RUNA: Object-level Out-of-Distribution Detection via Regional Uncertainty Alignment of Multimodal Representations

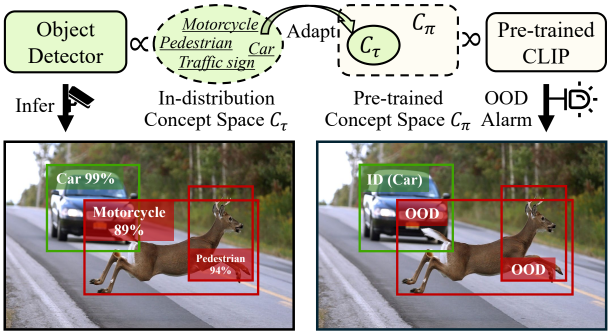
Bin Zhang[*1,2], **Jinggang Chen**[*1], **Xiaoyang Qu**[†2], **Guokuan Li**[1],
Kai Lu[†1], **Jiguang Wan**[1], **Jing Xiao**[2], **Jianzong Wang**[2]
[1]Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China
[2]Ping An Technology (Shenzhen) Co., Ltd, Shenzhen, China
{binz2398, chen.jinggang98, quxiaoy}@gmail.com, liguokuan@hust.edu.cn,
{kailu, jgwan}@hust.edu.cn, xiaojing661@pingan.com, jzwang@188.com

Abstract

Enabling object detectors to recognize out-of-distribution (OOD) objects is vital for building reliable systems. A primary obstacle stems from the fact that models frequently do not receive supervisory signals from unfamiliar data, leading to overly confident predictions regarding OOD objects. Despite previous progress that estimates OOD uncertainty based on the detection model and in-distribution (ID) samples, we explore using pre-trained vision-language representations for object-level OOD detection. We first discuss the limitations of applying image-level CLIP-based OOD detection methods to object-level scenarios. Building upon these insights, we propose RUNA, a novel framework that leverages a dual encoder architecture to capture rich contextual information and employs a regional uncertainty alignment mechanism to distinguish ID from OOD objects effectively. We introduce a few-shot fine-tuning approach that aligns region-level semantic representations to further improve the model's capability to discriminate between similar objects. Our experiments show that RUNA substantially surpasses state-of-the-art methods in object-level OOD detection, particularly in challenging scenarios with diverse and complex object instances.

Introduction

Identifying out-of-distribution (OOD) objects is vital for object detectors to safely deploy in an open-world environment. Most current models work in a closed-world environment, matching objects to the pre-defined in-distribution (ID) labels. Nevertheless, when deployed in an open-world environment, they acknowledge the possibility of encountering objects from unknown categories, which should not be naively assigned to any ID labels. It poses a risk to the security of object detection models. In high-stakes applications like autonomous driving, failure to detect OOD objects can lead to severe accidents (Nitsch et al. 2021). We can mitigate this risk if unknown objects are detected and the human driver is alerted to take control.



a) Over-confident Predictions b) OOD Calibrations

Figure 1: Object detectors in the open world tend to make erroneous decisions when facing unknown objects, threatening machine learning system security. To mitigate this, we adapt knowledge-rich vision-language representations into the ID concept space for object-level OOD detection.

The susceptibility to OOD inputs stems from insufficient knowledge about unknowns during training. It results in neural networks tending to abnormally generate overconfident predictions when semantic shifts occur within the samples (Hein, Andriushchenko, and Bitterwolf 2019), as shown in Figure 1. In object-level OOD detection, a line of prior works (Du et al. 2022a,c,b; Wu and Deng 2023; Wu, Deng, and Liu 2024) design additional modules to be integrated into the training process of detectors, leveraging the model's inherent uncertainty. Meanwhile, some estimation-based methods directly learning from training data (Lee et al. 2018b; Tack et al. 2020a; Sun et al. 2022) are also introduced into this domain as alternative solutions.

In recent years, advancements in contrastive multimodal pre-training methods, including ALIGN (Jia et al. 2021), CLIP (Radford et al. 2021), BLIP (Li et al. 2022) and InternVL (Chen et al. 2024b), have provided a novel perspective for detecting out-of-distributions. With extensive prior knowledge, vision-language (VL) representations can transferably detect distributional shifts in downstream image-level classification tasks given the ID textual class labels (Esmaeilpour et al. 2022; Ming et al. 2022). This observation prompts us to delve further: if this pre-trained alignment capability can be adapted to measure the regional uncertainty for individual objects, we would effortlessly unlock

[*] These authors contributed equally.
[†] Corresponding authors

a safety assistant for deployed detectors, replacing previous limited enhancement methods and potentially boosting the performance of object-level OOD detection. However, applying these pre-trained models to the object-level OOD detection task presents substantial challenges. Unlike image-level classification, where the entire image is considered, object detection focuses on localized regions. This localization process can lead to a loss of contextual information, making it difficult to accurately assess an object's anomaly.

Moreover, the effectiveness of pre-trained models is often affected by the variety and quality of the dataset of the object detector. Datasets like BDD-100K(Yu et al. 2020), which contain a wide range of object sizes, lighting conditions, and occlusions, can pose challenges for models trained on more generic datasets. For instance, an object detector trained on BDD-100K might misclassify small, occluded objects from other datasets as ID vehicles. This situation underscores the necessity for domain-specific adaptation to enhance the performance of these models in applied settings.

