Journal of Intelligent Computing and Networking
https://www.ffspub.com/index.php/jicn/index
ISSN 3079-9228 (print)
E-mail: jicn.office@ffspub.com

Article

Likelihood-Aware Semantic Alignment for Full-Spectrum Out-of-Distribution Detection

Fan LU, Kai ZHU, Wei ZHAI, Yang CAO, Zheng-Jun ZHA†
Department of Automation, University of Science and Technology of China, Hefei 230026, China
†*E-mail: zhazj@ustc.edu.cn*

Received: May 22, 2025 / Revised: July 6 2025 / Accepted: July 7, 2025 / Published online: July 18, 2025

Abstract: Full-spectrum out-of-distribution (F-OOD) detection aims to accurately recognize in-distribution (ID) samples while encountering semantic and covariate shifts simultaneously. However, existing out-of-distribution (OOD) detectors tend to overfit the covariance information and ignore intrinsic semantic correlation, inadequate for adapting to complex domain transformations. To address this issue, we propose a Likelihood-Aware Semantic Alignment (LSA) framework to promote the image-text alignment into semantically high-likelihood regions. LSA consists of an offline Gaussian sampling strategy which efficiently samples semantic-relevant visual embeddings from the class-conditional Gaussian distribution, and a bidirectional prompt customization mechanism that adjusts ID-related and negative context for a discriminative ID/OOD boundary. Extensive experiments demonstrate the remarkable OOD detection performance of our proposed LSA, especially on the intractable Near-OOD setting, surpassing existing methods by a margin of 15.26% and 18.88% on two F-OOD benchmarks, respectively.

Keywords: Out-of-distribution detection; multimodal learning; prompt tuning
https://doi.org/10.64509/jicn.11.10

1 Introduction

Deep visual models [1, 2] have demonstrated remarkable performance in closed-set environments. However, their performance significantly deteriorates when faced with out-of-distribution (OOD) samples in real-world scenarios, such as input from unknown classes [3]. To enhance the deployment security, OOD detection has received increasing research interest recently [4–7].

A rich line of OOD detection methods [6, 8–12] involves one in-distribution (ID) dataset in training while regarding all other datasets as OOD, where models tend to overfit the low-level covariate shift while ignoring the inherent semantic correlation across different datasets [13–15]. For instance, arbitrarily considering pet dogs and cartoon dogs as separate categories is unjust, as both share the semantic concept of 'dog', despite displaying distinct covariance information. Disregarding this fact in realistic scenarios can give rise to irreparable interference.

To comprehensively evaluate covariate shift and semantic shift simultaneously, the realistic F-OOD detection benchmarks are proposed in OpenOODv1.5 [15]. F-OOD benchmarks split OOD data into Near-OOD and Far-OOD based on their degrees of covariate shift, and introduce covariate-shifted ID (csID) data that retains ID semantics while being from different domains. Regrettably, the experimental evidence [15] proves that numerous existing OOD detectors have experienced notable performance decline on the F-OOD benchmarks. We analyze this phenomenon in Figure 1(a) and reveal that classical OOD detectors like EBO [8] assign lower confidence to portions of ID-related (*i.e.*, csID and ID) data than to OOD, further failing to uniformly aggregate them. Although several methods [16–19] leverage the generalization ability of CLIP models [20–22] to enhance the perception of ID semantics, we observe that naively incorporating it into F-OOD still fails to distinguish between OOD and csID clusters, which significantly overlap in low-confidence regions, as revealed in Figure 1(b). We illustrate the likelihood characteristics of CLIP image embedding space in Figure 2(a), that high-likelihood ID or csID samples which are close to class centers, share a more compact and consistent distribution despite the covariate shift introduced by csID. Moreover, as shown in Figure 2(b), the semantic-irrelevant textual context (*e.g.*, standing on grass) disrupts the CLIP models' alignment with intrinsic ID semantics. Motivated by this, we consider

† Corresponding author: Zheng-Jun Zha
*Academic Editor: Chunxiao Jiang

enhancing the ID image-text alignment of high-likelihood space to resilience against interference from covariate shift.

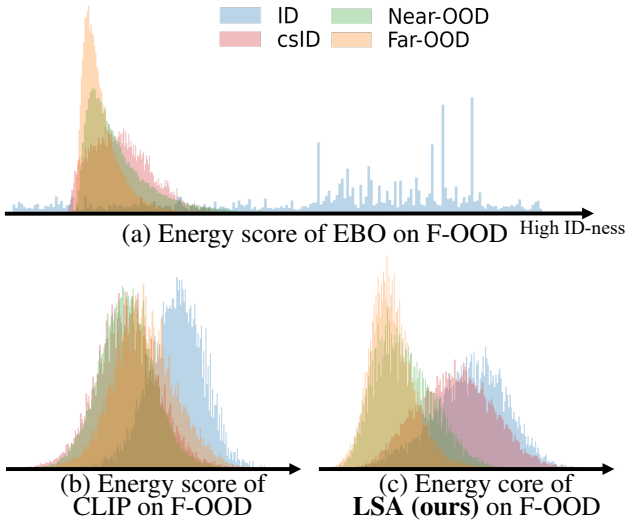


Figure 1: Different energy distributions on F-OOD. (a) Existing OOD methods such as EBO [8] confuse the semantic consistency of F-OOD on energy distribution, where a part of ID data obtains lower ID-ness than OOD and csID overlaps with OOD significantly. (b) Directly incorporating CLIP into F-OOD task does not eliminate the severe overlap between csID and OOD samples despite the more concentrated values of the ID cluster. **(c)** Our proposed LSA maintains consistent distribution between csID and ID data on energy score, which benefits from enhancing semantic alignment during context optimization. The energy scores in (b) and (c) are computed using image-text cosine similarity.

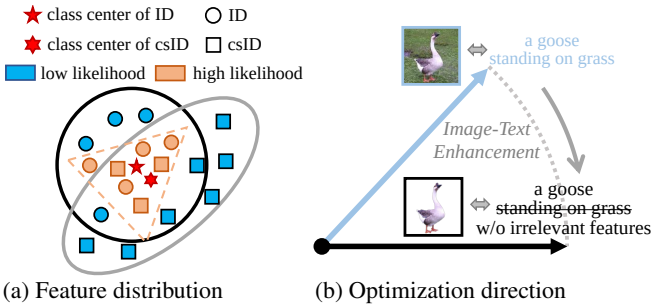


Figure 2: Motivation of the proposed method. (a) Compared to the discrepancy within the global distribution, high-likelihood features (*i.e.*, ones closer to the class center) from ID and csID clusters are more compact and consistent, which helps resist the interference of covariate shift. (b) Considering that the feature space of the CLIP-based model is jointly determined by the image-text pairs, our method simultaneously highlights the semantically high-likelihood visual regions and reduces the fitting to ID-irrelevant contexts.

To synergistically enhance the alignment of ID image and text semantics, we propose a likelihood-aware semantic alignment (LSA) framework that can be divided into visual and textual perspectives. For the design of the image branch, we construct an offline Gaussians sampling strategy, effectively modeling the whole probability density. Specifically, we first represent the ID images embeddings as a class-conditional Gaussian distribution utilizing the feature statistics of each class. Afterwards, we capture samples with the highest and lowest probability density from the Gaussians as ID semantic-relevant visual regions and OOD regularization regions, respectively. For the design of textual prompts, we present a bidirectional prompt customization mechanism, flexibly adjusting the ID/OOD boundary. Firstly, we enrich specific prompts for each ID class name, focusing on the ID-relevant semantics such as the foreground objects. Moreover, we learn unknown OOD prompts from scratch by aligning with the above OOD regularization regions, which serves as the 'negative anchor' role to enlarge the discrepancy between ID and OOD semantics in text space. Figure 1(c) illustrates that our method maintains consistency for data with the same semantic, benefiting from our implementation strategies.

Our contributions are summarized as follows: **1)** We explore and unleash the potential of the CLIP model for OOD detection under simultaneous semantic and covariate shifts. **2)** We present a customized prompt tuning method for the F-OOD task named LSA, which enhances the image-text alignment with semantically high-likelihood regions. **3)** Extensive experiments demonstrate that our proposed method achieves state-of-the-art performance on the realistic F-OOD benchmarks.

2 Related Work

2.1 Out-of-Distribution Detection

OOD samples in OOD detection tasks belong to different categories from the known ones and OOD detectors should concentrate on identifying semantic shift. Several classic works consider one dataset as ID and define all other datasets as OOD [6, 8–11, 23]. However, OOD detection models under this benchmark tend to overfit the low-level covariate shift and disregard the high-level semantic discrepancy between samples, conflicting with the goal of OOD detection.

