Adaptive Prompt Learning via Gaussian Outlier Synthesis for Out-of-distribution Detection

Anonymous ICCV submission

Paper ID 15742

Abstract

Out-of-distribution (OOD) detection aims to distinguish whether detected objects belong to known categories or not. Existing methods extract OOD samples from In-distribution (ID) data to regularize the model's decision boundaries. However, the decision boundaries are not adequately regularized due to the model's lack of knowledge about the distribution of OOD data. To address the above issue, we propose an Adaptive Prompt Learning framework via Gaussian Outlier Synthesis (APLGOS) for OOD detection. Specifically, we leverage the Vision-Language Model (VLM) to initialize learnable ID prompts by sampling standardized results from pre-defined Q&A pairs. Region-level prompts are synthesised in low-likelihood regions of class-conditional gaussian distributions. These prompts are then utilized to initialize learnable OOD prompts and optimized with adaptive prompt learning. Also, OOD pseudo-samples are synthesised via gaussian outlier synthesis. Similarity score between prompts and images is utilized to calculate contrastive learning loss in high-dimensional hidden space. The aforementioned methodology regularizes the model to learn more compact decision boundaries for ID and OOD categories. Extensive experiments show that our proposed method achieves state-of-the-art performance with less ID data on four mainstream datasets.

1. Introduction

Deep learning has made significant progress in recent years. It encompasses a multitude of research domains, including object detection [39, 51], autonomous driving [40, 50] and image generation [20, 41]. Various existing deep learning methods rely on large-scale datasets to regularize the model, enabling it to learn sufficient data distribution and supervision signals of the training data. In real-world scenarios, where the number of unknown categories is significantly greater than that in the training dataset, the model lacks knowledge about the distribution of unknown data in prac-

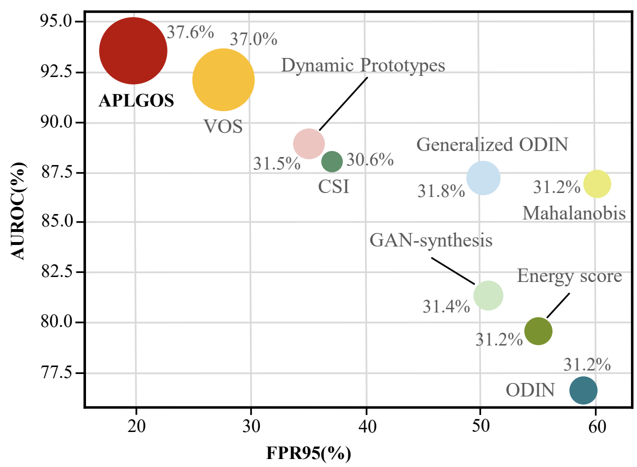

Figure 1. Quantitative comparisons with state-of-the-art OOD detection methods in terms of FPR95, AUROC and mAP metrics. Note that larger points denote higher mAP, and the numerical values are also given next to each point. Our APLGOS provides remarkable performance boost on all the metrics.

tical applications and struggles to learn compact decision boundaries that effectively distinguish between known and unknown categories. During the testing phase, unknown categories is likely to result in erroneous predictions accompanied by a high confidence score. This leads to severe safety risks in critical safety domains such as autonomous driving.

OOD detection [5, 21, 23, 32] is a research hotspot in recent years, which aims to enable the detectors to accurately distinguish not only seen categories, but also unseen categories during training. The detectors need to learn compact decision boundaries during training, ensuring low uncertainty for ID categories while maintaining high uncertainty away from them. To achieve this, existing OOD detection methods [12, 13, 33–35] provide sufficient supervision of OOD data for model training by extracting OOD pseudo-samples from ID data, helping the model better distinguish between known and unknown categories. However, due to the unpredictable quality of OOD pseudo-samples extracted from the ID data and the requirement for a large volume of ID data, the detector is not adequately regularized to learn compact decision boundaries for both ID and OOD categories. Therefore, synthesis-based methods [6, 18] have been proposed to generate OOD pseudo-samples. They synthesize out-of-distribution RGB images directly or virtual outliers in lower-dimensional hidden space, which to some extent mitigate the limitations of extracting pseudo-samples from ID data. In recent years, Vision-Language Models (VLMs) [29, 30, 47], owing to their powerful pre-trained knowledge and representation capabilities, have achieved considerable success and been applied across numerous fields.

In this paper, we propose APLGOS, a synthesis-based vision-language method that leverages the powerful learning and representation capabilities of VLMs to assist in synthesizing virtual outliers using ID data. APLGOS mainly consists of Prompt Learning Module (PLM) and Text-Image Alignment Module (TAM). PLM employs two distinct strategies to generate ID prompts and OOD pseudo-prompts, respectively, to assist in regularizing the model. For ID data, we first provide a pre-defined Q&A pair and templates with location and category names, e.g., "***Q:*** *What is in the region with coordinates <loc1>,<loc2>,<loc3>,<loc4>?* ***A:*** *That's a <CLS>*.", guiding the detector to incorporate location coordinates for more fine-grained observation. We guide ChatGPT through multiple rounds of standardization for the aforementioned prompts to generate a set of statements for the model to sample during training. The statements sampled from this set are then directly used to initialize the learnable ID prompts. In order to ensure that the generated OOD pseudo-samples better fit the distribution of OOD data, PLM generates OOD prompts through adaptive prompt learning via Gaussian Outlier Synthesis, where it samples virtual OOD prompts in the low-likelihood region of the class-conditional Gaussian distribution of ID prompts in high-dimensional hidden space. TAM calculates similarity scores for images and prompts and combines contrastive learning to align multimodal data, thereby regularizing model's decision boundaries.

In summary, the key contributions of this paper are as follows:

- We propose a vision-language OOD detection model namely APLGOS. Through adaptive prompt learning, APLGOS generates adaptive region-level prompts for ID and OOD images. Based on contrastive learning, APLGOS calculates similarity for images and prompts to ensure model learn compact decision boundaries.
- ID prompts, OOD prompts and OOD images are all virtual. ChatGPT standardizes pre-defined Q&A pairs with templates and instructions. Then we sample them to initialize learnable ID prompts. We synthesise virtual OOD prompts and OOD images in low-likelihood regions of class-conditional gaussian distribution.
- Extensive experiments on mainstream datasets show that APLGOS achieves state-of-the-art performance in terms of FPR95, AUROC, AUPR and mAP metrics. Compared to the baseline method [6], when using Berkley DeepDrive-100k as ID dataset and OpenImages as OOD dataset, our method reduces FPR95 by 7.76%.

