictly to the core feature regions of ID images. However, as shown in the middle column of Figure 1, the CLIP image encoder exhibits diffused attention in practice, spreading beyond semantically critical regions to non-in-distribution (non-ID) regions (background noise, irrelevant objects, blurred boundaries, etc.). We name this

*These authors contributed equally.
†Corresponding author.

non-targeted attention distribution phenomenon as Low Focus Attention, which comes from **(1) Semantic misalignment**. The image encoder's attention deviates from the core semantic regions of ID categories and spreads to non-target regions, directly causing misalignment between image features and text semantics, thus forming semantic misalignment. **(2) Inter-class feature confusion**. Non-target regions may contain similar visual elements to other ID categories. These elements trigger cross-category feature confusion and compress inter-class feature distances, which leads to inter-class feature confusion.

Thus, these two issues hinder the learning of discriminative feature representations, causing OOD detection failure. Semantic misalignment leads to redundant background noise disrupting key feature extraction, causing similar-background OOD samples to enter ID clusters and raising misclassification. Inter-class feature confusion allows OOD samples to "sneak into" ID feature spaces via shared semantics, blurring boundaries and exacerbating misclassification.

To address the semantic misalignment and inter-class feature confusion issues, we propose a novel fine-tuned OOD detection method with the Double loss constraint based on Optimal Transport (DOT-OOD). DOT-OOD integrates the Double Loss Constraint (DLC) module and Optimal Transport (OT) module. The DLC module comprises two loss functions: aligned image-text concept matching loss and negative samples repulsion loss. Specifically, the first loss aims to achieve precise cross-modal semantic alignment between image and text. Thus, ensuring that image features and text features are consistent in the semantic space. Meanwhile, the second loss reduces inter-class feature confusion by maximizing the feature distance between different categories. Building on the DLC module, to obtain enhanced image feature representation, we introduce the Optimal Transport (OT) (Cuturi 2013; Montesuma, Mboula, and Souloumiac 2025) to optimize feature distribution. This collaborative optimization strengthens the discriminative boundaries between ID and OOD, thereby boosting the model's OOD detection performance. As shown in Figure 1's right column, the DOT-OOD model activates core ID regions in red, with weak responses in non-ID regions. Our main contributions include:

1. We observe the Low Focus Attention phenomenon, which is from the semantic misalignment and inter-class feature confusion, resulting in OOD detection failure.
2. We propose DOT-OOD, which integrates the Double Loss Constraint module and Optimal Transport module to fine-tune CLIP for sharpening the discriminative boundaries between ID and OOD.
3. Extensive experiments in the few-shot scenario demonstrate that DOT-OOD has surpassed SOTA methods, enhancing OOD detection performance.

Related Work

Pre-trained Vision-Language Models

In recent years, pre-trained VLMs have become a research hotspot in AI, achieving breakthrough performance in cross-modal understanding tasks. Typical models such as CLIP (Radford et al. 2021), FILIP (Yao et al. 2022), BLIP (Li et al. 2022), and ALIGN (Jia et al. 2021) have shown excellent performance in various downstream tasks through training with massive image-text pairs. The core mechanism of these models is the adoption of contrastive learning (Aaron, Li, and Vinyals 2018; Khosla et al. 2020). This mechanism achieves semantic alignment between image and text representations by maximizing feature similarity of matched samples and minimizing that of unmatched ones.

OOD detection with VLMs

Zero-Shot OOD Detection. In zero-shot OOD detection (where ID images are unused in both training and inference (Miyai et al. 2024a)), researchers have developed several methods. For instance: MCM (Ming et al. 2022) separates ID/OOD samples via Softmax-calculated maximum matching between text and global image representations. GL-MCM (Miyai et al. 2023) enhances this by incorporating local image features into matching. NegLabel (Jiang et al. 2024) expands the label space with negative labels from large corpora, fusing image-label similarity for OOD scoring. CSP (Chen, Gao, and Xu 2024) boosts performance by constructing a semantic pool of modified superclass names to expand OOD candidate labels. CMA (Lee et al. 2025) introduces neutral prompts as agent, forming a vector triangular relationship among ID labels, inputs, and the agent to enhance robustness. Existing zero-shot OOD methods optimize pattern matching, label expansion, or prompts within the inherent framework of pre-trained models, whereas our approach focuses on addressing the semantic misalignment issue in VLMs.

Few-Shot OOD Detection. In few-shot OOD detection (where only a small number of ID images are used in training or inference (Miyai et al. 2024a)), researchers have proposed several innovative methods: LoCoOp (Miyai et al. 2024b) treats CLIP's ID-irrelevant local features as OOD features, pushing them away from ID text embeddings via entropy maximization to enhance ID-OOD separation. IDLike (Bai et al. 2024) discovers in-class outliers near ID features using CLIP, combining prompt learning to optimize OOD detection with minimal ID samples. NegPrompt (Li et al. 2024b) learns transferable negative prompts from ID data alone (based on CLIP), using joint optimization of three losses for few-shot and open-vocabulary OOD detection. LSN (Nie et al. 2024) learns class-specific negative prompts, integrating positive prompts to measure feature similarity and dissimilarity for better accuracy. SCT (Yu et al. 2024) introduces a modulation factor based on prediction uncertainty to adaptively adjust ID classification and OOD regularization, calibrating biases in OOD features from ID data. Local-Prompt (Zeng et al. 2025) freezes global prompts while adding local ones, combining global-guided negative enhancement and local regional regularization for coarse-to-fine tuning via local outlier knowledge. Existing few-shot OOD methods mainly optimize via prompt tuning or feature separation strategies based on limited ID data, whereas ours targets and resolves semantic misalignment in VLMs to enhance ID-OOD discriminability.