To overcome this issue, more realistic settings including SCOOD benchmarks [13] and F-OOD benchmarks [14, 15] are proposed. SCOOD benchmarks require models to distinguish ID and OOD samples mixed in an unlabeled dataset, UDG [13] and ET-OOD [24] respectively apply K-means clustering and energy-based optimal transport to make models understand the semantic knowledge hidden in the unlabeled dataset. The more challenging but practical F-OOD benchmarks [14, 15] introduce ID samples with changes in appearances like style, lighting or viewpoint, which can be easily misclassified as OOD just owing to the covariate shift. Unfortunately, OpenOODv1.5 [15] has proved that dozens of existing OOD detectors are so sensitive to covariate shift that they encounter comprehensively significant performance degradation in the F-OOD detection problem. All of these indicate that the fundamental challenge of OOD detection tasks, which is how to be robust to covariate shift and fully focus on the semantic shift between ID/OOD samples, necessitates increased research efforts. In this work, we explore how to highlight the semantic knowledge in the F-OOD detection task, and propose the LSA which enhances the image-text alignment with semantically high-likelihood regions.

2.2 Pre-trained Vision-Language Models

Pre-trained vision-language models (*e.g.*, CLIP [25]) align the rich multi-modal representations using large-scale image-text pairs during training instead of learning with only image supervision, which enables CLIP to perform open-world image recognition tasks such as open-vocabulary semantic segmentation and OOD detection. Several CLIP-based open-vocabulary semantic segmentation works [20–22] have successfully enhanced the perception and localization of target semantic regions by aligning semantic regions in images and specified textual concepts. Previous studies on CLIP-based OOD detection primarily explore the performance of zero-shot OOD detection with CLIP [16–18] and the impact of classic CLIP fine-tuning methods [26–28] on OOD detection performance [29]. Recent works including CoOp [26], CoCoOp [27] and MaPLe [30], extend prompt tuning originating from NLP [31, 32] to computer vision and aim to fine-tune foundation models to adapt to new tasks in a parameter-efficient way. Inspired by this, LoCoOp [19] applies prompt learning with CLIP in OOD detection by regarding ID-irrelevant regions as OOD regularization and pushing them away from the ID class text embedding. However, existing CLIP-based OOD detection methods ignore the semantic-relevant visual knowledge (i.e., foreground objects) and align them with the learnable ID class-specific context. In this paper, we propose LSA, which is a customized prompt tuning method for F-OOD, to unleash the ability of CLIP to focus on semantic shift.

3 Method

3.1 Problem Statement

During training in the F-OOD task, we can only access the in-distribution (ID) training set $\mathscr{D}_{\text{ID}}$, and the corresponding label set is $\mathscr{Y}_{\text{ID}}$. In addition to the ID testing set $\mathscr{T}_{\text{ID}}$, F-OOD introduces the data from different domains compared to $\mathscr{D}_{\text{ID}}$ while retaining ID semantics as covariate-shifted ID (csID) set formed as $\mathscr{T}_{\text{csID}}$ during test time, where $\forall(x_{\text{csID}}, y_{\text{csID}}) \sim \mathscr{T}_{\text{csID}}, y_{\text{csID}} \in \mathscr{Y}_{\text{ID}}$. Moreover, the OOD testing set $\mathscr{T}_{\text{OOD}}$ from the semantics not overlapping with $\mathscr{Y}_{\text{ID}}$ is split into Near-OOD ($\mathscr{T}_{\text{OOD}}^{\texttt{near}}$) and Far-OOD ($\mathscr{T}_{\text{OOD}}^{\texttt{far}}$) which represent two different levels of covariate shift. We desire to identify data from $\mathscr{T}_{\text{OOD}}$ as OOD while classifying samples from $\mathscr{T}_{\text{ID}}$ and $\mathscr{T}_{\text{csID}}$ correctly.

3.2 Overall Framework

The framework of our LSA is presented in Figure 3. We first model the ID visual embeddings as an offline class-conditional Gaussian distribution in Figure 3(a), and then select regions with the highest/lowest probability density from each ID class Gaussians, which will be involved in training. During training shown in Figure 3(b), we apply $\mathscr{L}_{ce}$ and $\mathscr{L}_{uni}$ to learn ID and OOD contexts through aligning them with the corresponding visual regions sampled from Gaussians, and $\mathscr{L}_{bin}$ further widen the discrepancy between ID and OOD semantics in text space.

3.3 Offline Gaussians Sampling Strategy

To obtain ID semantic-relevant regions and OOD regularization regions for alignment with class-specific contexts without extra segmentation models [33] or generative models [34], we represent the ID visual embeddings as a class-conditional Gaussian distribution and sample distinguishable likelihood regions from the feature Gaussians.

We assume the probability density of visual embeddings follow the class-conditional Gaussian distribution based on the hypothesis made in [35]: $p_\theta(\boldsymbol{\varphi}_v(\boldsymbol{x})|y=c) = \mathscr{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, where $\boldsymbol{\varphi}_v$ is the visual encoder of CLIP, $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are the mean and covariance matrix of the *c-th* ID class, respectively. To parameterize the ID class-conditional Gaussians with the feature statistics, we calculate the class mean $\widehat{\boldsymbol{\mu}}_c$ and covariance $\widehat{\boldsymbol{\Sigma}}_c$ as:

$$\widehat{\boldsymbol{\mu}}_c = \frac{1}{N_c} \sum_{y_i=c} \boldsymbol{\varphi}_v(x_i), \tag{1}$$

$$\widehat{\boldsymbol{\Sigma}}_c = \frac{1}{N_c} \sum_{y_i=c} (\boldsymbol{\varphi}_v(x_i) - \widehat{\boldsymbol{\mu}}_c)(\boldsymbol{\varphi}_v(x_i) - \widehat{\boldsymbol{\mu}}_c)^\top, \tag{2}$$

where $x_i \in \mathscr{D}_{\text{ID}}$ and N_c is the number of samples in ID class c. Then we can flexibly select regions from the estimated Gaussians based on the likelihood. Specifically, we sample N data from the above Gaussians for each ID class and obtain a class-specific probability density set $\{f(\boldsymbol{a}_c^1), f(\boldsymbol{a}_c^2), ..., f(\boldsymbol{a}_c^N)\}$, where $f(\cdot)$ is the function of Gaussian probability density, and the smaller probability density corresponds to the larger distance from the mean (*i.e.*, class center). As a result, we respectively capture data with the highest and lowest probability density as ID semantic-relevant and OOD regularization regions:

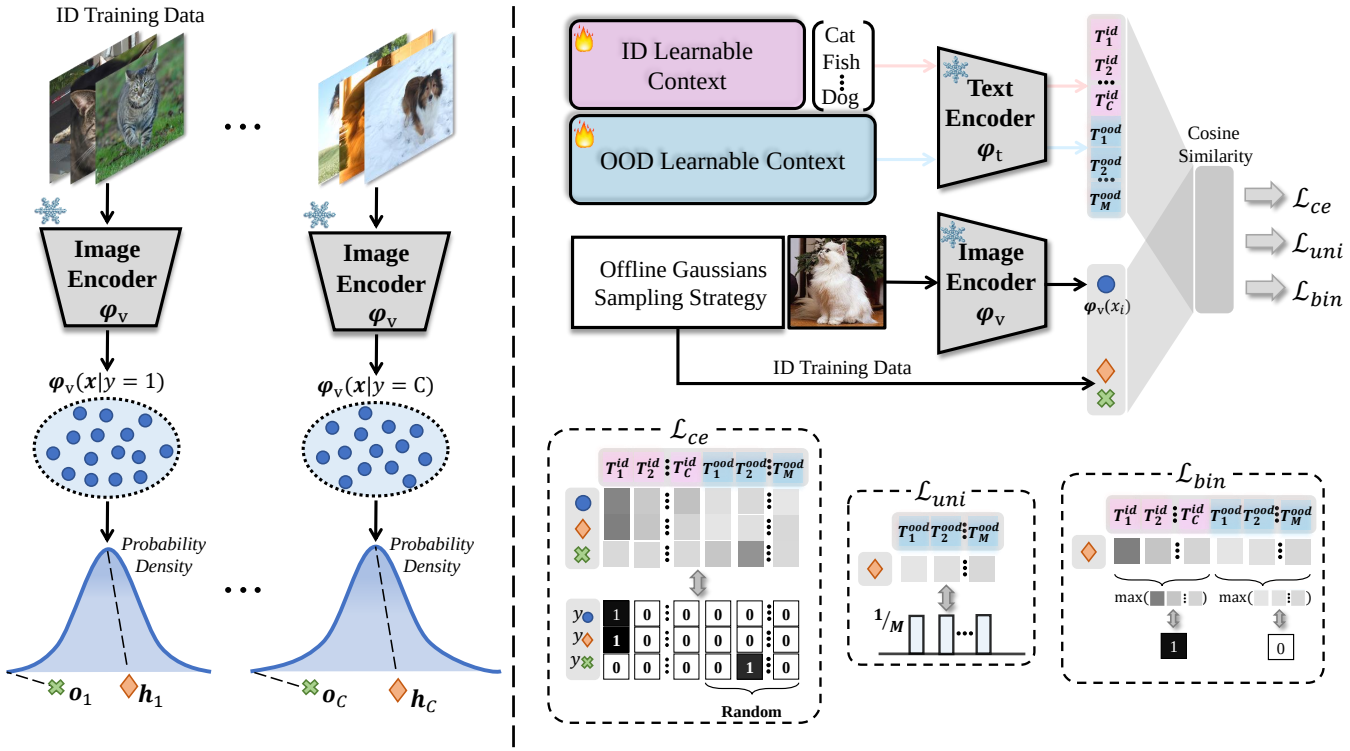
$$\boldsymbol{h}_c = \boldsymbol{a}_c^i, \text{ where } i = \arg\max_n f(\boldsymbol{a}_c^n), \tag{3}$$