2. Related Work

2.1. Out-of-distribution Detection

OOD detection [14, 25, 35, 38] aims to learn a compact decision boundary on training data that allows model to detect not only the categories with low uncertainty, that have been seen in training phase, but also the unseen categories with high uncertainty. Since in physical world, the number of unseen categories for the model is much bigger than seen categories, using large-scale dataset to regularize the model [13, 26] is difficult to fully cover all unseen categories of physical world. Liang *et al*. [22] use temperature scaling and add small perturbations to the input to separate the softmax score distributions between ID and OOD images. Based on energy theory [17], the work [25] replace traditional softmax score with energy score to distinguish ID and OOD images. Recently, outlier based methods are proposed, which utilize outliers exposure [28, 46] or generate virtual outliers in pixel [9, 18] space or hidden feature space [6] to regularize the model. Nevertheless, they are inefficient and the quality of the synthesised virtual outliers is worrying. With the emergence of vision-language models,vision-language model-based methods are proposed to address open-vocabulary problems [27, 36, 38]. To the best of our knowledge, no prior work has explored the use of prompt learning in OOD detection task.

2.2. Prompt Learning

Prompt learning is to view pre-trained language models, such as BERT [4], GPT [1, 2] and BLOOM [16] as knowledge bases, and use them to provide text prompts to optimize the performance of downstream tasks. In contrast to hand-designed prompts, the goal of prompt learning is to adaptively provide accurate prompts for downstream tasks. Zhou *et al*. [49] propose CoOp, which models a prompt's context words with learnable vectors while keeping pre-trained parameters fixed. To prevent CoOp from overfitting base classes, Zhou *et al*. [48] introduce CoCoOp, which uses conditional context optimization to generate an input-conditional token for each image, but this approach introduces high computational costs.At the same time, due to the effectiveness of prompt learning, there are various methods incorporating it with computer vision tasks [3, 8, 44, 45].

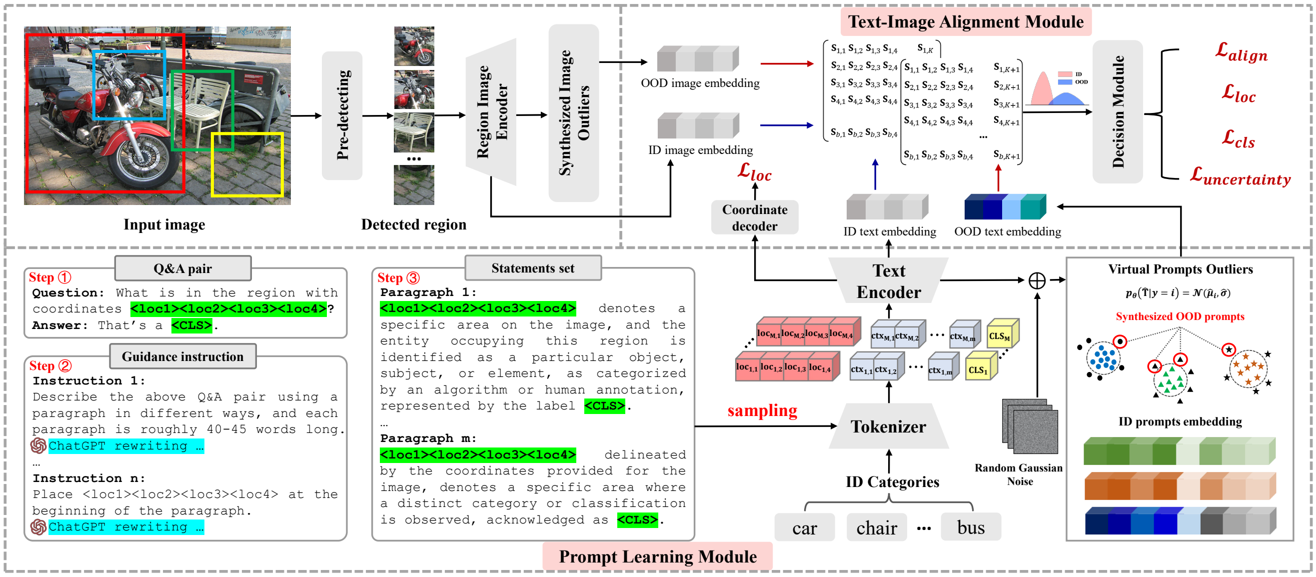

Figure 2. The proposed APLGOS network architecture. Prompt learning module is responsible for using ChatGPT to standardize Q&A pairs with guidance introduction and templates, then it generates a statements set. The module samples prompts from the statements set to initialize the learnable ID prompts, and synthesises virtual OOD prompts in low-likelihood regions of class-conditional gaussian distributions. The Text-Image Alignment Module computes similarity scores to align text and image embeddings in the hidden space.

3. Methodology

We propose an Adaptive Prompt Learning framework via Gaussian Outlier Synthesis for OOD Detection. As shown in Figure 2, APLGOS mainly consists of two modules, *i.e.* PLM and TAM. PLM leverages ChatGPT to standardize pre-defined Q&A pairs using guidance instructions and predefined templates, generating a set of statements. During training, PLM samples statements from this set to initialize the learnable prompts. For ID categories, APLGOS directly employs the initialized prompts as input to the text encoder, whereas for OOD categories, it synthesizes virtual OOD prompts and images within the low-likelihood region of the class-conditional Gaussian distribution of ID classes in the hidden space. Notably, only ID images are sourced from the dataset, while ID prompts, OOD prompts, and OOD images are all virtual and synthesized. This approach enables the model to enhance the quality of pseudo-samples with less ID data while better capturing the distribution of OOD data. Additionally, through contrastive learning, TAM computes similarity scores to align images and prompts within the high-dimensional hidden space.

For clarity, we omit the *batchsize* of data in the following description and consider a single batch as an example. The input to APLGOS consists of two modalities: detected region images $[\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_b]$ extracted from a raw RGB image $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, and text prompts $\mathbf{T} \in \mathbb{R}^{b\times l}$. Here, C, H, and W denote the number of channels, height, and width of the image, respectively. b represents the number of detected region images from a single raw RGB image. l indicates the length of the text prompts. The text input is given as $\mathbf{T} = [\mathbf{T}_1, \mathbf{T}_2, ..., \mathbf{T}_b]$, where the <CLS> token in the sampled prompts has been replaced with the corresponding labels.