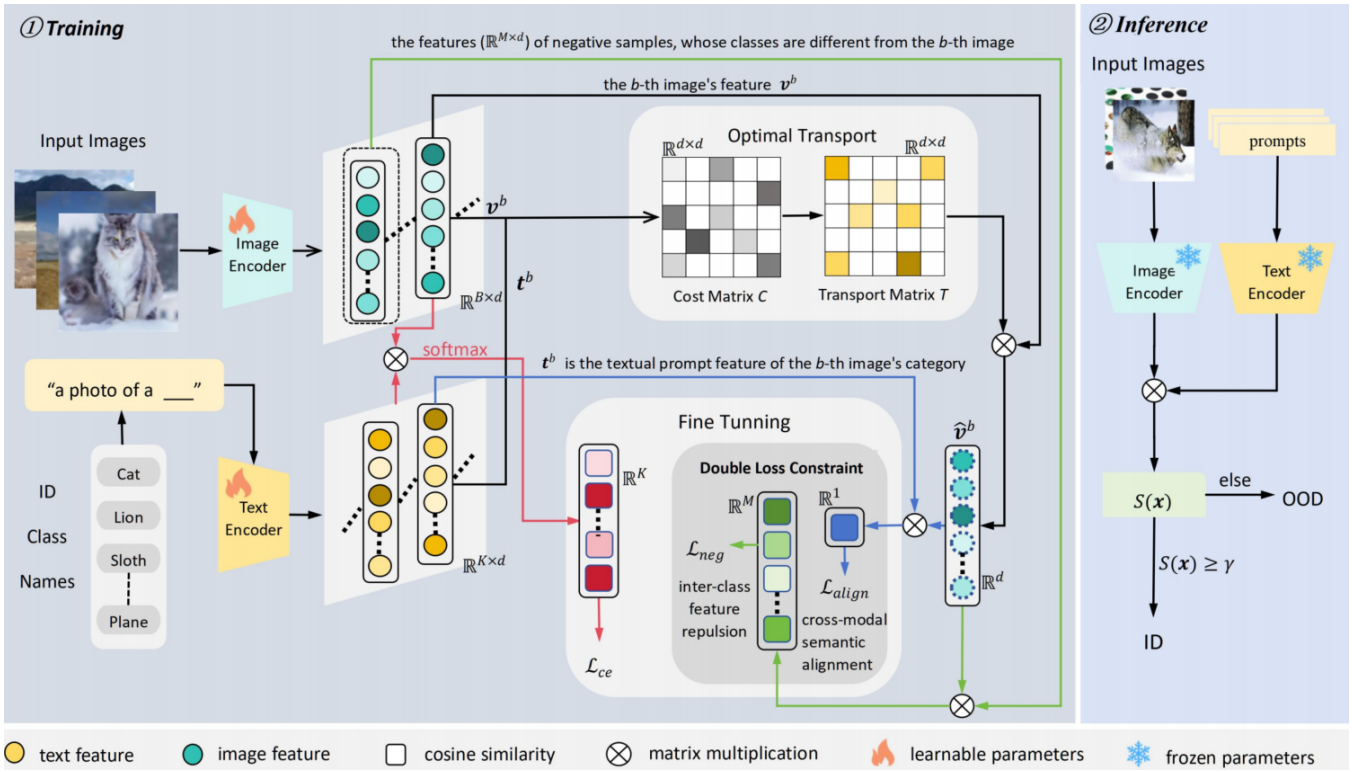

Figure 2: Overall framework of DOT-OOD. DOT-OOD integrates the OT module and the DLC module, with cross-entropy loss involved, for model fine-tuning collaboratively. OT achieves preliminary cross-modal feature alignment. $\mathcal{L}_{ce}$ secures ID classification accuracy, $\mathcal{L}_{align}$ reinforces cross-modal semantic consistency, and $\mathcal{L}_{neg}$ repulse inter-class feature.

Preliminaries

Few-shot out-of-distribution detection

OOD detection can be treated as a binary classification task, where the detector classifies input images into either ID or OOD categories. Training data is $\mathcal{D}^{\text{id}}_{\text{train}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, where $\boldsymbol{x}_i$ is the i-th image, $y_i \in \mathcal{Y}^{\text{id}}$ is the class label, $\mathcal{Y}^{\text{id}} \in \mathbb{R}^K$ is the ID class label space, K is the number of ID classes, and N is the number of training data. Under the few-shot OOD setting, a small number of images in each class (e.g., 1, 2, 4, or 16) are selected to fine-tune pre-trained models. $\mathcal{Y}^{\text{ood}}$ is the label space of OOD data, which is disjoint from the ID sample classes, i.e., $\mathcal{Y}^{\text{in}} \cap \mathcal{Y}^{\text{ood}} = \emptyset$. The model cannot obtain any OOD data during training.