$$\boldsymbol{o}_c = \boldsymbol{a}_c^j, \text{ where } j = \arg\min_n f(\boldsymbol{a}_c^n), \tag{4}$$

where $\boldsymbol{h}_c, \boldsymbol{o}_c \in \mathbb{R}^D$. D is the dimension of visual representation output by $\boldsymbol{\varphi}_v$, so $\boldsymbol{h}_c$ and $\boldsymbol{o}_c$ can directly join in training. We denote the corresponding sets as $\mathscr{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_C\}$ and $\mathscr{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_C\}$, and C is the number of ID classes. To access updated $\mathscr{H}$ and $\mathscr{O}$ during training, we extract the embeddings of $\mathscr{D}_{\text{ID}}$ before training and maintain an offline class-conditional queue with N_c extracted embeddings from each ID class. In each iteration, we randomly replace a part of embeddings in each class queue with the same number of new embeddings.

3.4 Bidirectional Prompt Customization

To match diverse ID objects effectively, for the *c-th* ID class, we concatenate the learnable class-specific prompts $\boldsymbol{W}_c \in \mathbb{R}^{K\times D}$ with the D-dimensional class word embeddings $\boldsymbol{e}_c$, where K is the number of tokens. The textual embedding of the *c-th* ID class is formed as $\boldsymbol{Z}_c = [\boldsymbol{W}_c^1, \boldsymbol{W}_c^2, ..., \boldsymbol{W}_c^K, \boldsymbol{e}_c]$. Moreover, to enlarge the discrepancy between ID/OOD semantic, we expand M learnable OOD (negative) prompts $\tilde{\boldsymbol{Z}} \in \mathbb{R}^{M\times K\times D}$ which to be aligned with samples in $\mathscr{O}$. Then we obtain the ID and OOD textual representation denoted as $\boldsymbol{T}^{id} = \boldsymbol{\varphi}_t(\boldsymbol{Z}) \in \mathbb{R}^{C\times D}$ and $\boldsymbol{T}^{ood} = \boldsymbol{\varphi}_t(\tilde{\boldsymbol{Z}}) \in \mathbb{R}^{M\times D}$, where



(a) Offline Gaussians Sampling Strategy (b) Training Pipeline

Figure 3: Overall framework of our proposed LSA. (a) The offline class-conditional Gaussian distribution is first modeled with the statistics of ID embeddings, then we select samples with the highest and lowest probability density in the Gaussians as ID semantic-relevant regions $\boldsymbol{h}$ and OOD regularization regions $\boldsymbol{o}$, respectively. (b) ID and OOD contexts are optimized through three losses. $\mathscr{L}_{ce}$ align ID context with ID image embedding and $\boldsymbol{h}$, and bring OOD context and $\boldsymbol{o}$ close, where every $\boldsymbol{o}$ obtains a random OOD label. $\mathscr{L}_{uni}$ unify the similarity of $\boldsymbol{h}$ with OOD text embedding. Meanwhile, a binary sigmoid loss $\mathscr{L}_{bin}$ is utilized to encourage ID and OOD textual embeddings to further separate and enlarge the ID/OOD semantic discrepancy. **Flames** and **Snowflakes** refer to learnable and frozen parameters, respectively.

φ_t is the text encoder of CLIP. And the completed txetual representation is written as $\boldsymbol{T} = [\boldsymbol{T}^{id}, \boldsymbol{T}^{ood}] \in \mathbb{R}^{(C+M)\times D}$.

With the updated $\mathscr{H}$ and $\mathscr{O}$ in every iteration, we first consist $\mathscr{D}_{\text{few}}$ with 16 images for each ID class from $\mathscr{D}_{\text{ID}}$ to limit the training cost. Then we form the visual embeddings set $\mathscr{V} = \{\varphi_v(x_{i'}) | x_{i'} \in \mathscr{D}_{\text{few}}\} \cup \mathscr{H}_s \cup \mathscr{O}$ for alignment with $\boldsymbol{T}$ through corss-entropy (CE) loss. Note that to maintain the class distribution in a batch of training data, we randomly select S samples from $\mathscr{H}$ to constitute $\mathscr{H}_s$ and S is half of batchsize ($S < C$). For one embedding $\boldsymbol{v} \in \mathbb{R}^D$ from $\mathscr{V}$, the prediction probability on the ID classes and the expanded OOD context is formed as:

$$p(\boldsymbol{y}|\boldsymbol{v}) = \frac{\exp(\boldsymbol{s_v}/\tau)}{\sum_{k=1}^{C+M}\exp(s_v^k/\tau)} \in \mathbb{R}^{(C+M)}, \tag{5}$$

where τ is a temperature parameter learned by CLIP. $\boldsymbol{s_v} = [s_v^1, s_v^2, ..., s_v^{C+M}]$ and $s_v^k = \cos(\boldsymbol{v}, \boldsymbol{T}_k)$. In terms of ground truth labels in CE loss, we assign random labels from $\mathscr{Y}_{\mathscr{O}} = \{C, C+1, ..., C+M-1\}$ to samples in $\mathscr{O}$, while labels $\mathscr{Y}_{\mathscr{H}}$ in $\mathscr{H}_s$ are the index of selected samples from $\mathscr{H}$ and $\mathscr{Y}_{\mathscr{H}} \in \mathscr{Y}_{\text{ID}} = \{0, 1, ..., C-1\}$. CE loss is formulated as:

$$\mathscr{L}_{ce} = -\mathbb{E}_{(\boldsymbol{v},y)\sim\mathscr{V}}[\ell_{\text{CE}}(p(\boldsymbol{y}|\boldsymbol{v}), y)]. \tag{6}$$

Meanwhile, we compute the similarity of ID semantic-relevant region $\boldsymbol{h} \in \mathscr{H}$ on OOD context as:

$$p(\boldsymbol{y}_{ood}|\boldsymbol{h}) = \frac{\exp(\boldsymbol{s_h}/\tau)}{\sum_{k=1}^{M}\exp(s_h^k/\tau)} \in \mathbb{R}^M, \tag{7}$$

where $\boldsymbol{s_h} = [s_h^1, s_h^2, ..., s_h^M]$ and $s_h^k = \cos(\boldsymbol{h}, \boldsymbol{T}_k^{ood})$. Then we employ the OE loss [9] to unify the prediction of $\boldsymbol{h}$ on the extended negative context:

$$\mathscr{L}_{uni} = -\mathbb{E}_{\boldsymbol{h}\sim\mathscr{H}}[\ell_{\text{OE}}(p(\boldsymbol{y}_{ood}|\boldsymbol{h}))], \tag{8}$$

where $\ell_{\text{OE}}(p(\boldsymbol{y}_{ood}|\boldsymbol{h})) = \frac{\mathbf{1}_M}{M} \cdot \log(\boldsymbol{y}_{ood}|\boldsymbol{h}))$, and $\frac{\mathbf{1}_M}{M}$ is the uniform posterior distribution over all of M OOD textual vectors.