3.1. Prompt Learning Module

ID Prompts. To enhance the model's representation ability and more effectively regularize its decision boundaries, we generate a set of statements for the Prompt Learning Module to sample from, rather than using a single invariant statement to initialize the learnable ID prompts. Specifically, we first predefine a Q&A pair, such as "***Q:*** *What is in the region with coordinates <loc1>,<loc2>,<loc3>,<loc4>?* ***A:*** *That's a <CLS>*.". We then input this Q&A pair into ChatGPT for standardization. During this process, we provide predefined templates and guiding instructions to ensure that ChatGPT standardizes the Q&A pair accordingly. The standardization process is illustrated below with an example prompt:

$$\Omega_0 = g(\mathrm{Q}^{\mathrm{A}} + \mathrm{M} + \mathrm{G}_0),\;\; \Omega_i = g(\Omega_{i-1} + \mathrm{G}_i), \tag{1}$$

where Ω_i denotes generated prompt result in i_{th} round, Q^{A} denotes Q&A pair, M denotes predefined template, G_i denotes guidance instruction for i_{th} standardizing round and g is ChatGPT's standardizing operation. We collect the statements from these t rounds to obtain statements set Ω_t. These statements are then used for sampling during the initialization of learnable ID prompts.

We introduce no extra character sets and vocabularies, and the generated prompts are represented in natural language. The learnable prompts follow the paradigm *e.g.* $<loc_1><loc_2><loc_3><loc_4><V_1><V_2>\cdots\cdots<V_m>$ $<CLS>$, which is initialized by sampled prompt. $<loc_1><loc_2><loc_3><loc_4>$ are learnable location tokens, which implicitly introduce location information into the prompts. $<V_1><V_2>\cdots<V_m>$ are learnable description tokens, and m is its length. <CLS>is class token.

$$\hat{\mathbf{T}} = f_\theta(h(r(g(\Omega_{t-1} + \mathrm{G}_t)))), \tag{2}$$

where $\hat{\mathbf{T}} = [\hat{\mathbf{T}}_1, \hat{\mathbf{T}}_2, ..., \hat{\mathbf{T}}_b]$, $\hat{\mathbf{T}}_i \in \mathbb{R}^{l'}$, t is rounds of standardizing operations, l' is length of prompt embedding. Here, for ease of understanding, we use one $\hat{\mathbf{T}}_i$ as an example to describe the subsequent operations, and standardize $\hat{\mathbf{T}}_i$ as $\hat{\mathbf{T}}$, $\hat{\mathbf{T}} \in \mathbb{R}^{l'}$. f_θ is transformer-based text encoder, h is tokenizer, r is replacement function for <CLS> token. We replace <CLS> directly with the category label of the object in the current region (i.e., the corresponding ID class label).

OOD Prompts. In the hidden space, distinct decision boundaries should be established between ID and OOD prompts. In the OOD detection task, we refine the decision boundaries as much as possible. By incorporating prompt learning, we synthesize region-level OOD pseudo-prompts using Gaussian outlier synthesis. Specifically, the Prompt Learning Module synthesizes virtual OOD prompts in the low-likelihood regions of class-conditional Gaussian distributions in hidden space. This allows the Text-Image Alignment Module to perceive the distribution difference between ID and OOD categories in hidden space and align images and prompts through contrastive learning. Provided that the quantity of data is large enough, we assume the ID prompts embedding from text encoder form a class-conditional multivariate Gaussian distribution:

$$p_\theta(\hat{\mathbf{T}}|y = i) = \mathcal{N}(\hat{\mu}_i, \hat{\sigma}), \tag{3}$$

where θ is the parameter of text encoder f_θ, y is ground truth label, $\hat{\mu}_i$ is empirical gaussian mean of i_{th} in-distribution category prompts embedding, and $i \in \{1, 2, ..., K\}$, K represents the number of in-distribution classes, $\mathcal{N}(\hat{\mu}_i, \hat{\sigma}) = \frac{1}{\sqrt{2\pi}\hat{\sigma}} e^{-\frac{(\hat{\mathbf{T}}-\hat{\mu}_i)^2}{2\hat{\sigma}^2}}$, $\hat{\sigma}$ denotes the tied covariance matrix.

First, we calculate the empirical gaussian mean of i_{th} ID category prompts embedding as follows:

$$\hat{\mu}_i = \frac{1}{|\mathcal{Q}_T|} \sum_{j=1}^{|\mathcal{Q}_T|} \hat{\mathbf{T}}_{i,j}, \tag{4}$$

where $|\mathcal{Q}_T|$ denotes the length of the prompts queue $\mathcal{Q}_T$ used to buffer ID prompts, and $\mathcal{Q}_T \in \mathbb{R}^{K \times |\mathcal{Q}_T|}$.

Then we calculate the tied covariance matrix of ID prompts embedding as follows:

$$\hat{\sigma} = \frac{1}{K|\mathcal{Q}_T|} \sum_{i=1}^{K} \sum_{j=1}^{|\mathcal{Q}_T|} (\hat{\mathbf{T}}_{i,j} + \alpha\varepsilon - \hat{\mu}_i)(\hat{\mathbf{T}}_{i,j} + \alpha\varepsilon - \hat{\mu}_i)^T + \beta\mathbf{E}, \tag{5}$$

where ε is learnable matrix initialized by randomly gaussian noise, $\mathbf{E}$ is unit matrix, α, β are hyper-parameters, $\hat{\sigma}$ is tied covariance matrix, and $\hat{\sigma} = [\hat{\sigma}_1, \hat{\sigma}_2, ..., \hat{\sigma}_K]^T$.