This study proposes a novel framework, RUNA, to address the abovementioned limitations for object-level OOD detection. RUNA leverages a dual encoder architecture to provide rich contextual information and employs a **R**egional **UN**certainty **A**lignment strategy to effectively calculate uncertainty scores for object regions, enabling accurate classification as ID or OOD. We employ few-shot fine-tuning to bridge the performance gap between generic and domain-specific datasets. The diverse and challenging nature of datasets like BDD-100K, characterized by varying lighting conditions, occlusions, and object appearances, necessitates tailored model adaptation. Moreover, the pre-trained VL models, primarily trained on scene-centric images, exhibit limited alignment capability with the unique characteristics of domain-specific datasets. Fine-tuning allows the model to acquire domain-specific characteristics, enhancing its capability to detect OOD objects efficiently.

Our contributions are as follows:

- We propose RUNA, a novel object-level OOD detection framework with a dual encoder architecture that captures global and local features for accurate regional uncertainty estimation.
- We develop a few-shot fine-tuning approach to efficiently align region-level ID semantics, substantially enhancing the model's capacity to differentiate between ID and OOD objects.
- Our approach remarkably improves object-level OOD detection performance compared to previous methods.

Preliminaries

Object-level OOD Detection. Object-level OOD detection aims to identify unknown objects that fall outside the recognized categories and may be misidentified by the model. Object-level OOD detection is more applicable to real-world machine learning systems than image-level OOD detection. However, it also presents more significant challenges, as it requires careful consideration of each object's uncertainties at a fine-grained level.

We define the ID space as $\mathcal{X}$, with the associated label space given by $\mathcal{Y}_{\text{in}} = \{y_1, y_2, y_3, ...y_K\}$. In object-level OOD detection literature(Du et al. 2022a,b,c; Wu and Deng 2023; Wu, Deng, and Liu 2024), regional objects $\hat{x}_b$ from an OOD sample x_{out} are considered to experience a semantic shift compared to ID objects, meaning their label space $\mathcal{Y}_{\text{out}}$ does not overlap with $\mathcal{Y}_{\text{in}}$.

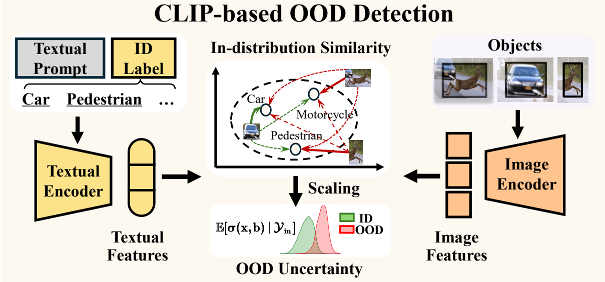


Figure 2: Framework of CLIP-based OOD Detection. Green arrows represent ID samples, while red arrows denote OOD samples. The solid line highlights the maximum similarity, and the dotted lines indicate other similarity measures.

For an unknown image x, the object detector f_θ predicts results as $D_x = \{b_i, y_i^p\}_{i=1}^m$. Here, $b_i \in \mathbb{R}^4$ represents the bounding box, and $y_i^p \in \mathcal{Y}_{\text{in}}$ is its ID semantic label. OOD detection is structured as a binary classification task with uncertainty estimation $\sigma(\cdot)$, distinguishing between ID and OOD objects. Given a bounding box b, the goal is to predict the uncertainty $\sigma(x, b)$:

$$G(\hat{x}_b, \mathcal{Y}_{\text{in}}) = \begin{cases} \text{in}, & \text{if} \quad \mathbb{E}[\sigma(x,b) \mid \mathcal{Y}_{\text{in}}] \leq \gamma \\ \text{out}, & \text{if} \quad \mathbb{E}[\sigma(x,b) \mid \mathcal{Y}_{\text{in}}] > \gamma \end{cases} \tag{1}$$

where γ is the threshold chosen such that a high fraction of ID data (e.g., 95%) falls below it.

Zero-shot CLIP-based OOD Detection. CLIP has excelled in zero-shot OOD detection tasks by utilizing extensive training data and large-scale models.

We outline the approach for performing zero-shot OOD detection utilizing existing CLIP-based zero-shot methodologies Maximum Concept Matching (MCM)(Ming et al. 2022). As depicted in Figure 2, the CLIP model employs an image encoder to extract features from images. While CLIP lacks an explicit classifier, we can use ID classes to create text inputs (e.g., "a photo of a dog"). The text encoder processes these text inputs to produce class-specific features, which act similarly to a classifier. Let $\mathcal{T}_{\text{in}}$ represent the collection of test prompts that include K class labels. To illustrate, Ming et al. calculates the score with the softmax score of the similarity between image features and textual features:

$$S_{\text{MCM}}(x) = \max_{t_i \in \mathcal{T}_{\text{in}}} \frac{e^{\text{Sim}(\mathcal{I}(x), \mathcal{T}(t_i))/\tau}}{\sum_{t_c \in \mathcal{T}_{\text{in}}} e^{\text{Sim}(\mathcal{I}(x), \mathcal{T}(t_c))/\tau}} \tag{2}$$

where x denotes the input image, t_i and t_c denotes the text to match. $\text{Sim}(\mathcal{I}(x), \mathcal{T}(t_i))$ denotes the similarity between image feature $\mathcal{I}(x)$ and textual feature $\mathcal{T}(t_i)$. If the MCM for a given image is below a predefined threshold, it is classified as ID; otherwise, it is considered OOD.