To motivate $\boldsymbol{T}^{id}$ and $\boldsymbol{T}^{ood}$ to be further apart without a metric loss needing a margin hyperparameter, we introduce a binary sigmoid loss formulated as:

$$\begin{aligned}\mathscr{L}_{bin} = \mathbb{E}_{\boldsymbol{h}\sim\mathscr{H}}\left[-\log(\text{S}(\max_k \cos(\boldsymbol{h}, \boldsymbol{T}_k^{id})))\right] + \\ \mathbb{E}_{\boldsymbol{h}\sim\mathscr{H}}\left[-\log(1-\text{S}(\max_k \cos(\boldsymbol{h}, \boldsymbol{T}_k^{ood})))\right],\end{aligned} \tag{9}$$

where $\text{S}(\cdot)$ is the sigmoid function.

Our overall objective can be expressed as Eq. 10 with the weights γ and λ.

$$\mathscr{L} = \mathscr{L}_{ce} + \gamma\mathscr{L}_{uni} + \lambda\mathscr{L}_{bin}. \tag{10}$$

3.5 Test-time OOD Score

During evaluation, we propose the D-energy score for OOD detection, which is the difference between ID-related energy and OOD-related energy:

$$\text{D-energy}(xi) = \text{E}^{id}(x_i) - \text{E}^{ood}(x_i). \tag{11}$$

$\text{E}^{id}(x_i)$ and $\text{E}^{ood}(x_i)$ are the energy functions written as:

$$\text{E}^{id}(x_i) = T \cdot \log \sum_{k=1}^{C} \exp(s_{id}^k/T), \tag{12}$$

$$\text{E}^{ood}(x_i) = T \cdot \log \sum_{k=1}^{M} \exp(s_{ood}^k/T), \tag{13}$$

where T is the temperature parameter and we apply its default value of 1. s_{id}^k and s_{ood}^k are the similarities of the test data with ID and OOD contexts, which are formed as:

$$\begin{aligned} s_{id}^k &= \cos(\boldsymbol{\varphi}_v(x_i), \boldsymbol{T}_k^{id}), k \in \{1,2,...,C\}, \\ s_{ood}^k &= \cos(\boldsymbol{\varphi}_v(x_i), \boldsymbol{T}_k^{ood}), k \in \{1,2,...,M\}. \end{aligned} \tag{14}$$

ID data from $\mathscr{T}_{\text{ID}}$ and $\mathscr{T}_{\text{csID}}$ will obtain higher $\text{E}^{id}(x_i)$ considering their larger similarity with $\boldsymbol{T}^{id}$ than OOD data, while the $\text{E}^{ood}(x_i)$ of ID data will be lower because $\boldsymbol{T}^{ood}$ is pushed away from $\boldsymbol{T}^{ood}$ through $\mathscr{L}_{bin}$ in Eq. 9. Finally, ID data will be promoted to produce higher ID-ness on the D-energy score which subtracts $\text{E}^{ood}(x_i)$ from $\text{E}^{id}(x_i)$.

In F-OOD, $\mathscr{T}_{\text{OOD}}^{\texttt{near}}$ is more complex to detect owing to its similar domain and style with $\mathscr{D}_{\text{ID}}$ while $\mathscr{T}_{\text{OOD}}^{\texttt{far}}$ is more tractable considering its covariate shift. Benefit from our learned prompts with sensitivity to semantic shift and robustness under covariate shift, we introduce the extra MCM score [17] to detect samples from $\mathscr{T}_{\text{OOD}}^{\texttt{near}}$:

$$\begin{aligned} \text{MCM}(xi) &= \max_k \cos(\boldsymbol{\varphi}_v(x_i), \boldsymbol{T}_k^{id}), \\ & k \in \{1,2,...,C\}. \end{aligned} \tag{15}$$

4 Experiments

4.1 Benchmarks and Compared Methods

The full-spectrum out-of-distribution (F-OOD) detection benchmarks are first proposed in SEM [14], but the benchmarks in SEM are limited by the data scale and are unavailable in the official repository. Hence, 'F-OOD' in this work refers to the full-spectrum benchmarks in OpenOODv1.5 [15] by default. F-OOD benchmarks include the large-scale ImageNet-200 benchmark and ImageNet-1K benchmark, which regard ImageNet-200 and ImageNet-1K [36] as ID training data $\mathscr{D}_{\text{ID}}$, respectively. They both incorporate ImageNet-C [37] with image corruptions, ImageNet-R [38] with style changes and ImageNet-V2 [39] with resampling bias as $\mathscr{T}_{\text{csID}}$. In terms of $\mathscr{T}_{\text{OOD}}^{\texttt{near}}$, the two benchmarks include SSB-hard [40] and NINCO [41]. Additionally, they consider iNaturalist [42], Textures [43] and OpenImage-O [44] as $\mathscr{T}_{\text{OOD}}^{\texttt{far}}$.

We first select methods, which achieve the best result on FPR@95, AUROC and AUPR-IN of Near-OOD and Far-OOD according to the results released by OpenOODv1.5 in full results, as compared to methods. Then three CLIP prompt tuning methods *e.g.,* CoOp [26], CoCoOp [27] and MaPLe [30] are included. In the area of OOD detection, the CLIP zero-shot method MCM [17] and prompt tuning method LoCoOp [19] are also chosen as comparison methods. Our proposed LSA applies CLIP based on ViT-B/16 in all experiments.

4.2 Results on Full-Spectrum Benchmarks

We compare the results of our proposed approach with the compared methods in Table 1. We only report the average metric values on the corresponding OOD datasets for each benchmark limited by space. Results show that our proposed LSA consistently obtains the best results across all OOD detection metrics. LSA demonstrates consistent superiority across all CLIP-based methods, dispelling skepticism that its enhanced OOD performance derives solely from CLIP's intrinsic capabilities. It's worth noting that, in the troublesome Near-OOD of the two benchmarks, LSA significantly outperforms LoCoOp by 28.45%/20.35% on FRP@95 and 15.26%/18.88% on AUROC.

4.3 Ablation Study and Qualitative Analysis

Effectiveness of losses and ID semantic-relevant regions $\boldsymbol{h}$. We first analyze the effect of each loss in our proposed LSA in Table 2. $\mathscr{L}_{ce}$ enhances the alignment of ID semantic-relevant regions $\boldsymbol{h}$ with $\boldsymbol{T}^{id}$ and initially learns OOD prompts with the OOD regularization regions $\boldsymbol{o}$, achieving better results on Near-OOD than LoCoOp. $\mathscr{L}_{uni}$ which flattens the similarity of $\boldsymbol{h}$ with $\boldsymbol{T}^{ood}$ and $\mathscr{L}_{bin}$ that pushes $\boldsymbol{T}^{id}$ and $\boldsymbol{T}^{ood}$ apart from each other, both further enlarge the advantages based on $\mathscr{L}_{ce}$. Finally, our complete method combining the three losses boosts the performance to the best.

Then we show the effect of the ID semantic-relevant regions $\boldsymbol{h}$ sampled from Gaussians in Table 3. Exp#5 is the version of LSA. In general, $\boldsymbol{h}$ will enhance the performance on Near-OOD $\mathscr{T}_{\text{OOD}}^{\texttt{near}}$ which can't be easily detected considering its homologous domain with $\mathscr{D}_{\text{ID}}$, while covariance information from global images can improve the detection to Far-OOD $\mathscr{T}_{\text{OOD}}^{\texttt{far}}$. Exp#1 which exploits no $\boldsymbol{h}$, performs well on $\mathscr{T}_{\text{OOD}}^{\texttt{far}}$ while failing on $\mathscr{T}_{\text{OOD}}^{\texttt{near}}$ compared to Exp#5. The comparison between Exp#2 – #4 and Exp#1 shows that the performance exhibits varying degrees of improvement on $\mathscr{T}_{\text{OOD}}^{\texttt{near}}$ when $\boldsymbol{h}$ is involved in training. Exp#5 achieves the best results on $\mathscr{T}_{\text{OOD}}^{\texttt{near}}$ utilizing $\boldsymbol{h}$ in all three losses. Although LSA is a bit inferior on AUROC of $\mathscr{T}_{\text{OOD}}^{\texttt{far}}$ compared to Exp#1, the result of this metric outperforms all compared methods as presented in Table 1. Moreover, we pay more attention to the performance on $\mathscr{T}_{\text{OOD}}^{\texttt{near}}$ considering the covariate shift in $\mathscr{T}_{\text{OOD}}^{\texttt{far}}$ is a 'shortcut' for OOD detection.

Table 1: Comparison between previous methods and LSA on two F-OOD benchmarks. In each row, we report the averaged values on two Near-OOD datastes and three Far-OOD datasets, while the dataset-specific values are shown in Appendix. 'Near'/'Far' means 'Near-OOD'/'Far-OOD'. ↑/↓ indicates higher/lower value is better, and the best results are in **bold**.