After computing the empirical Gaussian mean $\hat{\mu}$ and the tied covariance matrix $\hat{\sigma}$, the Prompt Learning Module samples virtual OOD prompts from the low-likelihood regions of the class-conditional Gaussian distributions in hidden space, based on the estimated multivariate distributions. Then, it selects the top-k prompts with the lowest probability from this ϵ-likelihood region:

$$\mathcal{V}_i = \Psi(\hat{\mathbf{T}}, \hat{\mu}_i, \hat{\sigma}), \tag{6}$$

where Ψ is class-conditional gaussian distribution probability density and satisfies the following relation:

$$\Psi(\hat{\mathbf{T}}, \hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_K, \hat{\sigma}) = \Psi(\hat{\mathbf{T}}, \hat{\mu}_1, \hat{\sigma})\Psi(\hat{\mathbf{T}}, \hat{\mu}_2, \hat{\sigma}) \cdots \Psi(\hat{\mathbf{T}}, \hat{\mu}_K, \hat{\sigma}), \tag{7}$$

For each $\Psi(\hat{\mathbf{T}}, \hat{\mu}_i, \hat{\sigma})$, its expansion can be formulated as:

$$\Psi(\hat{\mathbf{T}}, \hat{\mu}_i, \hat{\sigma}) = \{v_i | \frac{1}{\sqrt{2\pi}^{\frac{l'}{2}} |\hat{\sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(v_i - \hat{\mu}_i)^T \hat{\sigma}^{-} (v_i - \hat{\mu}_i)} < \epsilon\}, \tag{8}$$

where $v_i \sim \mathcal{N}(\hat{\mu}_i, \hat{\sigma})$ denotes sampled virtual prompt using i_{th} ID category prompts, $i = \{1, 2, ..., K\}$, and "$-$" denotes matrix inverse operation. The final synthesised OOD prompts are denoted as $\hat{\mathbf{T}}^\dagger$.

3.2. OOD Virtual Images Synthesis

Existing methods [12, 13, 33–35] directly extract OOD pseudo-samples from ID data. However, the extracted pseudo-samples are unable to fit the distribution of OOD data adequately. In this paper, we also use synthesis method to get OOD data. The principle of synthesizing OOD image is similar to Eq. 3 to Eq. 8. Compared with synthesizing OOD prompts, the input for calculating the empirical Gaussian mean and tied covariance is ID image embedding instead of ID prompts embedding. We define the final synthesised virtual images using current ID image embedding queue $\mathcal{Q}_I$ as $\hat{\mathbf{X}}^\dagger$.

3.3. Text-Image Alignment Module

We first encode ID and OOD images and prompts to generate their embeddings. Then, the similarity score between prompts embedding and image embeddings is computed as follows:

$$\mathbf{S} = \frac{||\hat{\mathbf{X}}||_p(||\hat{\mathbf{T}}||_p)^T}{e^{\omega}}, \tag{9}$$

where $\hat{\mathbf{X}}$ is embedding of detected region images in the second training phase, and $\hat{\mathbf{X}} = [\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, ..., \hat{\mathbf{X}}_b]$, $\hat{\mathbf{X}}_i$ is one detected region image embedding, $\hat{\mathbf{X}}_i \in \mathbb{R}^{l'}$. In the third training phase, the input is embedding of synthesised virtual image $\hat{\mathbf{X}}^{\dagger}$ instead of $\hat{\mathbf{X}}$, ω is hyper-parameters for scaling. $\mathbf{S}$ is similarity score. The prompts embedding in Eq. 9 is ID prompts embedding $\hat{\mathbf{T}}$ in the second phase and synthesised OOD prompts embedding $\hat{\mathbf{T}}^{\dagger}$ in the third phase, $||\cdot||_p$ is normalization, in addition, $||\hat{\mathbf{X}}_i||_p = \hat{\mathbf{X}}_i / \sqrt{\sum_{j=1}^{l'} |\hat{\mathbf{X}}_{i,j}|^2}$ and $||\hat{\mathbf{T}}_i^{\dagger}||_p = \hat{\mathbf{T}}_i^{\dagger} / \sqrt{\sum_{j=1}^{l'} |\hat{\mathbf{T}}_{i,j}^{\dagger}|^2}$.

3.4. Loss Function

Alignment loss $\mathcal{L}_{align}$ constrains the contrastive learning process during alignment, receiving ID or OOD data at different training phases. The similarity score between prompts embedding and image embeddings is used to calculate the alignment loss:

$$\mathcal{L}_{align}(\mathbf{S}, y) = -\sum_{i=1}^{K'} t_i log(\mathcal{R}_i(\mathbf{S})), \tag{10}$$

where t_i represents category label of the object contained in currently detected region. $\mathcal{R}_i$ represents the standardized prediction score. We treat all OOD categories as a single category, *i.e.*, "background". During the training phase, if the ID dataset contains a total of K classes, each detected region image is required to calculate similarity scores with $(K+1)$ text prompts, i.e., $K' = K + 1$.

Previous methods typically generate simple prompts that lack location information, such as "*a photo of a <CLS>*" [48, 49], or provide brief prompts with relative location information for the entire image [42]. We argue that these prompts lack the fine granularity needed for the model to learn essential location information in vision-language-based detection tasks. $\mathcal{L}_{loc}$ is designed to implicitly incorporate location information, enabling the generation of fine-grained prompts for detected image regions.

$$\mathcal{L}_{loc} = \frac{\lambda}{\Phi(\mathbf{B}_g)}[\sum_{i=1}^{z}(\sqrt{\mathbf{B}_{g_i}} - \sqrt{u(\mathbf{B}_r)_i})^2]^{\frac{1}{2}}, \tag{11}$$

where $\mathbf{B}_g$ represents ground truth coordinates of detected image region, $\mathbf{B}_r$ represents regression results of coordinates, and $\mathbf{B}_g \in \mathbb{R}^{b\times 4}$, $\mathbf{B}_r \in \mathbb{R}^{b\times 4}$, $z = 4$, u represents calculating absolute values, Φ represents calculating the dimension of vector, λ is hyper-parameter.

After incorporating the classification loss $\mathcal{L}_{cls}$ and the location loss $\mathcal{L}_{loc}$, the total loss can be expressed as:

$$\begin{aligned}\mathcal{L} =& \xi_1[\gamma_1\tau\mathcal{L}_{align}^{id} + \gamma_2(1-\tau)\mathcal{L}_{align}^{ood}] \\ &+ \gamma_3\xi_2[\kappa\mathcal{L}_{loc}^{id} + (1-\kappa)\mathcal{L}_{loc}^{ood}] \\ &+ \gamma_4\xi_3\mathcal{L}_{cls} + \gamma_5\xi_4\mathcal{L}_{reg} + \overline{\mathcal{W}}.\end{aligned} \tag{12}$$

Note that γ_1,γ_2,γ_3,γ_4,γ_5 are the hyper-parameters, ξ,τ,κ determine the loss functions used in the current training phase and $\xi_i = \{0, 1\}$, $\tau = \{0, 1\}$. In order to better regularize the model, in the actual implementation of $\mathcal{L}$, we also add the regularization term $\mathcal{W}$, and $\mathcal{W}_i = [\Delta^2_{(\mathcal{F}(\mathbf{O}_1),\mathcal{B}_1)_i} + \Delta^2_{(\mathcal{G}(\mathbf{O}_2),\mathcal{B}_2)_i}]$, $\overline{\mathcal{W}} = 1/N\sum_{i=1}^{N}\mathcal{W}_i$, $\Delta^2_{(a,b)}$ represents $(a-b)^2$, $\mathcal{F}, \mathcal{G}$ represent regression blocks, $\mathbf{O}_i$ represents regularization matrix, $\mathcal{B}_i$ represents bias matrix of regression block, $i = \{1, 2\}$.

