ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888875

VisTa: Visual-contextual and Text-augmented Zero-shot Object-level OOD Detection

Bin Zhang*†, Xiaoyang Qu†‡, Guokuan Li*‡, Jiguang Wan* and Jianzong Wang†
*Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China
†Ping An Technology (Shenzhen) Co., Ltd, Shenzhen, China
{binz2398, quxiaoy}@gmail.com, {liguokuan, jgwan}@hust.edu.cn, jzwang@188.com

***Abstract*—As object detectors are increasingly deployed as black-box cloud services or pre-trained models with restricted access to the original training data, the challenge of zero-shot object-level out-of-distribution (OOD) detection arises. This task becomes crucial in ensuring the reliability of detectors in open-world settings. While existing methods have demonstrated success in image-level OOD detection using pre-trained vision-language models like CLIP, directly applying such models to object-level OOD detection presents challenges due to the loss of contextual information and reliance on image-level alignment. To tackle these challenges, we introduce a new method that leverages visual prompts and text-augmented in-distribution (ID) space construction to adapt CLIP for zero-shot object-level OOD detection. Our method preserves critical contextual information and improves the ability to differentiate between ID and OOD objects, achieving competitive performance across different benchmarks.**

***Index Terms*—Zero-shot object-level OOD detection, visual prompt, vision-language representations.**

I. Introduction

As object detection models are being more frequently used in practical applications [1]–[3], ensuring their robustness in open-world scenarios is crucial. In such environments, models frequently encounter inputs not belonging to the training distribution, referred to as out-of-distribution (OOD) samples. In contrast, in-distribution (ID) samples refer to those on which the model has been trained. When deep neural networks encounter OOD samples, they often fail silently, leading to overconfident erroneous predictions [4]–[9]. This failure to correctly identify OOD samples can have severe consequences, such as misclassification and overconfidence in predictions. In safety-critical tasks like autonomous driving, where undetected OOD objects can cause accidents, these consequences can be particularly risky [10]–[12]. Therefore, the emergence of object-level OOD detection, which focuses on identifying anomalous objects at a granular level, is a crucial and rapidly evolving research area.

In object-level OOD detection, prior works [13]–[15] often require integrating additional modules into the training process of detectors, leveraging the model's inherent uncertainty. These approaches typically necessitate supervised training and the inclusion of extra components, such as uncertainty estimation heads, to identify potential OOD objects. Additionally, the SAFE method [16] avoids retraining the object detector. Instead, SAFE manullay extracts features from the pre-trained object detector and trains a separate MLP to identify OOD objects, reducing the need to modify the original detection pipeline and offering a more efficient and flexible alternative. Estimation-based methods [17]–[20], on the other hand, focus on identifying outliers through techniques like distance metrics or contrastive learning, directly learning from the training data.

Research is supported by the Key Research and Development Program of Guangdong Province (grant No. 2021B0101400003).
‡Corresponding authors.

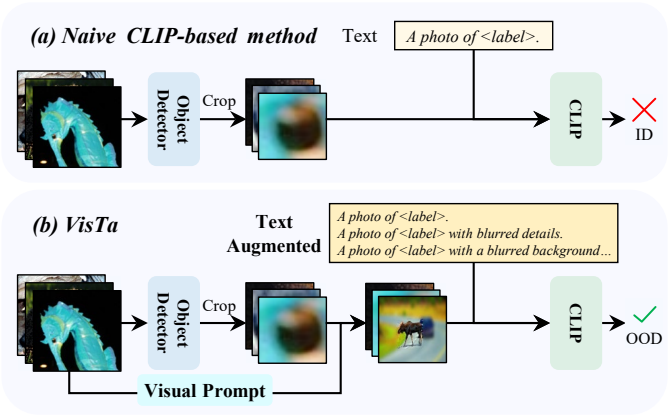


Fig. 1. Framework of (a) the naive CLIP-based method and (b) the proposed VisTa approach. In (a), zero-shot CLIP-based OOD detection is directly adapted for object-level OOD detection using cropping. In (b), we emphasize the two key components of our VisTa method.