Benchmark: ImageNet-200						
Methods	FPR@95 ↓		AUROC ↑		AUPR-IN ↑	
	Near	Far	Near	Far	Near	Far
MCD [11]	89.15	89.57	58.63	65.07	73.08	89.96
NPOS [45]+KNN [46]	94.99	65.04	51.19	78.99	69.44	93.24
PixMix [47]+RMDS [48]	91.30	93.60	61.91	65.43	74.79	90.35
Gram [49]	91.02	83.89	61.23	65.26	75.16	87.50
CLIP-based Models						
MCM [17]	87.24	62.43	70.07	88.99	79.58	97.20
CoOp [26]	91.84	67.60	63.69	85.75	76.11	95.92
CoCoOp [27]	89.07	70.29	67.17	88.42	77.37	96.93
MaPLe [30]	91.19	72.72	67.28	86.19	79.10	96.27
LoCoOp [19]	81.18	49.82	70.15	89.10	79.16	96.76
LSA(ours)	**52.73**	**32.41**	**85.41**	**90.97**	**90.26**	**97.49**

Benchmark: ImageNet-1K						
Methods	FPR@95 ↓		AUROC ↑		AUPR-IN ↑	
	Near	Far	Near	Far	Near	Far
DeepAug [38]+SHE [50]	83.26	68.70	68.27	78.85	89.11	96.49
StyAug [51]+GradNorm [52]	87.14	58.82	65.27	81.62	87.99	96.81
AugMix [53]+SHE [50]	84.45	60.26	69.66	83.06	89.29	97.24
ASH [54]	93.27	59.56	60.52	86.75	85.41	97.15
CLIP-based Models						
MCM [17]	94.74	77.47	58.11	82.56	83.87	97.84
CoOp [26]	95.82	80.31	56.36	82.41	83.76	97.64
CoCoOp [27]	96.23	79.59	53.11	74.48	82.81	94.87
MaPLe [30]	95.23	75.84	57.38	83.70	84.38	97.84
LoCoOp [19]	90.91	54.33	59.34	84.02	84.66	97.92
LSA(ours)	**70.56**	**48.06**	**78.22**	**86.85**	**93.05**	**98.10**

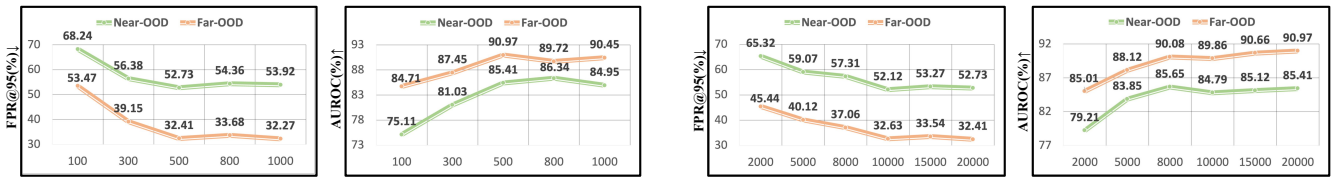
Table 2: Our completed method combines $\mathscr{L}_{ce}$, $\mathscr{L}_{uni}$ and $\mathscr{L}_{bin}$. 'Near'/'Far' means 'Near-OOD'/'Far-OOD'. The best results are in **bold**.

$\mathscr{L}_{ce}$	$\mathscr{L}_{uni}$	$\mathscr{L}_{bin}$	FPR@95 ↓		AUROC ↑	
			Near	Far	Near	Far
✓			68.94	55.03	77.29	88.44
✓	✓		60.52	40.63	83.86	89.39
✓		✓	59.35	38.96	82.27	89.70
✓	✓	✓	**52.73**	**32.41**	**85.41**	**90.97**

Table 3: $\mathscr{L}_{ce\text{-}G}$ means excluding $\boldsymbol{h}$ from $\mathscr{H}_s$ in $\mathscr{L}_{ce}$. $\mathscr{L}_{uni\text{-}G}$ and $\mathscr{L}_{bin\text{-}G}$ mean replacing $\boldsymbol{h} \in \mathscr{H}$ with the same number of global image embeddings in the corresponding loss. LSA combines $\mathscr{L}_{ce}$, $\mathscr{L}_{uni}$ and $\mathscr{L}_{bin}$. 'Near'/'Far' means 'Near-OOD'/'Far-OOD'. We refer to each experiment by its index for brevity. The best results are in **bold** and the second best are underlined.

	Strategy	FPR@95 ↓		AUROC ↑	
		Near	Far	Near	Far
1:	$\mathscr{L}_{ce\text{-}G}$+$\mathscr{L}_{uni\text{-}G}$+$\mathscr{L}_{bin\text{-}G}$	58.15	<u>33.12</u>	82.23	**91.50**
2:	$\mathscr{L}_{ce}$+$\mathscr{L}_{uni\text{-}G}$+$\mathscr{L}_{bin\text{-}G}$	56.92	33.93	82.50	90.13
3:	$\mathscr{L}_{ce}$+$\mathscr{L}_{uni}$+$\mathscr{L}_{bin\text{-}G}$	56.03	38.83	83.59	89.29
4:	$\mathscr{L}_{ce}$+$\mathscr{L}_{uni\text{-}G}$+$\mathscr{L}_{bin}$	55.07	36.95	84.06	89.46
5:	**LSA(ours)**	**52.73**	**32.41**	**85.41**	<u>90.97</u>

Analysis about hyper-parameters. Here we analyze the impact of the hyperparameters in the offline Gaussians sampling strategy, including the size of each class queue N_c and the number of sampled data N from each class Gaussians. In general, a larger N_c used to calculate the statistics is more beneficial to estimate the precise class-conditional Gaussian distribution. Figure 4(a) shows the performance fluctuation is slight when $N_c \geq 500$, and we choose $N_c = 500$. Although every class contains about 1,200 training samples in ImageNet-1k and ImageNet-200, the offline embedding queue extracted by the pre-trained encoder does not need to be optimized and the embeddings from a part of samples are sufficient to model an accurate Gaussian distribution.



(a) Comparison between different N_c (b) Comparison between different N

Figure 4: Our proposed method obtains a stably good result the size of each class queue $N_c \geq 500$ and the number of sampled data from each class Gaussians $N \geq 10000$, showing the robustness of our method. ↑/↓ indicates higher/lower value is better.

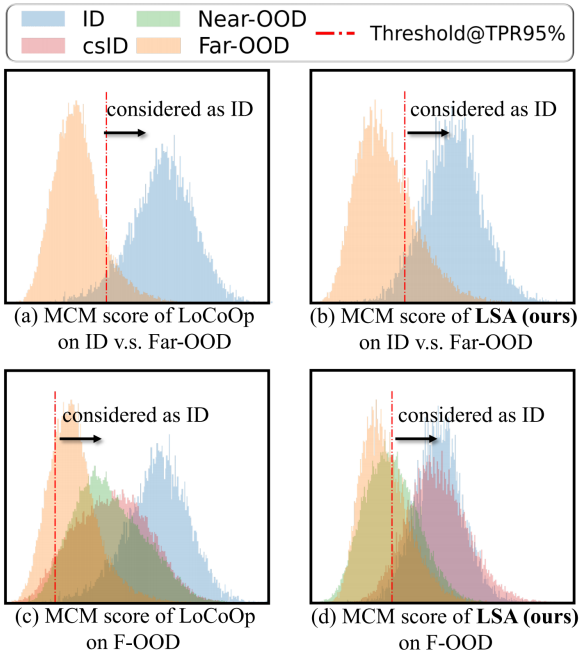
Table 4: MCM score is formulated in Equation 15. △ denotes the subtraction results between with/without MCM score during evaluation. 'Near'/'Far' means 'Near-OOD'/'Far-OOD'.

	FPR@95 ↓		AUROC ↑		AUPR-IN ↑	
	Near	Far	Near	Far	Near	Far
LoCoOp w/o MCM	**76.20**	66.79	**71.93**	79.57	**80.21**	93.63
LoCoOp w/ MCM	81.18	**49.82**	70.15	**89.10**	79.16	**96.76**
△	+4.98	-16.97	-1.78	+9.53	-1.05	+3.13
LSA w/o MCM	55.62	**32.41**	82.35	**90.97**	86.39	**97.49**
LSA w/ MCM	**52.73**	42.14	**85.41**	88.54	**90.26**	97.02
△	-2.89	+9.73	+3.06	-2.43	+3.87	-0.47

Moreover, we sample regions with the highest/lowest probability density from N data in the Gaussians, and the performance is not sensitive to N when $N \geq 10000$ as shown in Figure 4(b). N is not hard to choose considering the 3σ principle, which suggests about 99.7% of data points in Gaussians fall within three standard deviations (σ) of the mean (μ). As a result, we can precisely sample data located around μ and $\mu \pm 3\sigma$ when N is large enough.