4. Experiments

4.1. Datasets

We verify our proposed APLGOS on four commonly used datasets: PASCAL VOC, Berkley DeepDrive-100k, MS-COCO2017 and OpenImages. The PASCAL VOC [7] dataset contains 9963 images in 20 categories, split into 5011 training and 4952 test images, with a resolution of 500×375 (375×500). The BDD-100k [43] dataset consists of 100,000 high-resolution driving scenarios with detailed road object annotations. The MS-COCO2017 [24] dataset includes 328,000 images across 91 categories and 2.5 million instance tags, with 82 categories having more than 5000 tags. OpenImages V4 [15] contains 9.2 million images across 500 categories, commonly used for classification, object detection, and visual relationship detection. The above four datasets comprehensively evaluate our proposed method from different aspects and perspectives.

4.2. Implementation Details

We employ transformer as the backbone for the text encoder in the Prompt Learning Module. For the image encoder, we employ ResNet50 [10] and RegNetX4.0 [31] as backbones, respectively. We use ChatGPT-3.5 to standardize Q&A pairs. The ratio of ID data used for training to synthesised OOD data is approximately 1:1. We use PASCAL VOC and Berkeley DeepDrive-100K as ID datasets, and evaluate on two OOD datasets containing subsets randomly sampled from MS-COCO2017 and OpenImages, respectively. To ensure the fairness of the test, we manually exclude the categories in the OOD dataset that overlap with those in the ID dataset before evaluating on the OOD dataset. We set $B = 16$ and train APLGOS on PASCAL VOC for 18,000 iterations, and set $B = 8$ to train on Berkeley DeepDrive-100k for 90,000 iterations. We set the learning rate $lr = 0.01$. The length of prompt embedding and

ID Dataset	Method	FPR95 ↓	AUROC ↑	AUPR ↑	mAP (ID) ↑
		OOD: MS-COCO2017 / OpenImages			
PASCAL VOC	MSP [11]	70.99 / 73.13	83.45 / 81.91	-	48.7
	ODIN [22]	59.82 / 63.14	82.20 / 82.59	-	48.7
	Mahalanobis [19]	96.46 / 96.27	59.25 / 57.42	-	48.7
	Energy score [25]	56.89 / 58.69	83.69 / 82.98	-	48.7
	Gram matrices [32]	62.75 / 67.42	79.88 / 77.62	-	48.7
	Generalized ODIN [14]	59.57 / 70.28	83.12 / 79.23	-	48.1
	CSI [35]	59.91 / 57.41	81.83 / 82.95	-	48.1
	GAN-synthesis [18]	60.93 / 59.97	83.67 / 82.67	-	48.5
	VOS-ResNet50* [6]	48.28 / 52.14	87.65 / 85.3	98.76 / 96.98	47.8
	VOS-RegX4.0* [6]	50.53 / 50.27	88.10 / 87.08	98.92 / 97.80	49.1
	APLGOS (ResNet50)	**47.16 / 49.66**	**87.89 / 85.91**	**98.80 / 97.54**	**48.8**
	APLGOS (RegNetX4.0)	**45.96 / 47.10**	**89.19 / 88.49**	**99.00 / 98.30**	**49.4**
Berkeley DeepDrive-100k	MSP [11]	80.94 / 79.04	75.87 / 77.38	-	31.2
	ODIN [22]	62.85 / 58.92	74.44 / 76.61	-	31.2
	Mahalanobis [19]	57.66 / 60.16	84.92 / 86.88	-	31.2
	Energy score [25]	60.06 / 54.97	77.48 / 79.60	-	31.2
	Gram matrices [32]	60.93 / 77.55	74.93 / 59.38	-	31.2
	Generalized ODIN [14]	57.27 / 50.17	85.22 / 87.18	-	31.8
	CSI [35]	47.10 / 37.06	84.09 / 87.99	-	30.6
	GAN-synthesis [18]	57.03 / 50.61	78.82 / 81.25	-	31.4
	VOS-ResNet50* [6]	46.97 / 31.25	84.97 / 89.82	99.67 / 99.86	35.7
	VOS-RegX4.0* [6]	42.82 / 27.55	86.36 / 92.11	99.76 / 99.93	37.0
	Dynamic Prototypes [37]	45.72 / 35.05	85.14 / 88.92	-	31.5
	APLGOS (ResNet50)	**41.10 / 23.30**	**87.36 / 92.87**	**99.73 / 99.89**	**35.8**
	APLGOS (RegNetX4.0)	**39.48 / 19.79**	**87.47 / 93.59**	**99.79 / 99.94**	**37.6**

Table 1. Comparison with the state-of-the-art methods on mainstream datasets. Here we use PASCAL VOC and Berkeley DeepDrive-100k as ID datasets, MS-COCO2017 and OpenImages as OOD datasets, respectively. "-" denotes that the data is not available.