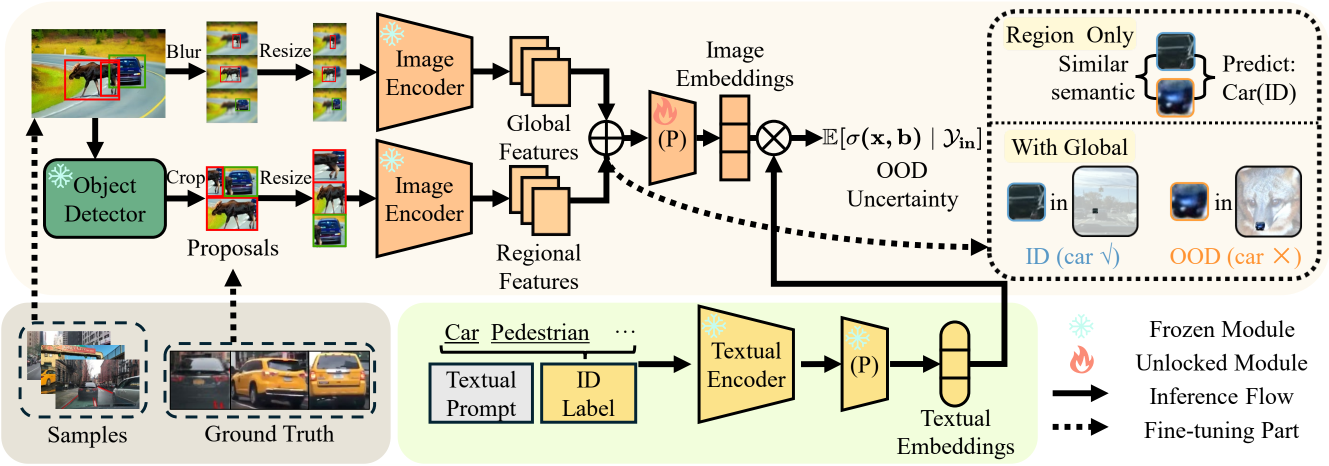


Figure 3: **Overview of RUNA Framework.** Our novel dual-encoder architecture computes regional object uncertainty by extracting global and regional image features and aligning them with text features. During fine-tuning, the image encoder handling regional images remain frozen, while only its projection layer (P) participate in the fine-tuning. The upper right dashed box highlights the importance of our feature fusion strategy: when encountering objects with similar semantics, limited local features can lead to incorrect decisions. By incorporating global features, the model can make more informed judgments.

Method

Overview. As in Figure 3, our framework, RUNA, utilizes a dual Image Encoder structure $\mathcal{I}^{(g)}, \mathcal{I}^{(r)}$ to collaboratively process visual information, where $\mathcal{I}^{(g)}$ captures global features from the entire image, and $\mathcal{I}^{(r)}$ focuses on regional features by processing specific objects or areas of interest. The outputs are subsequently fused to produce the final image embeddings. We employ a metric based on the maximum similarity score to quantify uncertainty in object detection. Notably, we exclusively fine-tune the projection layer of the visual encoder, denoted as $\mathcal{I}_P$, to tailor the pre-trained model for our particular task.

Given an image x and its corresponding region of interest (bounding box) $\hat{x}_b$, our framework first extracts global features from the entire image using $\mathcal{I}^{(g)}(x)$. Here, bounding box blurring is applied to keep contextual information while emphasizing the objects of interest, enhancing the extracted features' relevance. Concurrently, $\mathcal{I}^{(r)}(\hat{x}_b)$ processes the region of interest by cropping the specific area, enabling a more focused and detailed representation of the object.

To effectively integrate the semantic information extracted from the regional image with the global context, we propose a novel fusion strategy, which can be expressed as:

$$\mathcal{I}_t(x, \hat{x}_b) = \mathcal{I}_P(\lambda \cdot \mathcal{I}^{(r)}(\hat{x}_b) \oplus (1-\lambda) \cdot \mathcal{I}^{(g)}(x)) \tag{3}$$

where $\oplus$ denotes element-wise addition, and λ is a weighting coefficient that regulates the influence of each encoder. The resulting $\mathcal{I}_t(x, \hat{x}_b)$ represents the final image embedding, which synthesizes both regional and global features.

As illustrated in the top-right dashed box on the left in Figure 3, when encountering objects with similar semantics, relying solely on limited local features may cause the model to misinterpret subtle differences, leading to incorrect decisions. This is especially problematic when visually similar objects share overlapping attributes, making it challenging to distinguish between them using only localized cues. By incorporating global features, the model gains a broader context, enabling it better to grasp the overall scene structure and relational information. This holistic view allows for more informed and accurate judgments, as the model can integrate detailed local patterns and the larger contextual backdrop, leading to more robust OOD detection.