Test data consists of $\mathcal{D}^{\text{in}}_{\text{test}}$ and $\mathcal{D}^{\text{ood}}_{\text{test}}$. During testing, the OOD detector determines whether each input sample $\boldsymbol{x}$ belongs to ID or OOD by performing binary classification:

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

where $S(\boldsymbol{x})$ is the OOD scoring function, and γ is a threshold that ensures most (e.g., 95%) of the ID samples are correctly classified as ID (Hendrycks and Gimpel 2017; Yu et al. 2024; Zeng et al. 2025).

Prompt Learning with CLIP

The CLIP model consists of an image encoder f_{img} and a text encoder f_{txt}. In prompt learning, textual prompts are constructed via: handcrafted prompt templates (Radford et al. 2021; Zeng et al. 2025), or design learnable context words (Zhou et al. 2022; Zeng et al. 2025).

In image classification tasks, given $(\boldsymbol{x}, y) \in \mathcal{D}^{\text{id}}_{\text{train}}$, $\boldsymbol{v} = f_{\text{img}}(\boldsymbol{x})$ is the image feature obtained via CLIP's image encoder, and $\boldsymbol{t} = f_{\text{txt}}(\mathcal{T}(\hat{y})$ is the text prompt feature extracted by CLIP's text encoder. $(\mathcal{T}(\hat{y})$ is the text prompt constructed from the class name $\hat{y}$ of class y. For the image $\boldsymbol{x}$, the predicted probability p_{y_i} that image $\boldsymbol{x}$ belongs to class y_i is:

$$p_{y_i}(\boldsymbol{x}, \mathcal{Y}^{\text{id}}) = \frac{e^{\text{sim}\left(f_{\text{txt}}(\mathcal{T}(\hat{y}_i)), f_{\text{img}}(\boldsymbol{x})\right)/\tau}}{\sum_{c\in\mathcal{Y}^{\text{id}}} e^{\text{sim}\left(f_{\text{txt}}(\mathcal{T}(\hat{y}_c)), f_{\text{img}}(\boldsymbol{x})\right)/\tau}}, \tag{2}$$

where τ is the temperature, $\text{sim}(\cdot)$ is the cosine similarity.

Method

Overview

The overall framework of DOT-OOD (Figure 2) integrates two core modules: OT and DLC. These modules collaboratively optimize models. Test images are classified as ID or OOD via threshold judgment based on $S(\boldsymbol{x})$ computed by the outputs from well-trained models. Their specific implementation details will be explained subsequently.

Optimal Transport Module

OT Module maps image features to the text semantic space by minimizing transportation costs, achieving preliminary semantic alignment of cross-modal image and text features, and obtaining enhanced image feature representation.

The process based on OT is defined as follows:

Cost Matrix Construction. The cost matrix C measures the semantic distance between image and text features. Since both the image and text features have been L2-normalized, the element $\left(\boldsymbol{v}\boldsymbol{t}^\top\right)_{i,j}$ (i.e., $\boldsymbol{v}_i \cdot \boldsymbol{t}_j$) in the outer product matrix can be directly interpreted as the cosine similarity between the i-th and j-th components of $\boldsymbol{v}$ and $\boldsymbol{t}$. Specifically, the cost matrix element is defined as:

$$C_{i,j} = 1 - \left(\boldsymbol{v}\boldsymbol{t}^\top\right)_{i,j} = 1 - \boldsymbol{v}_i \cdot \boldsymbol{t}_j. \tag{3}$$

This cost matrix C is designed based on three considerations: (1) Cosine similarity, which captures vector semantic relevance and quantifies image-text consistency, guides OT's optimal distribution alignment. (2) Converting similarity to distance via $1 - \text{sim}(\cdot)$ fits OT's logic for low-cost matches (smaller values mean higher similarity). (3) Minimizing transport cost enables OT to prioritize aligning semantically close elements and enlarge distances between distant ones, thus achieving cross-modal alignment.

Transport Matrix. The transport matrix T is obtained by minimizing the Entropy-Regularized OT in Eq. 4:

$$\begin{aligned} \text{rOT}(\mu, \nu, C) := \min_{T} &\langle T, C\rangle - \epsilon \cdot \text{H}(T) \\ \text{s.t.} \quad &\{T \in \mathbb{R}_+^{d\times d} \mid T\mathbf{1}_d = \mu,\ T^\top \mathbf{1}_d = \nu\}, \end{aligned} \tag{4}$$

where $\langle T, C\rangle := \text{tr}(T^\top C)$ is defined as the Frobenius inner product of matrices, ϵ is a regularization parameter that controls the strength of the entropy term, d is the dimension of image/text features, $\mathbf{1}_d$ signifies an all-one vector of dimension d, and μ and ν are the probability simplex measures corresponding to the image/text features. In the experiment, to ensure equal weighting across feature dimensions, we set μ and ν as uniform probability measures: $\mu = \frac{1}{d}\mathbf{1}_d$ and $\nu = \frac{1}{d}\mathbf{1}_d$. $\text{H}(T)$ is the entropy, computed by:

$$\text{H}(T) = -\sum_{i,j} T_{ij} \log T_{ij}. \tag{5}$$

Entropy-regularized OT obtains T by introducing an entropy term, which can optimize semantic alignment, reduce computational complexity, and avoid extreme distributions of T (Cuturi 2013; Fan, Hu, and Huang 2023).