	Strategy	FPR95 ↓	AUROC ↑	AUPR ↑	mAP (ID) ↑
		OOD: MS-COCO2017 / OpenImages			
(a)	VOS-RegNetX4.0* [6]	50.53 / 50.27	88.10 / 87.08	98.82 / 97.80	49.1
(b)	[6] + <CLS>	50.12 / 49.50	88.56 / 86.83	98.91 / 97.79	48.2
(c)	[6] + "a region of a" + <CLS>	51.31 / 50.96	88.20 / 86.73	98.98 / 97.85	48.7
(d)	[6] + RP + <CLS>	49.50 / 49.40	88.49 / 86.73	98.82 / 97.77	48.9
(e)	[6] + <LOC>+ "a region of a" + <CLS>	49.56 / 47.60	88.23 / 87.07	98.89 / 97.87	49.1
(f)	[6] + <LOC>+ RP + <CLS> **(Ours)**	**45.96 / 47.10**	**89.19 / 88.49**	**99.00 / 98.30**	**49.4**

Table 2. Ablation studies for prompt strategies. "+" denotes the combination of strategies. "RP" represents sampled prompts from statements set, which is standardized by ChatGPT using Q&A pair and guidance instructions. **(b)** denotes the simplest prompt strategy, i.e., only providing the ground-truth label for the ID data, (for synthesised OOD image, we define its label as "background"). **(c)** denotes the original prompt strategy of CLIP [30]. **(d)** denotes that we replace the prompts in CLIP [30] with the statements by ChatGPT standardizing the Q&A pairs. **(e)** denotes adding location tokens <LOC> to **(c)**. **(f)** represents the prompts of our proposed APLGOS.

length of image embedding $l^{'} = 1024$. We use 1000 samples to estimate the class-conditional Gaussian distribution of ID image embeddings and 10000 samples for ID prompts embedding (i.e., $|\mathcal{Q}_I| = 1000$, $|\mathcal{Q}_T| = 10000$). The total length l of the standardized Q&A pair does not exceed 77. In the experimental tables, "*" denotes results from local replication based on open-source code. "↓" indicates that a smaller value is better, while "↑" indicates that a greater value is better.

4.3. Comparison with The State-of-the-Art

We report the results of our proposed framework with different image encoder backbones (ResNet50 and RegNetX4.0) on PASCAL VOC, Berkeley DeepDrive-100k, MS-COCO2017, and OpenImages datasets, as shown in Table 1. The best results for the same dataset and the same backbone settings are shown in **bold**. For the same evaluation metric on the same dataset, the best results are underlined. When using Transformer-based text encoder and ResNet50-based image encoder, APLGOS achieves an FPR95 of 47.16% and an mAP of 48.8% on PASCAL VOC (ID) with MS-COCO2017 as the OOD dataset. When OpenImages is used as the OOD dataset, FPR95 increases to 49.66%. Compared to the state-of-the-art OOD detection model [6], APLGOS reduces FPR95 by 1.12% and 2.48% on MS-COCO2017 and OpenImages, respectively.

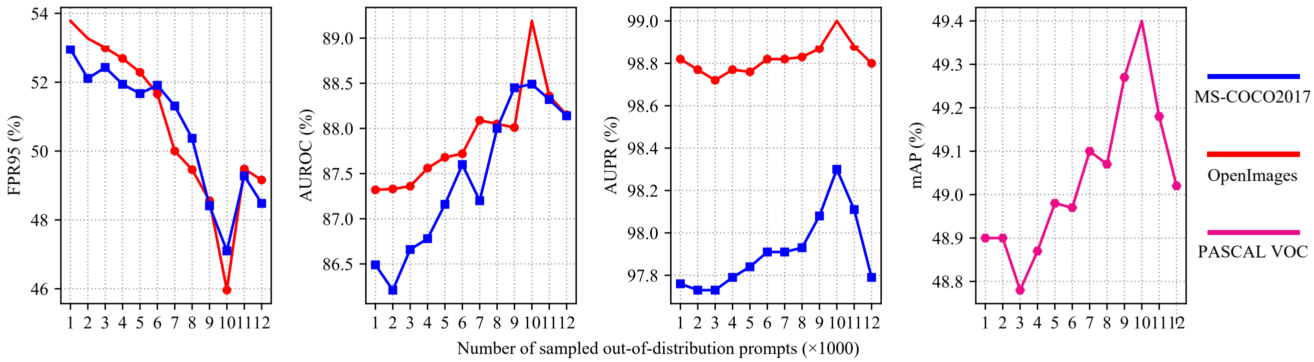

Figure 3. Ablation on number of sampled OOD prompts $\mathcal{K}$. The horizontal coordinate is the number of sampled ood prompts $\mathcal{K}$ ($\times 10^3$), while the vertical coordinates are, from left to right, FPR95, AUROC, AUPR, and mAP, respectively. **Red line** and **Blue line** represent using MS-COCO2017 and OpenImages as OOD datasets, respectively. **Pink line** represents using PASCAL VOC as ID dataset.

α	FPR95 ↓	AUROC ↑	mAP (ID) ↑
	OOD: MS-COCO2017 / OpenImages		
0	51.63 / 50.88	87.86 / 87.24	49.2
0.5	51.90 / 51.48	87.55 / 87.02	48.9
1.0	**45.96 / 47.10**	**89.19 / 88.49**	**49.4**
1.5	55.88 / 53.33	86.29 / 86.75	48.9
2.0	55.92 / 49.54	86.75 / 88.00	48.9

Table 3. The Ablation Experiments on The Strength of Random Gaussian Noise ε. α represents the strength of added gaussian noise. The value of α increases gradually from 0 to 2.0, and we take the value at 0.5 intervals.

Γ_1	FPR95 ↓	AUROC ↑	mAP (ID) ↑
	OOD: MS-COCO2017 / OpenImages		
1:4	50.11 / 58.38	87.71 / 85.67	49.1
1:3	49.40 / 55.12	87.91 / 86.38	49.2
1:2	47.98 / 54.49	88.40 / 85.94	49.2
1:1	**45.96 / 47.10**	**89.19 / 88.49**	**49.4**
2:1	48.25 / 50.20	88.30 / 87.76	49.2
3:1	50.95 / 53.94	86.81 / 84.70	47.5
4:1	50.20 / 51.56	86.70 / 84.89	47.3

Table 4. The ablation experiments on the ratio Γ_1 of ID and OOD data used during training. Our default parameters and results are shown in **bold**. Parameters and results of baseline [6] are shown with a dark base color.

With Transformer-based text encoder and RegNetX4.0-based image encoder, FPR95 decreases to 45.96% on MS-COCO2017 and 47.1% on OpenImages, while the mAP on PASCAL VOC improves to 49.4%. This setup further reduces FPR95 by 4.57% and 3.17% on MS-COCO2017 and OpenImages, respectively, compared to [6]. For Berkley DeepDrive-100k (ID), using ResNet50-based image encoder and Transformer-based text encoder, APLGOS achieves an FPR95 of 41.10% on MS-COCO2017 and 23.30% on OpenImages, with an mAP of 35.8%. When using RegNetX4.0-based image encoder instead, FPR95 further decreases to 39.48% on MS-COCO2017 and 19.79% on OpenImages, while mAP improves to 37.6%.