As illustrated in Figure 3, our approach functions as a post-hoc corrective technique that does not interfere with the training process of the object detection model. For a regional object $\hat{x}_b$ identified by the detector's prediction x, b, we transform the distributional uncertainty $\sigma(x, b)$ of $\hat{x}_b$ into a distance measure relative to the ID semantic space. Given the ID space as $\mathcal{P} \in \mathbb{R}^K$, the uncertainty of a predicted region $\hat{x}_b$ is represented as its deviation from this known distribution $\mathcal{P}$:

$$\mathbb{E}[\sigma(x, b) \mid \mathcal{Y}_{\text{in}}] = \mathbb{E}_{\mathcal{P}}[\mathcal{H}(\hat{x}_b, \mathcal{P})] \tag{4}$$

where $\mathcal{H}$ denotes the selected distance measurement function. In this framework, the uncertainty of an unknown object is converted into a distance-based estimation relative to the ID space $\mathcal{P}$. In line with prior research(Ming et al. 2022), we construct the K dimensions of $\mathcal{P}$ using the pre-trained CLIP text encoder $\mathcal{T}(\cdot)$.

Regional Uncertainty Alignment. Our goal is to quantify the discrepancy between the input region image $\hat{x}_b$ and the entire ID semantic space $\mathcal{P}$. This discrepancy is accomplished by assessing the similarity between the image features and predefined concept vectors corresponding to ID labels, with this similarity serving as a distance measure.

Initially, we consider that the semantic similarities between $\mathcal{I}_t(x, \hat{x}_b)$ and all ID concept vectors contribute to assessing its distance from the ID space $\mathcal{P}$ which means $\mathcal{T}(t_i)$. A straightforward method is to sum these similarities (Direct

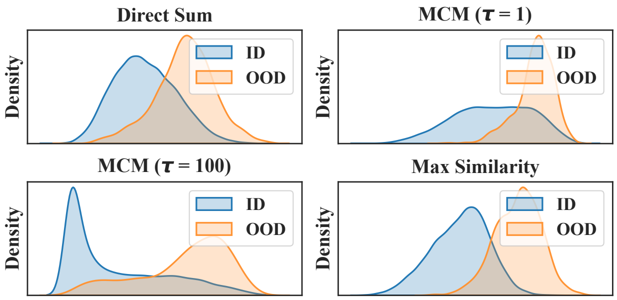


Figure 4: Distribution of uncertainty scores for Direct Sum, MCM(τ = 1), MCM(τ = 100) and Max Similarity.

Sum) to reflect the degree of deviation of $\{x, \hat{x}_b\}$:

$$\mathbb{E}[\sigma(x,b) \mid \mathcal{Y}_{\text{in}}] = -\sum_{i=1}^{K} \text{Sim}(\mathcal{I}_{\text{t}}(x,\hat{x}_b), \mathcal{T}(t_i)) \quad (5)$$

The negative sign indicates lower semantic similarity with ID concept vectors for OOD objects corresponds to a greater distance from $\mathcal{P}$, implying higher uncertainty.

However, during our evaluation, we observed that Direct Sum fails to effectively distinguish between ID and OOD objects in the VOC dataset. We attribute this issue to the limited variance in the cosine similarities outputted by CLIP, which do not exhibit significant differences between matching and non-matching situations. This results in the absolute values of the summed similarities, overshadowing the impact of actual differences.

To tackle this issue, we adapt the MCM in Eq.(2) by substituting $\mathcal{I}(x)$ with $\mathcal{I}_{\text{t}}(x,\hat{x}_b)$, thereby amplifying the differences between the similarities. However, MCM does not significantly improve the differentiation between ID and OOD objects, as shown in Figure 4. Interestingly, we notice that as the scaling factor of MCM increases, its performance improves. This phenomenon prompts us to investigate further. We find that as the differences between values expand, the influence of larger values on the scores also increases. When the factor approaches its limit, the score is predominantly influenced by the maximum similarity value. Consequently, we define the uncertainty estimation metric as follows:

$$\mathbb{E}[\sigma(x,b) \mid \mathcal{Y}_{\text{in}}] = -\max_{1\leq i\leq K} \text{Sim}(\mathcal{I}_{\text{t}}(x,\hat{x}_b), \mathcal{T}(t_i)) \quad (6)$$

where $\text{Sim}(\mathcal{I}_{\text{t}}(x,\hat{x}_b), \mathcal{T}(t_i))$ denotes similarity between the image feature $\mathcal{I}_{\text{t}}(x,\hat{x}_b)$ and the concept vector $\mathcal{T}(t_i)$ for each label i and K denotes the number of labels.

Few-shot Fine-tuning. Although the zero-shot method lays a solid groundwork for OOD detection, it lacks the nuanced understanding of ID data needed to distinguish between ID and OOD samples precisely. We introduce a fine-tuning strategy that utilizes few-shot learning for cost-effective and rapid adaptation to fill this gap. By randomly selecting a small set of images, our approach infuses the model with region-specific details, unlocking a more profound comprehension of ID characteristics.

For a given image x, we treat all ground truth bounding boxes $\{\hat{x}_b^i\}_{i=1}^m$ as potential fine-tuning candidates. From these, N shots of K kinds of objects are randomly drawn, denoted as $\{(\hat{x}_b^i, y_i)\}_{i=1}^{NK}$, where each $\hat{x}_b^i$ corresponds to a regional patch and y_i indicates its associated label. This process selectively exposes the model to key ID regions, enhancing its ability to align with the fine-grained semantic features vital for robust ID/OOD discrimination.