Effect of MCM scores in different methods. MCM score formed as Eq. 15 is the maximum similarity of data with ID text embedding $\boldsymbol{T}^{id}$ and can reflect the ability of learned ID context to capture semantics. We observe that employing the additional MCM score in LSA boosts the performance on $\mathcal{T}_{\text{OOD}}^{\texttt{near}}$ while encountering a decline on $\mathcal{T}_{\text{OOD}}^{\texttt{far}}$. This phenomenon occurs because the $\boldsymbol{T}^{id}$ in LSA is brought close to the semantic high-likelihood regions during optimization and may not take into account the non-semantic covariance information which improves the recognition of $\mathcal{T}_{\text{OOD}}^{\texttt{far}}$. Moreover, we analyze the effect of MCM score in LoCoOp, which is completely opposite to that of LSA as demonstrated in Table 4. This comparison indicates that LSA more efficiently obtains the prompts with sensitivity to semantic shift and robustness under covariate shift, and we can flexibly select OOD score at test times according to the difficulty of OOD. It notes that our method consistently outperforms LoCoOp irrespective of whether the MCM score are used.

We also present the distribution of data from F-OOD on the MCM score learned by LoCoOp and LSA in Figure 5. When only facing $\mathcal{T}_{\text{ID}}$ and $\mathcal{T}_{\text{OOD}}^{\texttt{far}}$, which clearly differ in both semantics and covariance, the MCM score from LoCoOp more significantly distinguishes between the data groups, and fewer samples from $\mathcal{T}_{\text{OOD}}^{\texttt{far}}$ are misclassified into ID as demonstrated by the comparison between Figure 5(a) and Figure 5(b). However, when F-OOD introduces $\mathcal{T}_{\text{csID}}$ and $\mathcal{T}_{\text{OOD}}^{\texttt{near}}$, LoCoOp suffers from the severe overlap between data with different semantics on its MCM score as shown in Figure 5(c). Gratifyingly, the MCM score based on our learned $\boldsymbol{T}^{id}$ captures the consistency of semantics as Figure 5(d) shows, two groups of ID data maintain a similar distribution, and the same goes for $\mathcal{T}_{\text{OOD}}^{\texttt{near}}$ and $\mathcal{T}_{\text{OOD}}^{\texttt{far}}$.



(a) MCM score of LoCoOp on ID v.s. Far-OOD (b) MCM score of **LSA (ours)** on ID v.s. Far-OOD

(c) MCM score of LoCoOp on F-OOD (d) MCM score of **LSA (ours)** on F-OOD

Figure 5: We compare distributions of MCM score learned by LoCoOp and LSA on F-OOD. The comparison between (a) and (c) shows when encountering csID and Near-OOD data, LoCoOp confuses csID samples with Near-OOD while splitting them from ID, and the two groups of OOD also exhibit distinct distributions. However, as shown in (d), csID and ID samples share a similar distribution, and Near-OOD maintains consistency with Far-OOD on the MCM score learned by our proposed LSA.

Influence of OOD scores. To exclude the possibility that the performance of our proposed LSA only gains from the

OOD score we applied, we compare the performance of LoCoOp [19] with LSA across several OOD scores, including MCM, MCM-GL and energy. As shown in Table 5, LSA consistently outperforms LoCoOp no matter which OOD score is chosen, even with the MCM-GL score boosting the OOD detection performance of LoCoOp. Moreover, our D-energy score in Eq. 11 which utilizes the $\boldsymbol{T}^{ood}$ away from ID semantic, further improves the OOD detection performance based on the energy score.

Table 5: 'MCM-GL' is the score applied in LoCoOp [19]. 'Energy' and 'D-energy' are shown in Equations 12 and 11.

Score	Methods	FPR@95 ↓		AUROC ↑	
		Near	Far	Near	Far
MCM	LoCoOp [19]	91.40	69.26	65.28	86.12
	LSA(ours)	56.61	52.78	81.53	84.44
MCM-GL	LoCoOp [19]	81.18	49.82	70.15	89.10
	LSA(ours)	57.31	40.75	80.24	89.74
Energy	LoCoOp [19]	88.46	90.74	63.03	72.51
	LSA(ours)	56.85	53.41	81.81	82.32
D-energy	**LSA(ours)**	**55.62**	**32.41**	**82.35**	**90.97**

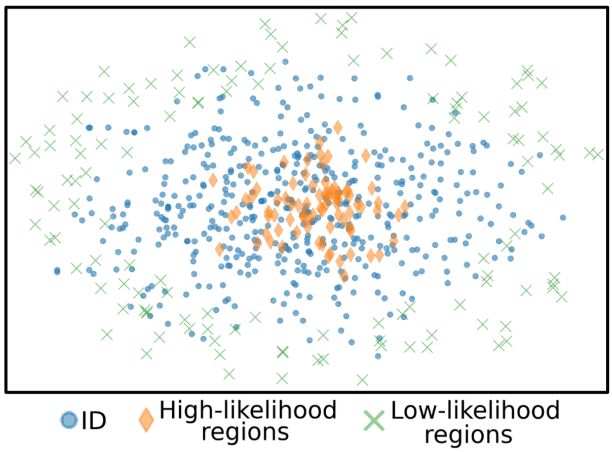


Figure 6: We visualize the distribution of 'goose' ID data, high-likelihood and low-likelihood regions from 'goose' Gaussians. High-likelihood regions are more compact than global ID features and more consistent with the class center, while low-likelihood regions are concentrated in the boundary far from the class center.

Visualization of the sampled regions from Gaussians. As illustrated in the visualization using UMAP in Figure 6, high-likelihood regions sampled from Gaussians are more compact and more proximate to the class feature center compared to global ID images, indicating that they are more suitable for alignment with ID class context. Moreover, low-likelihood regions can be exploited as OOD regularization since they are located far from the class center.

5 Conclusion

In this work, we focus on addressing the realistic F-OOD detection task, where models are required to precisely perceive ID semantics when disturbed by both semantic and covariate shifts. We reveal that high-likelihood regions concentrated around the ID or csID class centers maintain a solid semantic consistency, which contributes to mitigating the interference from covariate shift. We propose the LSA framework that efficiently captures semantically high-likelihood visual regions and adaptively optimizes the prompt-based decision boundary. Experimental results indicate that LSA effectively promotes the image-text alignment to focus on semantic shift, achieving superior performance than existing OOD detection and prompt learning methods on F-OOD benchmarks.

Funding

This work is supported by the National Natural Science Foundation of China (NSFC) under Grants 62225207, 62436008 and 62306295.

Author Contributions

Conceptualization, Fan LU and Kai ZHU; methodology, Fan LU.; software, Fan LU; validation, Fan LU, Kai ZHU and Wei ZHAI; formal analysis, Fan LU; investigation, Wei ZHAI and Yang CAO; resources, Yang CAO and Zheng-Jun ZHA; data curation, Kai ZHU and Wei ZHAI; writing—original draft preparation, Fan LU and Kai ZHU; writing—review and editing, Wei ZHAI, Yang CAO and Zheng-Jun ZHA; visualization, Fan LU; supervision, Yang CAO and Zheng-Jun ZHA; project administration, Zheng-Jun ZHA; funding acquisition, Zheng-Jun ZHA. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data Avaliable

The data and materials used in this study are availableupon request from the corresponding author.