4.4. Ablation Studies

Prompt strategies. To further validate the effectiveness of our prompt strategies, we conduct extensive ablation experiments on APLGOS's prompt strategies, and the results are shown in Table 2. Sampling from the statements set brings greater performance gains than simply initializing learnable prompts with "*a region of a*" (**(c)** ***vs*** **(d)**). Moreover, adding location tokens to prompts significantly improves performance, as it refines the scope of the prompts (**(c)** ***vs*** **(e)**). Compared to other prompt strategies, our APLGOS prompt strategy **(f)** integrates the advantages of the aforementioned strategies and achieves the best performance.

Number of Sampled OOD Prompts. APLGOS synthesises virtual prompts for OOD categories and for each ID category, APLGOS samples $\mathcal{K}$ virtual OOD prompts in low-likelihood regions of ID class-conditional gaussian distributions in high-dimensional hidden space. We conduct ablation experiments on $\mathcal{K}$, the results of its effect on performance are shown in Figure 3. When $\mathcal{K}$ is too small, it may fail to adequately cover the region outside the ID categories' decision boundaries in the hidden space. On the other hand, when $\mathcal{K}$ is too large, the excessive randomness in the sampled OOD prompts makes it difficult to effectively regularize the decision boundaries with the limited model parameters. Therefore, we set $\mathcal{K} = 10000$ as the default value.

Strength of Random Gaussian Noise ε. To enhance the size and diversity of the OOD prompts embedding sampling space and prevent the model from overly relying on the ID category distribution, we introduce a learnable matrix initialized with random Gaussian noise ε during the OOD prompt sampling stage (Eq. 5). We conduct ablation experiments on its strength α, and the results are shown in Table 3. A small value of α makes the sampling space of OOD prompts embedding too narrow, while a large value

Figure 4. Detection results on ID dataset. Here we use Berkley DeepDrive-100k dataset as ID dataset. We use RegNetX4.0 and Transformer as backbone. The **first row** is the detection results of baseline [6]. The **second row** is the detection result of our APLGOS. Our APLGOS rarely misclassifies the ID class as OOD class, and there is almost no missed detection.

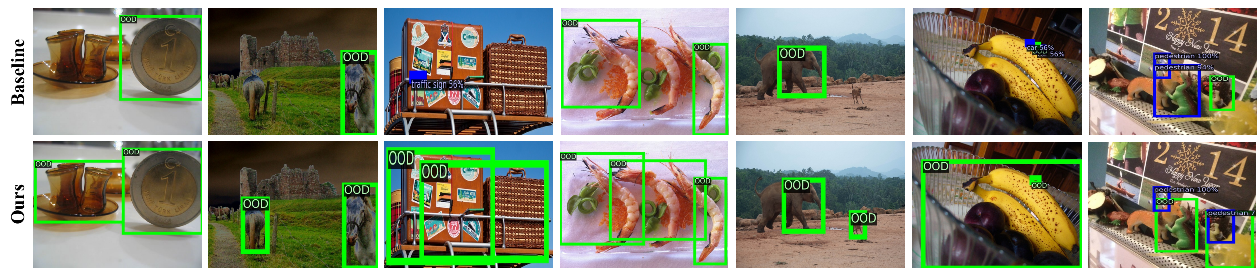

Figure 5. Detection results on OOD datasets. Here we use Berkley DeepDrive-100k dataset as ID dataset, MS-COCO2017 and OpenImages as OOD datasets. The **first row** is the detection results of baseline [6]. The **second row** is the detection results of our APLGOS. Compared to the baseline, APLGOS rarely misses detections and hardly produces overlapping boxes for the same object.

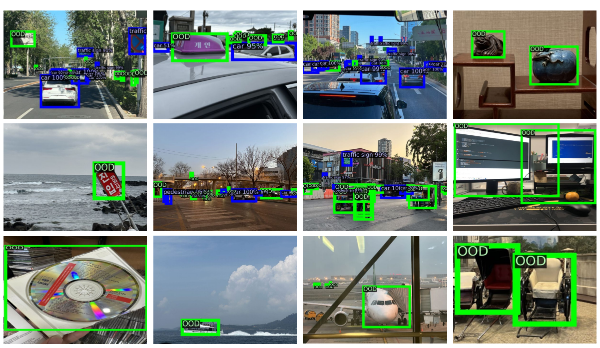

Figure 6. Detection results in Real World. Here we use Berkley DeepDrive-100k dataset as in-distribution dataset. Pictures we take ourselves with our phone as out-of-distribution dataset.

of α results in an overly large sampling space. Only by appropriately expanding the sampling space of OOD prompts embedding can the model's ability to fit the OOD distribution be effectively enhanced.

Ratio of ID and OOD Data Used During Training. To verify that APLGOS can achieve better performance with less ID data, we conduct ablation experiments on the amount of ID data used during training, and the results are shown in Table 4. By default, APLGOS adopts a ratio Γ_1 of 1:1 for ID and OOD data during training, whereas the baseline [6] uses a ratio of 2:1. However, in this case, the performance of APLGOS decreases instead.

Visualization of Detection Results. To better evaluate the performance of APLGOS, we visualize its detection results on ID datasets, OOD datasets, and real-world scenarios. The results are presented in Figures 4, 5 and 6. The images in real-world scenarios are captured using an iPhone 14 Pro Max. The visualization results demonstrate that APLGOS outperforms the baseline method in detecting ID and OOD categories. Moreover, the visualization of detection results in real-world scenarios further confirms its superior generalization ability.