Despite their effectiveness, these methods are constrained by their dependence on ID data and often require retraining or adding new training phases. Furthermore, many pre-trained models, like those offered by Hugging Face Model Hub, are trained on proprietary or large-scale datasets that are not publicly available. This limitation makes it difficult for practitioners to fully leverage these models with traditional OOD detection techniques, as they need access to the original training data.

To address these issues and make OOD detection more accessible, the zero-shot detection paradigm is a promising solution. Utilizing publicly available pre-trained models eliminates the need for ID-specific training data, providing a practical alternative for those with limited resources. Recently, the advent of pre-trained vision-language models (VLMs), such as CLIP [21], ALIGN [22], BLIP [23], and InternVL [24], has opened new avenues for OOD detection. Previous work [25]–[27] have focused on zero-shot OOD detection within image classification, mainly using CLIP as a general classifier. This study concentrates on zero-shot object-level OOD detection. As shown in Fig. 1(a), while the performance on image-level tasks is excellent, directly applying it in object-level tasks with the help of techniques like cropping has a poor performance due to significant loss of contextual information.

Given these challenges, as shown in Fig. 1(b), our proposed approach offers a novel zero-shot solution that leverages visual prompts to adapt CLIP for object-level OOD detection, eliminating the need for retraining or additional modules. Our prompting mechanism guides the model to emphasizes features at the object level while preserving essential contextual information lost during cropping,

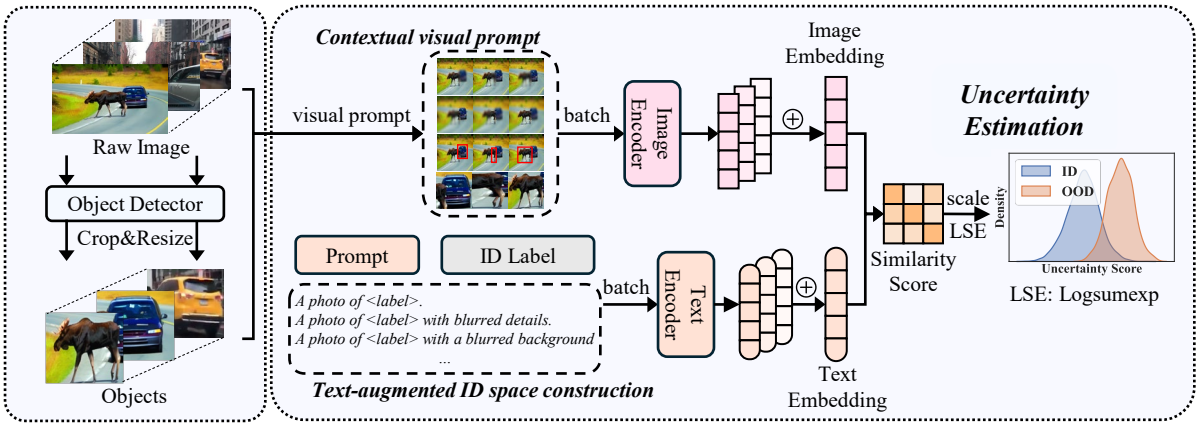


Fig. 2. **Overview of VisTa.** The ID embedding space $\mathcal{C}$ is built with CLIP's text encoder and augmented prompts. Image embeddings from contextual visual prompts are compared to $\mathcal{C}$ to compute similarity scores, which are used to calculate uncertainty and distinguish between ID and OOD samples.

allowing CLIP to transfer its pre-trained knowledge more effectively. Additionally, we introduce a text-augmented strategy to enhance the construction of the ID embedding space, aligning textual information with the visual prompts to expand the representative capacity of the ID space. This enriched embedding space enables more accurate differentiation between ID and OOD objects, significantly improving detection performance.

Our contributions are as follows:

- We propose a visual prompt-based method that effectively adapts CLIP for object-level OOD detection, preserving crucial contextual information and capturing features more accurately.
- We introduce a text-augmented approach that enhances the ID embedding space, aligning it with the visual prompts to improve the model's ability to differentiate between ID and OOD objects.
- We conduct extensive experiments demonstrating that our zero-shot approach achieves strong performance, surpassing existing methods on multiple object-level OOD detection benchmarks.