References

[1] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778 (2016). IEEE

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[3] Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 427-436 (2015)

[4] Park, S., Lee, K. H., Ko, B., Kim, N.: Unsupervised anomaly detection with generative adversarial networks in mammography. Scientific Reports **13**(1), 2925 (2023)

[5] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In 2012 IEEE conference on computer vision and pattern recognition, pp. 3354-3361 (2012). IEEE

[6] Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)

[7] Du, X., Wang, Z., Cai, M., Li, Y.: Vos: Learning what you don't know by virtual outlier synthesis. arXiv preprint arXiv:2202.01197 (2022)

[8] Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. In Proceedings of the 34th International Conference on Neural Information Processing Systems, pp. 21464-21475 (2020)

[9] Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. arXiv preprint arXiv:1812.04606 (2018)

[10] Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690 (2017)

[11] Yu, Q., Aizawa, K.: Unsupervised out-of-distribution detection by maximum classifier discrepancy. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 9518-9526 (2019)

[12] Regmi, S.: AdaSCALE: Adaptive Scaling for OOD Detection. arXiv preprint arXiv:2503.08023 (2025)

[13] Yang, J., Wang, H., Feng, L., Yan, X., Zheng, H., Zhang, W., Liu, Z.: Semantically coherent out-of-distribution detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8301-8309 (2021). IEEE

[14] Yang, J., Zhou, K., Liu, Z.: Full-spectrum out-of-distribution detection. International Journal of Computer Vision **131**(10), 2607-2622, (2023)

[15] Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Li, H.: Openood v1. 5: Enhanced benchmark for out-of-distribution detection. arXiv preprint arXiv:2306.09301 (2023)

[16] Esmaeilpour, S., Liu, B., Robertson, E., Shu, L.: Zero-shot out-of-distribution detection based on the pre-trained model clip. In Proceedings of the AAAI conference on artificial intelligence, pp. 6568-6576 (2022)

[17] Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., Li, Y.: Delving into out-of-distribution detection with vision-language representations. In Proceedings of the 36th International Conference on Neural Information Processing Systems, pp. 35087-35102, (2022)

[18] Miyai, A., Yu, Q., Irie, G., Aizawa, K.: Zero-shot in-distribution detection in multi-object settings using vision-language foundation models. arXiv preprint arXiv:2304.04521 (2023)

[19] Miyai, A., Yu, Q., Irie, G., Aizawa, K.: Locoop: Few-shot out-of-distribution detection via prompt learning. In Proceedings of the 37th International Conference on Neural Information Processing Systems, pp. 76298-76310 (2023)

[20] Zhou, C., Loy, C. C., Dai, B.: Extract free dense labels from clip. In European Conference on Computer Vision, Cham: Springer Nature Switzerland, pp. 696-712 (2022)

[21] Cha, J., Mun, J., Roh, B.: Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11165-11174 (2023)

[22] Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 7061-7070 (2023)

[23] Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Song, D.: Scaling out-of-distribution detection for real-world settings. arXiv preprint arXiv:1911.11132 (2019)

[24] Lu, F., Zhu, K., Zhai, W., Zheng, K., Cao, Y.: Uncertainty-aware optimal transport for semantically coherent out-of-distribution detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3282-3291 (2023)

[25] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sutskever, I.: Learning transferable visual models from natural language supervision. In International conference on machine learning, pp. 8748-8763 (2021)

[26] Zhou, K., Yang, J., Loy, C. C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337-2348, (2021)

[27] Zhou, K., Yang, J., Loy, C. C., Liu, Z.: Conditional prompt learning for vision-language models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16816-16825, (2022)

[28] Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930 (2021)

[29] Ming, Y., Li, Y.: How does fine-tuning impact out-of-distribution detection for vision-language models?. International Journal of Computer Vision **132**(2), 596-609, (2024)

[30] Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., Khan, F. S.: Maple: Multi-modal prompt learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp. 19113-19122 (2023)

[31] Jiang, Z., Xu, F. F., Araki, J., Neubig, G., How can we know what language models know? Transactions of the Association for Computational Linguistics **8**, 423-438, (2020)

[32] Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., Singh, S., Autoprompt: Eliciting knowledge from language models with automatically generated prompts. arXiv preprint arXiv:2010.15980 (2020)

[33] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Girshick, R.: Segment anything. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 4015-4026, (2023)

[34] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Bengio, Y.: Generative adversarial nets. In 28th Conference on Neural Information Processing Systems (NIPS 2014), pp.1-9 (2024)

[35] Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), pp. 1-11 (2018)

[36] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pp. 248-255 (2009). IEEE

[37] Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019)

[38] Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Gilmer, J.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 8340-8349 (2021)

[39] Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet?. In International conference on machine learning, pp. 5389-5400 (2019)

[40] Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Open-set recognition: A good closed-set classifier is all you need? In Tenth International Conference on Learning Representations, pp. 1-27 (2021)

[41] Bitterwolf, J., Mueller, M., Hein, M.: In or out? fixing imagenet out-of-distribution detection evaluation. arXiv preprint arXiv:2306.00826 (2023)

[42] Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Belongie, S.: The inaturalist species classification and detection dataset. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8769-8778 (2018)

[43] Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3606-3613 (2014)

[44] Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4921-4930 (2022)

[45] Tao, L., Du, X., Zhu, X., Li, Y.: Non-parametric outlier synthesis, arXiv preprint arXiv:2303.02966 (2023)

[46] Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In International Conference on Machine Learning, pp. 20827-20840 (2022)

[47] Hendrycks, D., Zou, A., Mazeika, M., Tang, L., Li, B., Song, D., Steinhardt, J.: Pixmix: Dreamlike pictures comprehensively improve safety measures. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16783-16792 (2022)

[48] Ren, J., Fort, S., Liu, J., Roy, A. G., Padhy, S., Lakshminarayanan, B.: A simple fix to mahalanobis distance for improving near-ood detection. arXiv preprint arXiv:2106.09022 (2021)

[49] Sastry, C. S., Oore, S.: Detecting out-of-distribution examples with gram matrices. In International Conference on Machine Learning, pp. 8491-8501 (2020)

[50] Zhang, J., Fu, Q., Chen, X., Du, L., Li, Z., Wang, G., Zhang, D.: Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In The Eleventh International Conference on Learning Representations, pp. 1-19 (2022)

[51] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., Brendel, W.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In International conference on learning representations, pp.1-22 (2018)

[52] Huang, R., Geng, A., Li, Y.: On the importance of gradients for detecting distributional shifts in the wild. In Proceedings of the 35th International Conference on Neural Information Processing Systems, pp. 677-689 (2021)

[53] Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple data processing method to improve robustness and uncertainty. arXiv preprint arXiv:1912.02781 (2019)

[54] Djurisic, A., Bozanic, N., Ashok, A., Liu, R.: Extremely simple activation shaping for out-of-distribution detection. arXiv preprint arXiv:2209.09858 (2022)

[55] Dan H., Steven B., Norman M., Saurav K., Frank W., Evan D., Rahul D., Tyler Z., Samyak P., Mike G., *et al.*: The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. arXiv preprint arXiv:2006.16241 (2021)

[56] Asano, Y. M., Rupprecht, C., Vedaldi, A.: Self-labelling via simultaneous clustering and representation learning. arXiv preprint arXiv:1911.05371 (2019)

[57] Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In Proceedings of the 27th International Conference on Neural Information Processing Systems, pp. 2292-2300 (2013)

[58] Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 2556-2565 (2018)

Appendix

A Experiment Details

Following existing works [18, 19, 26], the CLIP based on ViT-B/16 [25] is employed for all experiments, which is trained by an SGD optimizer with a weight decay of 0.0005 and a momentum of 0.9. We use the cosine annealing learning rate starting at 0.004, taking total 100 epochs. The dataloader is prepared with a batch-size of 64 for the ID training set $\mathscr{D}_{\text{ID}}$. For the size of each class queue N_c and the number of sampled data N from each class Gaussians, we use 500 and 20000, respectively. The number of the expanded OOD contexts M is set as 15 and we apply the number of tokens $K = 3$ for both learnable ID and OOD contexts. The learnable ID and OOD contexts are both initialized randomly.

For the training objective of our proposed LSA is denoted as:

$$\mathscr{L} = \mathscr{L}_{ce} + \gamma\mathscr{L}_{uni} + \lambda\mathscr{L}_{bin}, \tag{A1}$$

where we set $\gamma = 0.5$ and $\lambda = 0.1$ for all experiments.

B More Analysis of ID Semantic-Relevant Regions

We explore the effectiveness of the sampled ID semantic-relevant regions $\boldsymbol{h}$ in other prompt tuning methods including CoOp [26] and LoCoOp [19] here. As shown in Table B1, $\boldsymbol{h}$ improves all results of CoOp and LoCoOp on the OOD detection metrics, again indicating the efficacy of $\boldsymbol{h}$ to promote models to overcome the interference of covariate shift and enhance OOD detection performance. It is worth noting that the enhancement of $\boldsymbol{h}$ to Far-OOD is significantly weaker than that to Near-OOD, in both CoOp and LoCoOp. This phenomenon is consistent with the analysis in Table. 3 of our main text, which suggests that $\boldsymbol{h}$ can purely reflect the attributes of semantics and will significantly enhance the detection performance on Near-OOD which can't be easily detected considering its slight covariate shift, while the distinct covariance information from global images can improve the detection to Far-OOD.