Given the pre-trained image encoder's capacity for intense feature extraction, our fine-tuning selectively focuses on refining the projection layer after the image encoder handles regional images. We aim to align visual embeddings with their corresponding label embeddings closely. To this end, we employ a contrastive loss that sharpens the model's intra-ID discriminative power:

$$\mathcal{L}_{\text{ID}} = -\sum_{\hat{x}_b\in\mathcal{B}} \log \frac{\exp(\text{Sim}(\mathcal{I}(\hat{x}_b), \mathcal{T}(t_i))/\tau)}{\sum_{j=1}^{K}\exp(\text{Sim}(\mathcal{I}(\hat{x}_b), \mathcal{T}(t_j))/\tau)} \quad (7)$$

where τ is a temperature to scale cosine similarities. This contrastive loss formulation hones the model's precision within the ID space and optimizes its ability to differentiate between subtle category variations, driving a more context-aware and semantically aligned fine-tuning process.

Experiments

Experimental Settings

Datasets and metrics. We use PASCAL-VOC(Everingham et al. 2010) and BDD-100K(Yu et al. 2020) as ID datasets and evaluate on two OOD datasets sourced from MS-COCO(Lin et al. 2014) and OpenImages(Kuznetsova et al. 2020), ensuring no label overlap with ID datasets. The object detection model is pre-trained on the ID datasets. We evaluate using three metrics: 1) **FPR95**: False Positive Rate at 95% True Positive Rate for ID samples, indicating the proportion of misclassified OOD objects; 2) **AUROC**: Area Under the ROC curve, where higher values indicate superior performance; 3) **mAP**: We do not report mAP as the object detection model is not affected by the integreted RUNA OOD detector.

Models and Baselines. We utilize the Detectron2 platform (Wu et al. 2019) with Faster R-CNN(Ren et al. 2015) (using ResNet50(He et al. 2016) as the backbone) as the frozen object detection model. We adopt CLIP (VIT-B/16) (Radford et al. 2021) for the vision-language model. Our CLIP-based approaches are evaluated against several widely adopted image-level approaches, including MSP (Hendrycks and Gimpel 2017), ODIN (Liang, Li, and Srikant 2018), Mahalanobis (Lee et al. 2018b), Generalized ODIN (Hsu et al. 2020), CSI (Tack et al. 2020b), Gram matrices (Sastry and Oore 2020), Energy score (Liu et al. 2020), and CLIP-based MCM(Ming et al. 2022). Additionally, we compare with object-level OOD detection methods such as VOS (Du et al. 2022c), SIREN (Du et al. 2022a), SAFE (Wilson et al. 2023), TIB (Wu and Deng 2023), and PCA-based method (Wu, Deng, and Liu 2024).

Implementation details. We exclusively fine-tune the projection layer of the regional image encoder individually, while the global image encoder is off-the-shelf. For few-shot learning, we perform fine-tuning using 10-shot samples. For ID discriminative fine-tuning, we employ a batch size of 256

In-distribution datasets	Detection Method	OpenImages		MSCOCO	
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
Berkeley DeepDrive-100k	MSP (Hendrycks and Gimpel 2017)	79.04	77.38	80.94	75.87
	ODIN (Liang, Li, and Srikant 2018)	58.92	76.61	62.85	74.40
	Mahalanobis (Lee et al. 2018b)	60.16	86.88	57.66	84.92
	Energy score (Liu et al. 2020)	54.97	79.60	60.06	77.48
	Gram matrices (Sastry and Oore 2020)	77.55	59.38	60.93	74.93
	Generalized ODIN (Hsu et al. 2020)	50.17	87.18	57.27	85.22
	CSI (Tack et al. 2020b)	37.06	87.99	47.10	84.09
	SIREN (Du et al. 2022a)	37.19	87.87	39.54	88.37
	VOS (Du et al. 2022c)	35.61	88.46	44.13	86.92
	MCM* (Ming et al. 2022)	45.37	88.46	53.79	86.92
	SAFE (Wilson et al. 2023)	16.04	94.64	32.56	88.96
	TIB (Wu and Deng 2023)	24.00	92.54	36.85	88.47
	PCA-based (Wu, Deng, and Liu 2024)	35.05	88.92	45.72	85.14
	RUNA (Ours)	**9.95**	**96.76**	**16.85**	**93.92**
PASCAL-VOC	MSP (Hendrycks and Gimpel 2017)	73.13	81.91	70.99	83.45
	ODIN (Liang, Li, and Srikant 2018)	63.14	82.59	59.82	82.20
	Mahalanobis (Lee et al. 2018b)	96.27	57.42	96.46	59.25
	Energy score (Liu et al. 2020)	58.69	82.98	56.89	83.69
	Gram matrices (Sastry and Oore 2020)	67.42	77.62	62.75	79.88
	Generalized ODIN (Hsu et al. 2020)	70.28	79.23	59.57	83.12
	CSI (Tack et al. 2020b)	57.41	82.95	59.91	81.83
	SIREN (Du et al. 2022a)	49.12	87.21	54.23	86.89
	VOS (Du et al. 2022c)	50.79	85.42	47.29	88.35
	MCM* (Ming et al. 2022)	48.73	80.16	50.43	78.22
	SAFE (Wilson et al. 2023)	**20.36**	92.28	47.40	80.30
	TIB (Wu and Deng 2023)	47.19	88.09	41.55	90.36
	PCA-based (Wu, Deng, and Liu 2024)	50.56	85.71	44.54	89.40
	RUNA (Ours)	26.07	**93.63**	**30.67**	**92.48**

Table 1: **Main results.** ↑ denotes that higher values are considered superior, while ↓ signifies that lower values are desirable. All results are presented as percentages. **Bold** numbers represent superior results, and the second-best performance is marked with an underline. $*$ means adapted with our dual-encoder architecture.

and use the AdamW optimizer, conducting fine-tuning over 100 epochs with a base learning rate of 5×10^{-6}. We use ”a photo of a {label}” for the textual prompt. We set the dual encoder's fusion coefficient λ to 0.5 and the blur radius R of Gaussian Blur to 1.