II. Method

A. Overview

This work introduces a novel approach to enhance object-level OOD detection by adapting the CLIP model, which is traditionally used for image-level tasks. The key challenge in applying CLIP to object-level tasks is the loss of contextual information when focusing on individual objects. We propose two main components to address this: context-aware visual prompts and text-augmented ID space construction. The visual prompts guide CLIP to preserve contextual information better, improving its ability to detect OOD samples at the object level. Additionally, the text-augmented approach enhances the construction of the ID embedding space, expanding its representational capacity to align with the applied visual prompts. Together, these components enable a robust, zero-shot approach to object-level OOD detection, leveraging the strengths of multimodal representations and providing a reliable solution to the challenge of OOD detection.

The pipeline of our proposed approach is shown in Fig. 2 and can be outlined as follows: (1) We first construct the ID embedding space $\mathcal{C}$ by encoding the K ID class labels $\mathcal{Y}_{\text{in}} = \{y_i,\ i = 1,\ 2,\ 3,\ ...,\ K\}$ using CLIP's text encoder, enhanced with augmented prompts to improve its expressiveness; (2) Next, consider an image x along with its corresponding bounding box b, we generate image embeddings using contextual visual prompts, ensuring that crucial contextual information is preserved; (3) Finally, these image embeddings are compared with the text-augmented ID embedding space to compute the similarity score, which is then used to calculate the final uncertainty score, distinguishing between ID and OOD samples:

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

where $\sigma(x,b)$ represents the extracted feature, $\mathbb{E}[\sigma(x,b) \mid \mathcal{C}]$ denotes the uncertainty score, and $G(x,b)$ is the predicted outcome for OOD detection. The threshold γ is determined based on the distribution of ID data, ensuring that the majority (e.g., 95%) of the ID data can be correctly distinguished from OOD data.

B. Contextual Visual Prompt

Our approach to object-level OOD detection begins by introducing carefully designed visual prompts. These prompts are layered on top of a fixed cropping operation that isolates the object of interest. The cropping step is consistently applied to all inputs, providing a uniform basis upon which additional visual prompts are added. The visual prompts are crucial for enhancing the model's ability to retain and leverage essential context, addressing the loss of such information when applying CLIP to object-level tasks. Each visual prompt emphasizes different aspects of the object and its surroundings, ultimately enriching the resulting visual features.

For example, the blur outside prompt focuses on the object by blurring the background, ensuring that the model concentrates on the object itself. The blur inside prompt blurs the details within the object while keeping the background clear, subtly highlighting the object's immediate environment. Additionally, drawing a distinctive color (e.g., red) around the bounding box creates a visual boundary that explicitly defines the object, enhancing the model's ability to differentiate between the object and its surrounding context.

Given an image x and its bounding box b, each visual prompt $\phi_V^p(\cdot)$ generates specific visual features via the CLIP image encoder $\mathcal{I}(\cdot)$, denoted as:

$$z_p = \mathcal{I}(\phi_V^p(x,b)) \tag{2}$$

This formulation captures various aspects of the object's context and appearance. Then, these features are combined through element-wise addition to form a comprehensive visual representation:

$$Z = \sum_{p=1}^{n} z_p / \Big\| \sum_{p=1}^{n} z_p \Big\| \tag{3}$$

TABLE I

COMPARISON WITH OTHER COMPETITIVE OOD DETECTION METHODS. THE COMPARISON METRICS ARE FPR95 AND AUROC, WHERE ↑ AND ↓ DENOTE PREFERRED DIRECTIONS. ALL RESULTS ARE PRESENTED IN PERCENTAGES, WITH **BOLD** NUMBERS INDICATING SUPERIOR RESULTS.