Feature Enhancement and Normalization. By leveraging T obtained in Eq. 4, we can align the original image features with the text semantics, thereby obtaining the preliminary enhanced image feature $\boldsymbol{v}_e$. The formula is:

$$\boldsymbol{v}_e = T^\top \boldsymbol{v}. \tag{6}$$

To ensure geometric consistency, we normalize the enhanced feature to a unit hypersphere:

$$\hat{\boldsymbol{v}} = \frac{\boldsymbol{v}_e}{\|\boldsymbol{v}_e\|_2}. \tag{7}$$

In this part, we propose an OT-based image feature enhancement method. It enables preliminary alignment between an image feature and the text prompt feature through image-text feature mapping, yielding the enhanced image feature $\hat{\boldsymbol{v}}$.

Fine Tunning

To fine-tune the model, we propose two novel loss functions based on enhanced image features obtained by the OT module alongside the cross-entropy loss.

In-distribution Loss. To ensure the classification performance on ID, we use the standard cross-entropy loss to measure the discrepancy between the predicted label probabilities and the ground-truth labels of ID samples:

$$\mathcal{L}_{ce} = \mathbb{E}_{(\boldsymbol{x},y)\sim D^{id}_{train}}\big[-\log p_{y_i}(\boldsymbol{x}, \mathcal{Y}^{\text{id}}, f_{\text{txt}}, f_{\text{img}})\big]. \tag{8}$$

Double Loss Constraint Module. To mitigate **semantic misalignment** and **inter-class feature confusion**, we propose two targeted losses formulated on $\hat{\boldsymbol{v}}$:

(1) Aligned Image-Text Concept Matching Loss. While the OT module enables structured mapping between feature spaces, fine-grained semantic consistency between the enhanced image features and the target text features remains to be strengthened. Thus, we propose the *Aligned Image-Text Concept Matching Loss* $\mathcal{L}_{align}$, which further constrains semantically relevant Image-Text concepts to be closer. This can benefit ID samples to form compact clusters that OOD samples cannot fit into.

$$\mathcal{L}_{align} = \frac{1}{B}\sum_{b=1}^{B}\big(1 - \text{sim}(\boldsymbol{t}^b, \hat{\boldsymbol{v}}^b)\big), \tag{9}$$

where B denotes the number of samples in a batch, $\hat{\boldsymbol{v}}^b$ is the b-th enhanced image feature in the batch, $\boldsymbol{t}^b$ is the textual prompt feature of the b-th image's category. The $\mathcal{L}_{align}$ works by forcing core ID feature learning to suppress background noise, ensuring the learning of ID-specific semantic representations (e.g., object components).

(2) Negative Samples Repulsion Loss. While $\mathcal{L}_{align}$ reduces the distance between the enhanced image feature and its class text feature, it fails to differentiate distinct class features explicitly. To avoid inter-class confusion from semantic overlap, we propose the *Negative Samples Repulsion Loss* $\mathcal{L}_{neg}$ to constrain inter-class similarity:

$$\mathcal{L}_{neg} = \frac{1}{M}\sum_{b=1}^{B}\sum_{m=1}^{M}|\text{sim}(\hat{\boldsymbol{v}}^b, \boldsymbol{v}^m_{neg})|, \tag{10}$$

where $\boldsymbol{v}^m_{neg}$ is the m-th negative sample feature (same-batch images of different classes from $\hat{\boldsymbol{v}}^b$), and M denotes the total number of negative samples in the batch. By minimizing inter-class similarity, $\mathcal{L}_{neg}$ maximizes their distance, enhancing ID feature discriminability and indirectly widening the "semantic gap" between ID clusters. This makes OOD samples more likely to fall into these gaps and be detected as OOD, thereby creating a larger decision margin for OOD and strengthening the discriminative boundary.

Training Objective. The overall loss for fine-tuning models balances above losses through hyperparameters α and β:

$$\mathcal{L} = \mathcal{L}_{ce} + \alpha \cdot \mathcal{L}_{align} + \beta \cdot \mathcal{L}_{neg}. \quad (11)$$

The training objective forms a collaboration of "classification supervision, semantic alignment, and inter-class feature repulsion". $\mathcal{L}_{ce}$ is the common classification training loss. $\mathcal{L}_{align}$ is designed to promote the cross-modal semantic consistency, which suppresses non-ID noise and condenses ID features into tight clusters. $\mathcal{L}_{neg}$ is designed to maximize the ID inter-class distances, which carves semantic gaps between ID and OOD. The collaborative optimization of three losses ensures ID classification performance and sharpens the decision boundary between ID and OOD, thereby enhancing OOD detection performance.

Inference

For the ID classification task, we follow the strategy of CLIP: the category corresponding to the maximum probability after temperature-scaled softmax of image-text cosine similarity. Considering that the core positioning of DOT-OOD is to fine-tune models, and it does not specifically design an OOD score for OOD detection, we directly adopt the CSP scoring function proposed in (Chen, Gao, and Xu 2024) to distinguish abnormal samples in the OOD detection task.

Experiments

Experimental Setup

Datasets. Following common benchmarks in prior works, we utilize ImageNet-1K (Deng et al. 2009) as the ID dataset. For OOD datasets, we select iNaturalist, SUN, Places, and Texture, which encompass diverse class scenes ranging from natural organisms to man-made environments. In line with methodologies (Ming et al. 2022; Yu et al. 2024; Zeng et al. 2025), we further conducted hard-OOD experiments by extracting subsets from ImageNet-1K: ImageNet-10 and ImageNet-20.