Main Results

As illustrated in Table 1, our proposed CLIP-based approach, RUNA, demonstrates advantages over previous methods. Notably, on the autonomous driving dataset BDD-100K, our fine-tuning approach significantly enhances the detection of OOD objects. In tests on the OOD dataset OpenImages, RUNA achieves an FPR95 of **9.95%**, marking a **6.09%** reduction compared to the previously best-performing method, SAFE. On the OOD dataset MSCOCO, our method achieved an FPR95 of **16.85%**, improving by **15.71%** compared to SAFE. When VOC serves as the ID dataset and OpenImages as the OOD dataset, SAFE achieves a lower FPR95 compared to our method. However, on the OOD dataset MSCOCO, our method outperforms SAFE with an FPR95 of **30.67%**, achieving an improvement of **16.73%** over SAFE and **10.88%** over SOTA method TIB. Our approach provides detector-agnostic performance without requiring manual feature selection, making it more flexible and broadly applicable across various detection scenarios. In contrast, previous methods explicitly designed for object-level OOD detection require retraining the target detection model, which may affect its original detection performance.

Furthermore, our fine-tuning framework showcases a remarkable improvement over the zero-shot approach. For instance, on VOC, RUNA shows a substantial FPR95 improvement of **12.90%** on OpenImages and **20.70%** on MSCOCO. Specifically, on BDD-100K, RUNA shows a substantial FPR95 improvement of **19.47%** on OpenImages and **20.11%** on MSCOCO.

Ablation Study

Ablation on the dual encoder and ID fine-tuning. We examine the impact of the dual encoder and ID fine-tuning components within our object-level OOD detection framework. The fine-tuning strategy is designed to improve the

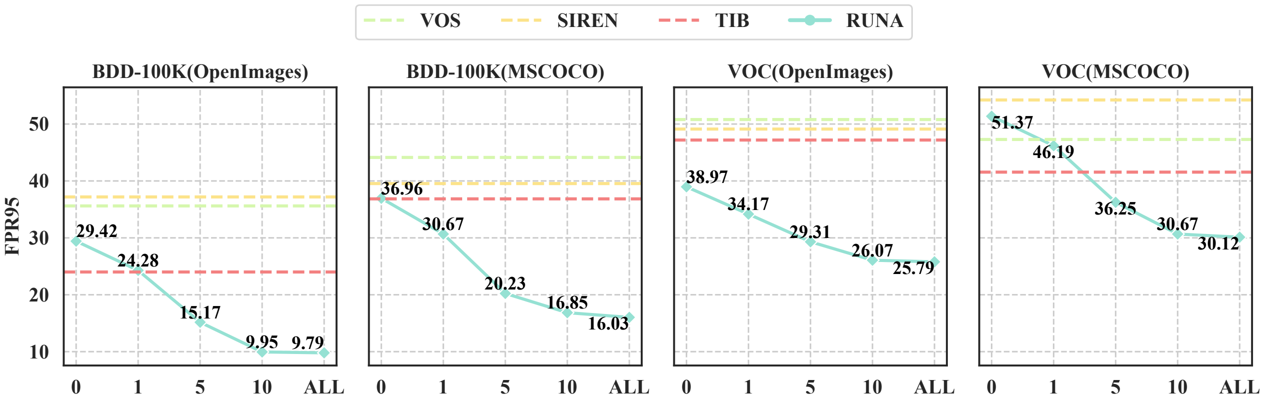


Figure 5: Ablation study on the number of fine-tuning samples (shots). This study examines how varying the number of shots affects detection performance, showing the trade-offs between data efficiency and detection quality.

ID	Method			OpenImages/MSCOCO	
	RE	GE	FT	FPR95 ↓	AUROC ↑
BDD-100k	✓	-	-	58.09/62.17	70.94/69.10
	-	✓	-	31.78/39.24	89.21/84.19
	✓	✓	-	29.42/36.96	89.90/85.37
	✓	-	✓	15.71/22.14	93.34/92.26
	✓	✓	✓	**9.95/16.85**	**96.76/93.92**
VOC	✓	-	-	42.75/55.26	91.19/84.65
	-	✓	-	39.22/52.49	91.45/87.49
	✓	✓	-	38.97/51.37	91.78/88.92
	✓	-	✓	30.76/34.31	92.01/90.17
	✓	✓	✓	**26.07/30.67**	**93.63/92.48**

Table 2: Ablation study on our regional encoder (RE), global encoder (GE) and few-shot fine-tune (FT).

discriminative capability of the pre-trained model with respect to ID semantics. At the same time, the dual encoder focuses on strengthening sensitivity to ID objects via enhanced feature representation. The evaluation results are presented in Table 2. We compare the performance of models with and without the dual encoder to assess its contribution to feature extraction. Additionally, we assess the impact of fine-tuning on ID data, analyzing its role in adapting the model to the unique features of the ID space and improving OOD detection accuracy. The results demonstrate that the dual encoder significantly enhances the model's capability to distinguish between ID and OOD samples. ID fine-tuning further refines this capability, leading to more robust OOD detection.