ID dataset	**BDD-100K**				**PASCAL-VOC**			
OOD dataset	**OpenImages**		**MSCOCO**		**OpenImages**		**MSCOCO**	
Detection method	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑
MSP [28]	79.04	77.38	80.94	75.87	73.13	81.91	70.99	83.45
ODIN [29]	58.92	76.61	62.85	74.40	63.14	82.59	59.82	82.20
Energy score [30]	54.97	79.60	60.06	77.48	58.69	82.98	56.89	83.69
Gram matrices [31]	77.55	59.38	60.93	74.93	67.42	77.62	62.75	79.88
Generalized ODIN [32]	50.17	87.18	57.27	85.22	70.28	79.23	59.57	83.12
Mahalanobis [17]	60.16	86.88	57.66	84.92	96.27	57.42	96.46	59.25
Vim [18]	53.80	86.49	54.58	87.17	88.40	68.73	83.47	71.94
KNN [19]	44.50	88.37	47.28	87.45	55.73	85.08	54.50	86.07
CSI [20]	37.06	87.99	47.10	84.09	57.41	82.95	59.91	81.83
MCM [25]	92.22	57.05	95.56	55.82	71.52	81.45	62.47	83.15
SIREN [13]	37.19	87.87	39.54	88.37	49.12	87.21	54.23	86.89
VOS [14]	35.61	88.46	44.13	86.92	50.79	85.42	47.29	88.35
TIB [15]	24.00	92.54	36.85	88.47	47.19	88.09	41.55	90.36
SAFE [16]	13.98	95.97	21.69	93.91	17.69	94.38	36.32	87.03
VisTa (Ours)	**8.68**	**97.76**	**15.27**	**93.92**	**9.49**	**97.91**	**25.74**	**94.47**

where n is the number of visual prompts and $\|\cdot\|$ denotes L2-norm. The integrated representation offers a holistic view of the object, encompassing fine-grained details and overall structure.

C. Text-Augmented ID Space Construction

Given the complexity and richness introduced by the visual prompts, a simple textual description like "*a photo of* {*label*}" is insufficient to represent the diversity of the ID space fully. To address this limitation, we propose a text-augmented approach to constructing the ID embedding space $\mathcal{C}$.

This approach enhances CLIP's text encoder by introducing text prompts aligned with the visual prompts applied earlier. For instance, to complement the blur outside visual prompt, we use a text prompt like "*A photo of* {*label*} *with a blurred background*" ensuring that the textual embedding accurately reflects the visual emphasis on the object within its context. Similarly, the blur inside prompt is paired with a text prompt such as "*A photo of* {*label*} *with blurred details*" to capture the nuanced focus on the object's environment.

By aligning these text prompts with the visual cues, we enrich the ID space $\mathcal{C}$ with a more expressive and contextually relevant set of embeddings. Specifically, $\mathcal{C}$ is constructed as:

$$t_i^p = \phi_T^p(y_i), \tag{4}$$

$$\mathcal{L}_i = \sum_{p=1}^{n} \mathcal{T}(t_i^p) / \Big\| \sum_{p=1}^{n} \mathcal{T}(t_i^p) \Big\|, \tag{5}$$

$$\mathcal{C} = \{\mathcal{L}_i, i \in \{1, 2, ..., K\}\} \tag{6}$$

where $\phi_T^p(\cdot)$ is the p-th prompt operation, t_i^p is the p-th text prompt of label y_i, $\mathcal{T}(\cdot)$ is the CLIP text encoder, n is the number of visual prompts, $\mathcal{L}_i$ is the augmented text embedding of label y_i and K is the number of ID classes. This enhanced ID space becomes crucial for accurately distinguishing between ID and OOD samples, allowing the model to understand better and interpret the contextual relationships inherent in the data.

D. Uncertainty Estimation

With the visual representation Z and the enriched ID space $\mathcal{C}$ constructed, we compute the similarity score between them: $S_i = \frac{Z \cdot \mathcal{L}_i}{\|Z\| \cdot \|\mathcal{L}_i\|}$. This similarity represents the degree of matching between the object and the ID label. Then, we compute the uncertainty score:

$$\mathbb{E}[\sigma(x, b) \mid \mathcal{C}] = -\log \sum_{i=1}^{K} e^{\tau \cdot S_i} \tag{7}$$

where τ is a temperature parameter. The final uncertainty score is then used to differentiate between ID and OOD samples as in (1).

TABLE II

IMPACT OF OUR CONTEXTUAL VISUAL PROMPT (VP) AND TEXT-AUGMENTED ID SPACE CONSTRUCTION (TA).