OOD Detection Baselines. We conduct a comprehensive comparison of the DOT-OOD method against a series of few-shot tuning-based OOD detection methods: CoOp (Zhou et al. 2022), LoCoOp (Miyai et al. 2024b), ID-Like (Bai et al. 2024), NegPrompt (Li et al. 2024b), LSN (Nie et al. 2024), SCT (Yu et al. 2024), and Local-Prompt (Zeng et al. 2025). Following SCT, we report the results of the combination of LSN and LoCoOp in Table 1. And Local-Prompt in Table 1 is combined with LoCoOp.

Implementation Details. Following the setup of prior works (Ming et al. 2022; Yu et al. 2024; Zeng et al. 2025), we employ CLIP-B/16 as the backbone for the main experiments. In the few-shot learning scenario, we randomly sample 1, 2, 4, and 16 samples per class for model fine-tuning. Hyperparameters ϵ, α, and β are set to 0.3, 0.5, and 0.5, respectively. The temperature τ in training is 0.01.

Considering that shallow networks have effectively extracted low-level features of images and texts, and their parameters have been sufficiently optimized during pre-training, excessive fine-tuning may lead to catastrophic forgetting and overfitting (Song et al. 2023). Therefore, we choose to focus on fine-tuning the deep modules of the CLIP encoder as well as the image and text projection layers.

The initial learning rate is set to 2×10^{-3} with a cosine annealing schedule for dynamic adjustment, using SGD as the optimizer. Training configurations are as follows: batch size of 128 for the 16-shot setting and 32 for other shot sizes, with 15 training epochs across all experiments. All experiments are conducted on a single NVIDIA 4090, and reported results are averaged over three independent runs.

Evaluation Metrics. We adopt two common evaluation metrics to assess OOD detection performance (Hendrycks and Gimpel 2017; Yu et al. 2024; Zeng et al. 2025): (1) the area under the receiver operating characteristic curve **(AUROC)**; (2) false positive rate of OOD samples when true positive rate of ID samples is 95% **(FPR95)**; Additionally, we evaluate the model's classification accuracy on ID datasets using ID accuracy **(ID-ACC)** (Miyai et al. 2024b; Zhou et al. 2022; Zeng et al. 2025).

Main Results

We conducted experiments on the ImageNet-1k benchmark and hard OOD tasks to comprehensively evaluate the performance of the DOT-OOD method.

ImageNet-1k Benchmark. Here, we compare DOT-OOD with SOTA methods to verify DOT-OOD's effectiveness. Due to space limitations, we primarily present experimental results under the 16-shot setting.

The results in Table 1 reveal that, in the few-shot fine-tuning scenario, compared to the SOTA method Local-Prompt, DOT-OOD's average FPR95 decreases by over 10%, and the average AUROC increases by over 2%. The above results strongly confirm the effectiveness and superiority of DOT-OOD in the task of OOD sample detection.

Hard OOD Task. Hard-OOD tasks involve semantically similar categories, making them more challenging. Following previous works, we evaluate the performance of DOT-OOD in Hard-OOD tasks, with the results shown in Table 2. The results indicate that DOT-OOD achieved the best performance across all settings, except in the scenario where ImageNet-10 is used as ID and ImageNet-20 as OOD, where the AUROC is slightly lower than that of Local-Prompt by 0.06%. This demonstrates that DOT-OOD also possesses strong competitiveness in Hard-OOD tasks.

Ablations

In this part, we systematically analyze the effectiveness mechanism of the proposed DOT-OOD through multiple ablation studies. Due to space limitations, we only present the average results on four OOD datasets here.

Ablation Study on Loss. The results of fine-tuning the model using different losses (as shown in Table 3) reveal that, in comparison with the baseline model (which only uses $\mathcal{L}_{\text{ce}}$, the introduction of either $\mathcal{L}_{\text{align}}$ or $\mathcal{L}_{\text{neg}}$ can enhance both OOD detection performance and ID classification accuracy. When the three losses are jointly employed for fine-tuning, the OOD detection performance (with the