Ablation on the number of fine-tuning samples (shots). We examine the effect of changing the number of samples used for model fine-tuning. The evaluation results are presented in Figure 5. The zero-shot method depends entirely on the pre-trained model, providing a competitive baseline performance but struggling with object-level OOD detection due to the lack of ID supervision. Introducing 1-shot learning shows immediate improvements, leveraging a single example to better align the model with the target task. With 5-shot learning, the model demonstrates significant gains, as many examples facilitate a more comprehensive understanding of the data distribution. Finally, 10-shot learning further enhances performance, capturing even more nuances and variations within the data. This study illustrates the clear benefits of incorporating a few labeled examples, with each incremental increase in sample size resulting in notable improvements in the model's accuracy and robustness. However, we also perform fine-tuning on the entire dataset, which yields only minimal improvements while incurring significantly more computational costs. This indicates that while few-shot learning provides substantial benefits, full dataset fine-tuning offers diminishing returns compared to the increased computational demands.

Ablation on different backbones of the visual encoder. We assess the influence of various ViT backbones on the performance of the CLIP model, as illustrated in Table 3. The analysis reveals that while the more extensive backbones, ViT-L/14 and ViT-L/14-336, provide slight improvements in FPR95 and AUROC, they substantially increase runtime and parameter count. Specifically, the ViT-L/14 and ViT-L/14-336 backbones, despite their slightly better performance, significantly increase computational demands. On the other hand, ViT-B/16 offers a good balance between efficiency and performance, demonstrating that the additional computational cost of larger models does not proportionally enhance performance. Consequently, we choose ViT-B/16 as the backbone for its optimal trade-off between performance gains and resource efficiency. This choice ensures that our framework remains computationally feasible while delivering high accuracy in OOD detection.

Related Work

OOD Detection for Classification

OOD detection distinguishes the unknown inputs that deviate from ID training data during the testing phase. The employment of the maximum softmax probability (MSP) (Hendrycks and Gimpel 2017) serves as a common base-

Backbone	BDD-100k		VOC		Run time(ms)			Params
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	1-image	5-image	10-image	
VIT-B/32	59.47/64.11	68.02/70.13	43.59/54.37	89.19/80.23	62.9	90.7	129.1	0.15B
VIT-B/16	58.09/62.17	70.94/69.10	42.75/55.26	91.19/84.65	83.8	220.1	410.4	0.15B
VIT-L/14	57.24/61.01	71.89/71.24	42.38/53.49	91.56/86.67	309.1	863.7	1559.5	0.43B
VIT-L/14(336)	56.99/60.12	72.80/73.01	42.01/53.35	91.98/87.13	533.3	2086.5	4673.9	0.43B

Table 3: Ablation on the effect of different CLIP configurations on performance and run time. We evaluate the effects upon zero-shot regional encoder only method. The results are presented for two OOD datasets, with OpenImages followed by MSCOCO.

line approach; however, it can yield excessively high values for OOD inputs (Hein, Andriushchenko, and Bitterwolf 2019). Various enhancements have been suggested, such as ODIN (Liang, Li, and Srikant 2018), Mahalanobis (Lee et al. 2018b), Energy score (Liu et al. 2020), Gram matrices (Sastry and Oore 2020), CSI (Tack et al. 2020b), GODIN (Hsu et al. 2020), etc. While traditional OOD detection methods (Dhamija, Günther, and Boult 2018; Lee et al. 2018a; Hendrycks, Mazeika, and Dietterich 2018; Li and Vasconcelos 2020; Chen et al. 2024a; Kingma and Dhariwal 2018; Schirrmeister et al. 2020) largely stemmed from image-level classification tasks, some unique challenges posed by object detection require specialized approaches.

Object-level OOD Detection

OOD detection has garnered increasing attention in object detection to ensure the robustness of visual systems. Mainstream approaches (Du et al. 2022a,c) primarily focus on model regularization to achieve optimal representations. In contrast, the SAFE method enhances OOD detection by selecting sensitivity-aware features from the object detector and inputting them into an auxiliary MLP network (Wilson et al. 2023). Recent advancements include the Two-Stream Information Bottleneck method (Wu and Deng 2023), which utilizes dual information streams to identify unfamiliar objects without relying on auxiliary data, and the PCA-Driven Dynamic Prototype Enhancement technique, which dynamically refines prototypes for improved OOD discrimination using Principal Component Analysis (Wu, Deng, and Liu 2024). While many conventional methods depend on model uncertainty for OOD detection, we leverage a pre-trained vision-language model. The enhanced alignment knowledge from this model enables us to overcome the cognitive limitations of the original detection framework, resulting in improved OOD detection performance.