5. Conclusion

In this paper, we propose a vision-language method, Adaptive Prompt Learning via Gaussian Outlier Synthesis (APLGOS) for Out-of-distribution Detection. Through prompt learning approach, APLGOS provides adaptive region-level prompts with location information for ID / OOD images. We use ChatGPT to standardize pre-defined Q&A pairs and generate a statements set. During training, only ID images are from the dataset, while ID prompts, OOD prompts and OOD images are all virtual. We sample statements from the statements set to initialize learnable ID prompts. We samples virtual OOD prompts and OOD images in the low-likelihood region of the class-conditional gaussian distribution in high-dimensional hidden space. Similarity score between prompts and images is utilized to calculate contrastive learning loss in high-dimensional hidden space, which guarantees the quality of virtual outliers as well as better regularization of the model. Through comprehensive experimental evaluations, we demonstrated the effectiveness of the proposed APLGOS.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems (NeurIPS)*, 33:1877–1901, 2020. 2

[3] L. Chen, G. Wang, L. Yuan, K. Wang, K. Deng, and P. Torr. Nerf-vpt: Learning novel view representations with neural radiance fields via view prompt tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1156–1164, 2024. 2

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[5] X. Du, Z Wang, M. Cai, and S Li. Towards unknown-aware learning with virtual outlier synthesis. In *International Conference on Learning Representations (ICLR)*, 2022. 1

[6] X. Du, Z. Wang, M. Cai, and Y. Li. Vos: Learning what you don't know by virtual outlier synthesis. *International Conference on Learning Representations (ICLR)*, 2022. 2, 6, 7, 8

[7] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88:303–338, 2010. 5

[8] Zhongbin Fang, Xiangtai Li, Xia Li, Joachim M Buhmann, Chen Change Loy, and Mengyuan Liu. Explore in-context learning for 3d point cloud understanding. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[9] M. Grcić, P. Bevandić, and S. Šegvić. Dense open-set recognition with synthetic outliers generated by real nvp. *arXiv preprint arXiv:2011.11094*, 2020. 2

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[11] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations (ICLR)*, 2016. 6

[12] D Hendrycks, M. Mazeika, and T. Dietterich. Deep anomaly detection with outlier exposure. *International Conference on Learning Representations (ICLR)*, 2018. 1, 4

[13] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, and D. Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019. 1, 2, 4

[14] Y. Hsu, Y. Shen, H. Jin, and Z. Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 10951–10960, 2020. 2, 6

[15] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision (IJCV)*, 128 (7):1956–1981, 2020. 5

[16] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. Luccioni, F. Yvon, M. Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. 2023. 2

[17] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 2

[18] K. Lee, H. Lee, K. Lee, and J. Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *International Conference on Learning Representations (ICLR)*, 2017. 2, 6

[19] K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems (NeurIPS)*, 31, 2018. 6

[20] D. Li, J. Li, and S. Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024. 1

[21] J. Li, P. Chen, Z. He, S. Yu, S. Liu, and J. Jia. Rethinking out-of-distribution (ood) detection: Masked image modeling is all you need. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 11578–11589, 2023. 1

[22] S. Liang, Y. Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *International Conference on Learning Representations (ICLR)*, 2017. 2, 6

[23] J. Liao, X. Xu, M. Nguyen, A. Goodge, and C. Foo. Coft-ad: Contrastive fine-tuning for few-shot anomaly detection. *IEEE Transactions on Image Processing (TIP)*, 2024. 1

[24] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 5

[25] W. Liu, X. Wang, J. Owens, and Y. Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems (NeurIPS)*, 33:21464–21475, 2020. 2, 6

[26] X. Liu, Y. Lochman, and C. Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23946–23955, 2023. 2

[27] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:35087–35102, 2022. 2

[28] S. Mohseni, M. Pitale, J. Yadawa, and Z. Wang. Self-supervised learning for generalizable out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 5216–5223, 2020. 2

[29] S. Ning, L. Qiu, Y. Liu, and X. He. Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23507–23517, 2023. 2

[30] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, pages 8748–8763, 2021. 2, 6

[31] I. Radosavovic, R. Kosaraju, R. Girshick, K. He, and P. Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10428–10436, 2020. 5

[32] C. Sastry and S. Oore. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning (ICML)*, pages 8491–8501, 2020. 1, 6

[33] B. Su, H. Zhang, and Z. Zhou. Hsic-based moving weight averaging for few-shot open-set object detection. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, pages 5358–5369, 2023. 1, 4

[34] B. Su, H. Zhang, J. Li, and Z. Zhou. Toward generalized few-shot open-set object detection. *IEEE Transactions on Image Processing (TIP)*, 33:1389–1402, 2024.

[35] J. Tack, S. Mo, J. Jeong, and J. Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems (NeurIPS)*, 33:11839–11852, 2020. 1, 2, 4, 6

[36] H. Wang, Y. Li, H. Yao, and X. Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1802–1812, 2023. 2

[37] A. Wu, C. Deng, and W. Liu. Unsupervised out-of-distribution object detection via pca-driven dynamic prototype enhancement. *IEEE Transactions on Image Processing (TIP)*, 2024. 6

[38] X. Wu, F. Zhu, R. Zhao, and H. Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 7031–7040, 2023. 2

[39] C. Xie, Z. Zhang, Y. Wu, F. Zhu, R. Zhao, and S. Liang. Described object detection: Liberating object detection with flexible expressions. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024. 1

[40] X. Xiong, Y. Liu, T. Yuan, Y. Wang, Y. Wang, and H. Zhao. Neural map prior for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17535–17544, 2023. 1

[41] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024. 1

[42] Z. Yang, Z. Gan, J. Wang, X. Hu, F. Ahmed, Z. Liu, Y. Lu, and L. Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision (ECCV)*, pages 521–539. Springer, 2022. 5

[43] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 2636–2645, 2020. 5

[44] Y. Zhai, Y. Zeng, Z. Huang, Z. Qin, X. Jin, and D. Cao. Multi-prompts learning with cross-modal alignment for attribute-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 6979–6987, 2024. 2

[45] C. Zhang, X. Chen, S. Chai, C. Wu, D. Lagun, T. Beeler, and F. De la Torre. Iti-gen: Inclusive text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3969–3980, 2023. 2

[46] J. Zhang, N. Inkawhich, Y. Chen, and H. Li. Fine-grained out-of-distribution detection with mixup outlier exposure. *arXiv preprint arXiv:2106.03917*, 2(5), 2021. 2

[47] J. Zhang, J. Huang, S. Jin, and S. Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024. 2

[48] K. Zhou, J. Yang, C. Loy, and Z. Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 16816–16825, 2022. 2, 5

[49] K. Zhou, J. Yang, C. Loy, and Z. Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 130(9):2337–2348, 2022. 2, 5

[50] Z. Zhu, Y. Zhang, H. Chen, Y. Dong, S. Zhao, W. Ding, J. Zhong, and S. Zheng. Understanding the robustness of 3d object detection with bird's-eye-view representations in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21600–21610, 2023. 1

[51] O. Zohar, K. Wang, and S. Yeung. Prob: Probabilistic objectness for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11444–11453, 2023. 1