Method	OOD Datasets									
	iNaturalist		SUN		Places		Texture		AvG	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Few-shot tuning-based methods (16-shot)										
CoOp†	$28.25^{\pm 4.68}$	$93.92^{\pm 1.17}$	$31.15^{\pm 0.89}$	$93.13^{\pm 0.38}$	$39.12^{\pm 1.05}$	$90.50^{\pm 0.37}$	$41.86^{\pm 1.88}$	$90.40^{\pm 0.54}$	$35.09^{\pm 1.60}$	$91.99^{\pm 0.42}$
LoCoOp†	$17.58^{\pm 2.22}$	$96.30^{\pm 0.57}$	$22.82^{\pm 0.34}$	$95.20^{\pm 0.06}$	$32.21^{\pm 0.53}$	$92.03^{\pm 0.20}$	$45.27^{\pm 0.95}$	$88.86^{\pm 0.26}$	$29.47^{\pm 0.29}$	$93.10^{\pm 0.03}$
ID-like†	$9.71^{\pm 0.60}$	$98.05^{\pm 0.07}$	$38.93^{\pm 0.10}$	$90.54^{\pm 0.68}$	$47.06^{\pm 1.44}$	$88.06^{\pm 0.90}$	$32.82^{\pm 5.12}$	$91.89^{\pm 1.49}$	$32.12^{\pm 1.09}$	$92.14^{\pm 0.01}$
NegPrompt†	$37.79^{\pm 0.11}$	$90.49^{\pm 0.01}$	$32.11^{\pm 3.77}$	$92.25^{\pm 1.00}$	$35.52^{\pm 0.41}$	$91.16^{\pm 0.03}$	$43.93^{\pm 9.09}$	$88.38^{\pm 3.31}$	$37.34^{\pm 1.41}$	$90.57^{\pm 0.59}$
LSN†	$36.17^{\pm 4.81}$	$92.66^{\pm 1.16}$	$34.27^{\pm 0.44}$	$93.53^{\pm 0.20}$	$41.47^{\pm 0.85}$	$90.52^{\pm 0.37}$	$46.43^{\pm 0.60}$	$89.38^{\pm 0.24}$	$39.59^{\pm 0.73}$	$91.53^{\pm 0.09}$
SCT†	$13.94^{\pm 0.68}$	$95.86^{\pm 0.28}$	$20.55^{\pm 1.07}$	$95.33^{\pm 0.12}$	$29.86^{\pm 0.67}$	$92.24^{\pm 0.05}$	$41.51^{\pm 0.48}$	$89.06^{\pm 0.09}$	$26.47^{\pm 0.39}$	$93.37^{\pm 0.07}$
Local-Prompt*	8.63	98.07	23.23	95.12	31.74	92.42	34.50	92.29	24.52	94.48
DOT-OOD	$\mathbf{1.22}^{\pm 0.01}$	$\mathbf{99.71}^{\pm 0.01}$	$\mathbf{10.84}^{\pm 0.09}$	$\mathbf{97.37}^{\pm 0.03}$	$\mathbf{21.77}^{\pm 0.34}$	$\mathbf{94.25}^{\pm 0.06}$	$\mathbf{22.54}^{\pm 0.34}$	$\mathbf{94.95}^{\pm 0.06}$	$\mathbf{14.09}^{\pm 0.14}$	$\mathbf{96.57}^{\pm 0.03}$

Table 1: Comparison Results of ImageNet-1k Benchmark. We conducted experiments using the CLIP-B/16 network. Bold numbers are superior results. ↑ indicates that a larger value is better, and ↓ indicates that a smaller value is better. Results marked with † are from (Yu et al. 2024), and results marked with * are cited from the corresponding research papers.

Method	ID / OOD	ImageNet-10 / ImageNet-20		ImageNet-20 / ImageNet-10	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑
MCM		5.00	98.71	12.51	97.70
GL-MCM		10.10	98.04	9.00	98.62
NegLabel		5.00	98.80	11.60	98.04
CSP		3.30	99.02	3.40	98.79
LoCoOp		11.20	97.49	12.00	97.79
Local-Prompt		3.90	**99.06**	6.20	98.84
DOT-OOD		**3.10**	99.00	**3.07**	**99.06**

Table 2: Comparison results of hard OOD detection tasks. Bold numbers represent superior results.

average FPR95 reduced by more than 4% and AUROC increased over 1%) and ID-ACC (74.04%) are optimal. This verifies the synergistic optimization effect of loss functions on OOD detection and ID classification.

Loss	OOD		ID
	FPR95↓	AUROC↑	ACC↑
Baseline ($\mathcal{L}_{ce}$)	18.55	95.39	71.99
$+\mathcal{L}_{align}$	18.49	95.39	72.40
$+\mathcal{L}_{neg}$	15.90	96.21	73.86
$+\mathcal{L}_{align} + \mathcal{L}_{neg}$	**14.09**	**96.57**	**74.04**

Table 3: Impact of fine-tuning the CLIP model with different losses. Bold numbers represent superior results.

The effectiveness of the OT module. To verify the role of the OT module in DOT-OOD, we conducted an ablation experiment, with results in Table 4. In this experiment, when the OT module is removed (i.g, w/o OT), the enhanced image features $\hat{\boldsymbol{v}}$ in $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{neg}}$ are directly replaced with original image features $\boldsymbol{v}$. The results indicate that, without the OT module, the average FPR95 and AUROC, and ID-ACC are even worse than those of the baseline (using only $\mathcal{L}_{\text{ce}}$). Upon introducing OT, the three metrics improve by about 10%, 3%, and 5%, respectively. This demonstrates that the OT module effectively enhances both the OOD detection and ID classification performance of the model.

Method	OOD		ID
	FPR95↓	AUROC↑	ACC↑
Baseline ($\mathcal{L}_{ce}$)	18.55	95.39	71.99
w/o OT	23.66	94.04	69.39
w/ OT	**14.09**	**96.57**	**74.04**

Table 4: Impact of the OT module on DOT-OOD. Bold numbers represent superior results.

Fine-tuning deep modules of CLIP is valid. In the main experiment, we selectively fine-tune the deep modules of CLIP's image and text encoders and their projection layers instead of all modules. To verify this strategy, we conduct an ablation experiment, with results in Table 5. Comparing the first row (fine-tuning from middle layers) with the last row (fine-tuning from deep layers), the former has the smallest impact on OOD detection but harms ID accuracy. Additionally, training from the middle layers increases training costs. Fine-tuning only projection layers (the middle three rows: fine-tuning only the visual projection layer, the text projection layer, and both together) impairs both tasks. Therefore, fine-tuning deep modules instead of full-model fine-tuning enhances the model's semantic representation capabilities while preserving the stability of low-level features.