OOD Detection with Vision-language Models

Recent vision-language models, such as CLIP (Radford et al. 2021), have significantly advanced computer vision by aligning images and text in a shared feature space using a self-supervised contrastive objective. In OOD detection, Esmaeilpour et al. proposed ZOC, integrating a transformer-based decoder with CLIP's image encoder (Radford et al. 2021), tackling the challenge of sourcing OOD candidate labels—an issue our method bypasses. Building on this, Ming et al. introduced a CLIP-based OOD detection approach using MCM, while (Miyai et al. 2023) explored zero-shot ID detection to determine whether all objects in an image are ID. Unlike image-level studies, our work harnesses vision-language models for object-level OOD detection. The core challenge is to localize and identify region-level out-of-distribution objects within images. This task becomes particularly complex when most pre-trained models are designed for image-level representations. In contrast to previous studies that heavily rely on textual prompting, we concentrate on visual prompting to effectively guide CLIP's attention to the target of interest. By integrating few-shot learning, we greatly enhance the efficiency of the fine-tuning process. This approach allows us to achieve optimal performance with a minimal amount of data, thereby improving the model's flexibility and robustness across various contexts and reassuring the reader of the reliability of our approach.

Conclusion

This paper focuses on detecting OOD objects using pre-trained vision-language representations. Our investigation begins with evaluating the effectiveness of CLIP-like representations in identifying object-level OOD instances and proposes the zero-shot baseline. Additionally, we propose RUNA, a novel Regional UNcertainty Alignment strategy, which significantly enhances detection performance by adapting vision-language models to ID semantic space and guiding vision-language models to be more sensitive to ID concepts. Overall, experimental results demonstrated substantial improvements over previous methods, emphasizing the effectiveness and promise of our proposed approaches. In the future, we intend to investigate more refined visual prompting techniques to enhance the model's capacity to capture subtle details and variations within the data. Furthermore, we aim to explore the deployment of our model to edge services, enabling real-time object detection and OOD detection in resource-constrained environments.

Acknowledgement

This research was funded by the Key Research and Development Program of Guangdong Province (grant No. 2021B0101400003) and the National Key Research and Development Program of China under (grant No.2023YFB4502701). This work was done while Bin Zhang was interning at Ping An Technology.

References

Chen, J.; Li, J.; Qu, X.; Wang, J.; Wan, J.; and Xiao, J. 2024a. GAIA: Delving into Gradient-based Attribution Abnormality for Out-of-distribution Detection. *Advances in Neural Information Processing Systems*, 36.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

Dhamija, A. R.; Günther, M.; and Boult, T. 2018. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31.

Du, X.; Gozum, G.; Ming, Y.; and Li, Y. 2022a. Siren: Shaping representations for detecting out-of-distribution objects. *Advances in Neural Information Processing Systems*, 35: 20434–20449.

Du, X.; Wang, X.; Gozum, G.; and Li, Y. 2022b. Unknown-Aware Object Detection: Learning What You Don't Know from Videos in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13678–13688.

Du, X.; Wang, Z.; Cai, M.; and Li, Y. 2022c. VOS: Learning What You Don't Know by Virtual Outlier Synthesis. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Esmaeilpour, S.; Liu, B.; Robertson, E.; and Shu, L. 2022. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 6568–6576.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hein, M.; Andriushchenko, M.; and Bitterwolf, J. 2019. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 41–50.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. 2018. Deep Anomaly Detection with Outlier Exposure. In *International Conference on Learning Representations*.

Hsu, Y.-C.; Shen, Y.; Jin, H.; and Kira, Z. 2020. 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*, 10951–10960.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Kingma, D. P.; and Dhariwal, P. 2018. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31.

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

Lee, K.; Lee, H.; Lee, K.; and Shin, J. 2018a. Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples. In *International Conference on Learning Representations*.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018b. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, Y.; and Vasconcelos, N. 2020. Background data resampling for outlier-aware classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13218–13227.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33.

Ming, Y.; Cai, Z.; Gu, J.; Sun, Y.; Li, W.; and Li, Y. 2022. Delving into out-of-distribution detection with vision-language representations. *Advances in Neural Information Processing Systems*, 35: 35087–35102.

Miyai, A.; Yu, Q.; Irie, G.; and Aizawa, K. 2023. Zero-Shot In-Distribution Detection in Multi-Object Settings Using Vision-Language Foundation Models. *arXiv preprint arXiv:2304.04521*.

Nitsch, J.; Itkina, M.; Senanayake, R.; Nieto, J.; Schmidt, M.; Siegwart, R.; Kochenderfer, M. J.; and Cadena, C. 2021. Out-of-distribution detection for automotive perception. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 2938–2943. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Sastry, C. S.; and Oore, S. 2020. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning*, 8491–8501. PMLR.

Schirrmeister, R.; Zhou, Y.; Ball, T.; and Zhang, D. 2020. Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features. *Advances in Neural Information Processing Systems*, 33: 21038–21049.

Sun, Y.; Ming, Y.; Zhu, X.; and Li, Y. 2022. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, 20827–20840. PMLR.

Tack, J.; Mo, S.; Jeong, J.; and Shin, J. 2020a. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33: 11839–11852.

Tack, J.; Mo, S.; Jeong, J.; and Shin, J. 2020b. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33: 11839–11852.

Wilson, S.; Fischer, T.; Dayoub, F.; Miller, D.; and Sünderhauf, N. 2023. SAFE: Sensitivity-aware features for out-of-distribution object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23565–23576.

Wu, A.; and Deng, C. 2023. TIB: Detecting Unknown Objects via Two-Stream Information Bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wu, A.; Deng, C.; and Liu, W. 2024. Unsupervised Out-of-Distribution Object Detection via PCA-Driven Dynamic Prototype Enhancement. *IEEE Transactions on Image Processing*.

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

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