ID dataset	**Module**		**OpenImages/MSCOCO**	
	TA	**VP**	**FPR95** ↓	**AUROC** ↑
BDD-100k	-	-	64.12 / 69.16	72.01 / 71.83
	✓	-	60.12 / 63.18	75.29 / 74.37
	-	✓	23.42 / 30.96	91.90 / 90.37
	✓	✓	**8.68 / 15.27**	**97.76 / 93.92**
VOC	-	-	46.80 / 55.54	88.23 / 76.59
	✓	-	38.97 / 51.37	91.78 / 88.92
	-	✓	20.76 / 30.31	92.01 / 91.17
	✓	✓	**9.49 / 25.74**	**97.91 / 94.47**

III. EXPERIMENTS

A. Experimental setup

Datasets. We perform object detection tasks using the predefined ID/OOD splits outlined in [14]. The two ID datasets are derived from the widely used PASCAL-VOC [33] and BDD-100K [34] datasets. For the OOD datasets, we provide subsets of the MS-COCO [35] and OpenImages [36] datasets, ensuring that classes present in the custom ID datasets are excluded.

Evaluation metrics. Following [14], we adopt two key metrics: **FPR95** and **AUROC**. VisTa serves as a supplementary component to an already trained object detection network and does not influence the base model's performance regarding the mean average precision (mAP) metric; therefore, we do not include mAP in our reporting as done in [14].

Implementation details. We implement the Faster-RCNN detector with ResNet-50 using the Detectron2 library [37] and employ CLIP (ViT-B/16 [38]) as the VLM. For the visual prompts, we adopt crop,

TABLE III
IMPACT OF DIFFERENT CLIP BACKBONE. (CROP ONLY)

Backbone	RN50	RN101	VIT-B/32	VIT-B/16	VIT-L/14
FPR95	52.34	50.23	48.31	46.80	46.59
AUROC	84.97	85.79	87.19	88.23	88.37
Runtime(ms)	86.7	154.2	91.3	223.7	866.9

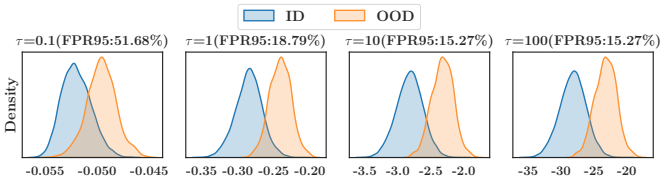


Fig. 3. Impact of different temperature parameter τ.

blur inside, blur outside, and box. We configured the Gaussian blur with a standard deviation of 2 and chose red for the box color.

B. Main Results

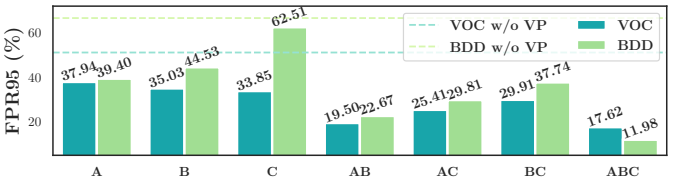
As shown in Table I, our proposed zero-shot CLIP-based approach, VisTa, demonstrates advantages over previous methods. Notably, on the autonomous driving dataset BDD-100K, VisTa greatly enhances the identification of OOD objects. In tests on the OOD dataset OpenImages, VisTa achieves an FPR95 of **8.68%**, a **5.30%** reduction compared to the previously best-performing method, SAFE. On the OOD dataset MSCOCO, VisTa achieves an FPR95 of **15.27%**, improving by **6.42%** compared to SAFE. When VOC serves as the ID dataset and OpenImages as the OOD dataset, VisTa performs the best, achieving an FPR95 of **9.49%**, which is an improvement of **8.20%** compared to SAFE. On the OOD dataset MSCOCO, VisTa achieves an FPR95 of **25.74%**, improving by **10.58%** compared to SAFE. These results highlight VisTa's robustness and strong OOD detection capabilities in both zero-shot and non-zero-shot scenarios.

C. Analysis

Impact of visual prompt and text-augmented ID space. Table II illustrates the impact of visual prompt and text-augmented ID space construction on OOD detection performance. The results show that while visual prompts alone significantly improve detection accuracy, adding the text-augmented ID space without corresponding visual prompts offers minimal performance gains. This minimal gain is because the textual enhancements are designed to complement the visual cues, and without the visual prompts, the model cannot fully leverage the added semantic layer. Visual prompts are essential for preserving contextual information, which is crucial for distinguishing objects at the local level. Combined with the text-augmented ID space, which adds a complementary semantic layer, the two components synergistically improve feature representation, directly supporting our goal of leveraging visual and textual cues for robust OOD detection.