DOT-OOD Improves OOD Score Performance Across CLIP Models. To verify the effectiveness of DOT-OOD on different parameterized CLIP models, we compared MCM, GL-MCM, NegLabel, and CSP before and after integrating DOT-OOD on CLIP-B/16 and CLIP-L/14 (Table 6). Results show all scoring methods gain better OOD detection

Module	OOD		ID
	FPR95↓	AUROC↑	ACC↑
$\text{Encoder}^{\text{M}}$	14.10	96.56	73.66
$\text{Projection}_{\text{v}}$	17.04	95.98	70.31
$\text{Projection}_{\text{t}}$	18.85	95.54	71.96
$\text{Projection}_{\text{v,t}}$	16.98	95.77	71.85
$\text{Encoder}^{\text{D}}$	**14.09**	**96.57**	**74.04**

Table 5: Impact of fine-tuning different CLIP modules with DOT-OOD. Bold numbers represent superior results.

Method	CLIP-B/16		CLIP-L/14	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MCM	42.61	90.63	36.28	91.93
+ **DOT-OOD**	**32.93**	**92.14**	**32.14**	**92.06**
GL-MCM	35.06	90.90	35.73	91.51
+ **DOT-OOD**	**32.17**	**91.36**	**32.04**	**91.99**
NegLabel	25.40	94.21	24.70	94.69
+ **DOT-OOD**	**16.06**	**96.20**	**18.59**	**95.49**
CSP	17.51	95.76	23.35	94.90
+ **DOT-OOD**	**14.09**	**96.57**	**16.78**	**95.87**

Table 6: Performance comparison of different OOD scoring methods on different CLIP models before and after integrating DOT-OOD. Bold numbers represent superior results.

after integrating DOT-OOD, regardless of CLIP's parameter scale: FPR95 down 3%-11%, AUROC up 1%-2%. This confirms DOT-OOD's generalization in enhancing detection boundary division for scoring methods across CLIP models.

To demonstrate the optimization effect of DOT-OOD, we take CLIP-B/16's MCM score as an example (Figure 3): ID (ImageNet-1k) and OOD (iNaturalist) distributions are more concentrated, with sharper peaks, larger separation, and smaller overlaps. CLIP-L/14 shows similar gains (Table 6), clarifying boundaries, boosting OOD discrimination, and enabling more accurate detection.

Visualization

In this part, taking CLIP-B/16 as an example, we utilize Grad-CAM (Selvaraju et al. 2017) and T-SNE (Maaten and Hinton 2008) to present the effects of DOT-OOD visually.

Attention Distribution of Image Encoder. In Figure 1, we observe that after DOT-OOD fine-tuning, the attention of CLIP's image encoder is more concentrated on ID regions, with reduced focus on non-ID regions. This enhances the model's capacity to extract key features of ID samples, strengthening the boundary between ID and OOD.

ID Class Feature Clustering Distribution. In Figure 4, we visualize ID image feature clustering distributions (T-SNE (Maaten and Hinton 2008)) before (left) and after (right) DOT-OOD fine-tuning. The ID dataset of the first

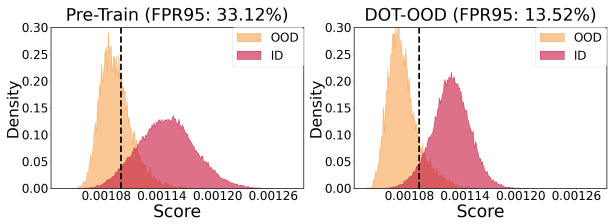

Figure 3: Comparison of MCM score distributions for ID and OOD before (left) and after (right) integrating DOT-OOD. The grey vertical line indicates the threshold that enables 95% of the ID samples to be correctly classified as ID.

and second row is ImageNet-10 and ImageNet-20, respectively. After DOT-OOD fine-tuning, the image features exhibit: Intra-class features aggregate more compactly, and inter-class features separate more clearly, leaving more semantic gaps. These gaps make OOD samples easier to fall into and thus be correctly detected as OOD, reducing feature confusion and enhancing OOD detection performance.

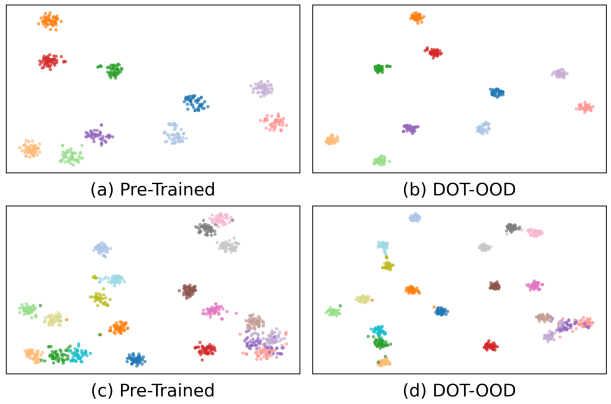

Figure 4: The clustering distributions of ID image features before (a,c) and after (b,d) DOT-OOD fine-tuning. Each color corresponds to a different class.

Conclusion

This paper targets the Low Focus Attention phenomenon in Vision-Language Models. This phenomenon is caused by semantic misalignment and inter-class feature confusion. To address these issues, we propose the DOT-OOD method, combining the Double Loss Constraint module and Optimal Transport module. This integration enhances the model's feature extraction and ID/OOD discriminative abilities. In the 16-shot scenario of the ImageNet-1k benchmark, DOT-OOD outperforms SOTA methods, with FPR95 reduced by over 10% and AUROC reaching 96.57%, confirming DOT-OOD gains in OOD detection performance. Ablation experiments quantify each module's contribution, validating DOT-OOD's rational design and effectiveness.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (Grant No.62002137), in part by the China Postdoctoral Science Foundation (Grant No. 2022M711360), Shandong Provincial Natural Science Foundation (ZR2025QC1570), and the National Key Laboratory of Ship Structural Safety (Grant No. 450324300).