Table B1: '+$\boldsymbol{h}$' denotes that we introduce ID semantic-relevant regions $\boldsymbol{h}$ into the cross-entropy losses employed in CoOp [26] and LoCoOp [19]. 'Near'/'Far' means 'Near-OOD'/'Far-OOD'. ↑/↓ indicates higher/lower value is better. △ is the subtraction results between with/without $\boldsymbol{h}$. The best results are in **bold**

Strategy	FPR@95 ↓		AUROC ↑		AUPR-IN ↑	
	Near	Far	Near	Far	Near	Far
CoOp	91.84	67.60	63.69	85.75	76.11	95.92
CoOp+$\boldsymbol{h}$	**75.94**	**65.22**	**72.07**	**85.82**	**80.26**	**96.34**
△	-15.90	-2.38	+8.38	+0.07	+4.15	+0.42
LoCoOp	81.18	49.82	70.15	89.10	79.16	96.76
LoCoOp+$\boldsymbol{h}$	**75.73**	**48.47**	**74.84**	**89.51**	**82.35**	**97.24**
△	-5.45	-1.35	+4.69	+0.41	+3.19	+0.48

Moreover, the advantages of ID semantic-relevant regions $\boldsymbol{h}$ sampled from the class-conditional feature Gaussians are: **1)** Sampling semantically high-likelihood regions from the low-dimensional feature space is more tractable than from pixel or patch level, especially when a fixed number of patches need to be chosen but the sizes of foreground objects (semantic-relevant regions) vary. **2)** We can flexibly select desired regions from the Gaussians based on their probability density. **3)** These sampled regions can directly participate in training with the same dimensions as image embeddings.

C Effect of Learnable OOD Context

For the text branch of our proposed LSA, we design the bidirectional prompt customization mechanism which expands the additional learnable OOD (negative) context to enlarge the discrepancy between ID and OOD semantics in text space. Here we analyze the effect of the expanded OOD context. Specifically, we abandon the OOD context in LSA and accordingly reconstruct the training objective as:

$$\mathscr{L}' = \mathscr{L}'_{ce} + \gamma\mathscr{L}'_{uni}, \tag{C2}$$

where γ is still set as 0.5, $\mathscr{L}'_{ce}$ and $\mathscr{L}'_{uni}$ are formulated as Eq. C3 and Eq. C4, respectively.

$$\mathscr{L}'_{ce} = -\mathbb{E}_{(\boldsymbol{v}',y)\sim\mathscr{V}'}[\ell_{\text{CE}}(p(\boldsymbol{y}|\boldsymbol{v}'), y)], \tag{C3}$$

$$\mathscr{L}'_{uni} = -\mathbb{E}_{\boldsymbol{o}\sim\mathscr{O}}[\ell_{\text{OE}}(p(\boldsymbol{y}_{id}|\boldsymbol{o}))], \tag{C4}$$

where $\mathscr{V}' = \{\varphi_v(x_{i'})|x_{i'} \in \mathscr{D}_{\text{few}}\} \cup \mathscr{H}_s$, and $p(\boldsymbol{y}_{id}|\boldsymbol{o})$ is the similarity of OOD regularization regions $\boldsymbol{o}$ with all ID contexts. The definitions of $\mathscr{D}_{\text{few}}$, $\mathscr{H}_s$ and ℓ_{OE} are the same as that in main text.

As the comparison in Table C2 shows, expanding OOD context during optimization significantly enhances the performance of LSA. It indicates that, in addition to exploiting the feature of pure semantic regions, it is also essential to design appropriate methods to explicitly learn the discrepancy between ID/OOD semantics for OOD detection. Note that LSA can still achieve state-of-the-art performance even without using the learnable OOD context.

Table C2: Our proposed LSA tunes the CLIP with OOD context, and the training object is written as Eq. C2 when deprecating OOD context. 'Near'/'Far' means 'Near-OOD'/'Far-OOD'. The best results are in **bold**.

Strategy	FPR@95 ↓		AUROC ↑		AUPR-IN ↑	
	Near	Far	Near	Far	Near	Far
w/o OOD context	68.63	48.24	79.30	89.97	84.10	97.36
w OOD context	**52.73**	**32.41**	**85.41**	**90.97**	**90.26**	**97.49**

D Detailed Results and More Backbones

In this section, we first present the detailed results among all datasets of LSA in Table D3. Far-OOD datasets are generally easier to detect than Near-OOD, and their experimental results are better. However, the empirical difficulty of Textures [43] is more significant than other Far-OOD datasets and comparable to Near-OOD as shown in Table D3. It can be explained that the Textures [43] dataset has flat backgrounds compared to ImageNet and lacks clear semantic information (*i.e.*, specific foreground objects), which will be identified as OOD relying on the resulting covariate shift, but it may be not so easy for LSA that focuses on semantic shift.

Then we compare the performance of LSA with different visual encoder backbones. In general, LSA favors Vision Transformer as the visual encoder and achieves the best results with ViT-B/ on the two F-OOD benchmarks. However, it does not mean larger-scale models will necessarily improve performance. As shown in Table D4, LSA obtains the best results employing ResNet-50 and ViT-B/16 with fewer parameters in terms of CNN and Vision Transformer architectures, respectively. Notably, LSA with ResNet-50 still outperforms all ResNet-based methods in Tab.1 (the first four rows of each benchmark table) on 11 out of 12 metrics. All of these prove that employing a large-scale and strong architecture is not the key to success in OOD detection tasks, as indicated in OpenOODv1.5 [15].

Table D3: **The dataset-specific results of our proposed LSA on two F-OOD benchmarks.** ↑/↓ indicates higher/lower value is better, and the average results of Near-OOD and Far-OOD are denoted as 'Mean' and in **bold**.

Benchmark: ImageNet-200					
Near-/Far-OOD	Dataset	FPR@95 ↓	AUROC ↑	AUPR-IN ↑	AUPR-OUT ↑
Near-OOD	SSB-hard [40]	49.84	85.63	83.20	87.86
	NINCO [41]	55.62	85.19	97.32	51.29
	Mean	**52.73**	**85.41**	**90.26**	**69.58**
Far-OOD	iNaturalist [42]	12.34	97.30	99.29	91.32
	Textures [43]	52.58	84.18	97.29	48.78
	OpenImage-O [44]	32.31	91.43	95.89	84.66
	Mean	**32.41**	**90.97**	**97.49**	**74.92**
Benchmark: ImageNet-1K					
Near-/Far-OOD	Dataset	FPR@95 ↓	AUROC ↑	AUPR-IN ↑	AUPR-OUT ↑
Near-OOD	SSB-hard [40]	69.13	79.57	88.11	66.72
	NINCO [41]	72.00	76.88	97.98	20.34
	Mean	**70.56**	**78.22**	**93.05**	**43.53**
Far-OOD	iNaturalist [42]	28.53	93.90	99.24	72.35
	Textures [43]	65.21	79.41	98.39	24.18
	OpenImage-O [44]	50.44	85.44	96.62	61.01
	Mean	**48.06**	**86.85**	**98.10**	**52.51**

Table D4: ↑/↓ indicates higher/lower value is better. The best results of CNN (ResNet-50 & ResNet-101) and Vision Transformer (ViT-B/16 & ViT-B/32) are in **bold**. Our proposed LSA employs the ViT-B/16 following [18, 19, 26] and the performance with ViT-B/16 achieves the best on two F-OOD benchmarks.

Benchmark: ImageNet-200						
Backbones	FPR@95 ↓		AUROC ↑		AUPR-IN ↑	
	Near-OOD	Far-OOD	Near-OOD	Far-OOD	Near-OOD	Far-OOD
ResNet-50	**70.32**	**54.32**	**74.97**	**82.84**	**85.31**	**94.94**
ResNet-101	71.33	60.06	72.77	81.18	83.80	94.33
ViT-B/16	**52.73**	**32.41**	**85.41**	**90.97**	**90.26**	**97.49**
ViT-B/32	59.01	41.59	75.73	85.57	89.22	96.16
Benchmark: ImageNet-1K						
Backbones	FPR@95 ↓		AUROC ↑		AUPR-IN ↑	
	Near-OOD	Far-OOD	Near-OOD	Far-OOD	Near-OOD	Far-OOD
ResNet-50	**79.80**	**57.16**	**72.85**	**83.27**	**90.45**	**97.74**
ResNet-101	80.45	60.22	71.46	81.77	87.93	96.41
ViT-B/16	**70.56**	**48.06**	**78.22**	**86.85**	**93.05**	**98.10**
ViT-B/32	76.87	54.81	73.28	83.01	90.78	97.59