References

Aaron, v. d. O.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Bai, Y.; Han, Z.; Cao, B.; Jiang, X.; Hu, Q.; and Zhang, C. 2024. ID-like Prompt Learning for Few-Shot Out-of-Distribution Detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17480–17489.

Chen, M.; Gao, J.; and Xu, C. 2024. Conjugated Semantic Pool Improves OOD Detection with Pre-trained Vision-Language Models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2292–2300.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5–6.

Dong, H.; Zhang, X.; Xu, J.; Ai, R.; Gu, W.; Lu, H.; Kannala, J.; and Chen, X. 2022. Superfusion: Multilevel Lidar-Camera Fusion for Long-Range HD Map Generation. arXiv:2211.15656.

Du, Y.; Wei, F.; Zhang, Z.; Shi, M.; Gao, Y.; and Li, G. 2022. Learning to Prompt for Open - Vocabulary Object Detection with Vision - Language Model. In *CVPR*.

Fan, C.; Hu, J.; and Huang, J. 2023. Few-Shot Multi-Agent Perception With Ranking-Based Feature Learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10): 11810–11823.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 3.

Jiang, X.; Liu, F.; Fang, Z.; Chen, H.; Liu, T.; Zheng, F.; and Han, B. 2024. Negative label guided ood detection with pretrained vision-language models. In *The Twelfth International Conference on Learning Representations*.

Karimi, D.; and Gholipour, A. 2022. Improving Calibration and Out-of-Distribution Detection in Deep Models for Medical Image Segmentation. *IEEE Transactions on Artificial Intelligence*, 4(2): 383–397.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, 18661–18673.

Lee, Y.; Cao, X.; Guo, J.; Ye, W.; Guo, Q.; and Chang, Y. 2025. Concept Matching with Agent for Out-of-Distribution Detection. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 4562–4570. AAAI Press.

Li, J.; Li, B.; Tu, Z.; Liu, X.; Guo, Q.; Xu, F. J.; Xu, R.; and Yu, H. 2024a. Light the Night: A Multi - Condition Diffusion Framework for Unpaired Low - Light Enhancement in Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15205–15215.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 12888–12900. PMLR.

Li, T.; Pang, G.; Bai, X.; Miao, W.; and Zheng, J. 2024b. Learning Transferable Negative Prompts for Out-of-Distribution Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605.

Ming, Y.; Cai, Z.; Gu, J.; Sun, Y.; Li, W.; and Li, Y. 2022. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems*.

Miyai, A.; Yang, J.; Zhang, J.; Ming, Y.; Lin, Y.; Yu, Q.; Irie, G.; Joty, S.; Li, Y.; Li, H.; Liu, Z.; Yamasaki, T.; and Aizawa, K. 2024a. Generalized Out-of-Distribution Detection and Beyond in Vision Language Model Era: A Survey. *arXiv preprint arXiv:2407.21794*.

Miyai, A.; Yu, Q.; Irie, G.; and Aizawa, K. 2023. Zero-shot in-distribution detection in multi-object settings using vision-language foundation models. *arXiv preprint arXiv:2304.04521*.

Miyai, A.; Yu, Q.; Irie, G.; and Aizawa, K. 2024b. Locoop: Few-shot Out-of-Distribution Detection via Prompt Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Montesuma, E. F.; Mboula, F. N.; and Souloumiac, A. 2025. Recent Advances in Optimal Transport for Machine Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2): 1161–1180. Epub 2025 Jan 9.

Nie, J.; Zhang, Y.; Fang, Z.; Liu, T.; Han, B.; and Tian, X. 2024. Out-of-Distribution Detection with Negative Prompts. In *Proceedings of the Twelfth International Conference on Learning Representations, ICLR 2024*, Proceedings of the International Conference on Learning Representations,

ICLR, 1–20. International Conference on Learning Representations.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 1.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626.

Song, K.; Ma, H.; Zou, B.; Zhang, H.; and Huang, W. 2023. FD-Align: Feature Discrimination Alignment for Fine-tuning Pre-Trained Models in Few-Shot Learning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 43579–43592. Curran Associates, Inc.

Tao, L.; Du, X.; Zhu, J.; and Li, Y. 2022. Non-parametric Outlier Synthesis. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*.

Xiang, X.; Zhang, Z.; and Chen, X. 2024. Curricular-Balanced Long-Tailed Learning. *Neurocomputing*, 571: 127121.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2022. FILIP: fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 3.

Yu, G.; Zhu, J.; Yao, J.; and Han, B. 2024. Self-Calibrated Tuning of Vision-Language Models for Out-of-Distribution Detection. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 56322–56348. Curran Associates, Inc.

Zeng, F.; Cheng, Z.; Zhu, F.; Wei, H.; and Zhang, X.-Y. 2025. Local-Prompt: Extensible Local Prompts for Few-Shot Out-of-Distribution Detection. In *The Thirteenth International Conference on Learning Representations*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348. 3, 4, 5, 6.