Impact of different CLIP backbone. We evaluate ResNet50, ResNet101, ViT-B/32, ViT-B/16, and ViT-L/14 using VOC (ID) and OpenImages (OOD) datasets with batch size 8. ResNet50 is fastest but provides weaker features, while ResNet101 improves performance at higher cost. ViT-B/32 balances performance and efficiency, and ViT-B/16 further enhances performance with increased computation. ViT-L/14 achieves the best results but with marginal gains over ViT-B/16 and significantly higher costs. ViT-B/16 is optimal, balancing performance and efficiency, while ViT-L/14 highlights the trade-off between performance and cost.

(a) Illustration of different visual prompts.



(b) Qualitative results of different visual prompts.

Fig. 4. Impact of different visual prompts. We report average results. Text-augmented ID space is constructed with corresponding visual prompts.

Impact of temperature parameter τ. Using BDD [34] as ID and MSCOCO [35] as OOD datasets, we study the effect of τ (0.1, 1, 10, 100) on ID-OOD separation. Fig. 3 shows that moderately increasing τ improves separation, but beyond a threshold, further increases provide no benefit. This indicates τ is most effective within a moderate range, optimizing similarity computation without over-smoothing.

Impact of different visual prompts. In our ablation study, we evaluate three specific visual prompting techniques: (A) blur inside, (B) blur outside, and (C) box, analyzed alongside a fixed cropping operation to isolate the object of interest. Fig. 4(a) illustrates the visual effects of these prompts, along with (D) grayscale, (E) colorful box, and the standard (F) crop operation. Blur inside preserves the surrounding context, while blur outside emphasizes the object by softening its background. The colorful bounding box enhances visual salience. Although we explore grayscale and colorful boxes, their minimal impact on feature extraction and occasional performance degradation leads us to exclude their data from the main bar chart.

The bar chart in Fig. 4(b) quantifies the primary visual prompts' impact on OOD detection accuracy. Blur inside shows the most improvement, likely due to VLMs' pre-training on "Bokeh"-style datasets. Blur outside also performs well, while the colorful box shows modest gains, reflecting its limited influence compared to blur methods.

IV. CONCLUSION

We present an innovative zero-shot method for object-level OOD detection that combines visual prompts with text-augmented ID space construction. Our method enhances CLIP's ability to preserve crucial contextual information and enriches the ID embedding space with aligned textual cues. This integration enables strong performance across multiple benchmarks in a zero-shot setting. Experimental results show that our approach consistently surpasses prior methods. Furthermore, unlike existing methods that require retraining the detector or adding new networks, our approach works directly with pre-trained models, maintaining high ID accuracy while significantly improving OOD detection without additional training.

REFERENCES

[1] O. Wosner, G. Farjon, and A. Bar-Hillel, "Object detection in agricultural contexts: A multiple resolution benchmark and comparison to human," *Computers and Electronics in Agriculture*, vol. 189, p. 106404, 2021.

[2] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9592–9600, 2019.

[3] A. M. Roy, J. Bhaduri, T. Kumar, and K. Raj, "Wildect-yolo: An efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection," *Ecological Informatics*, vol. 75, p. 101919, 2023.

[4] D. Miller, L. Nicholson, F. Dayoub, and N. Sünderhauf, "Dropout sampling for robust object detection in open-set conditions," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3243–3249, IEEE, 2018.

[5] A. Dhamija, M. Gunther, J. Ventura, and T. Boult, "The overlooked elephant of object detection: Open set," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1021–1030, 2020.

[6] A. Rosenfeld, R. Zemel, and J. K. Tsotsos, "The elephant in the room," *arXiv preprint arXiv:1808.03305*, 2018.

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

[8] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1563–1572, 2016.

[9] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*, pp. 1321–1330, PMLR, 2017.

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

[11] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.

[12] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, *et al.*, "The limits and potentials of deep learning for robotics," *The International journal of robotics research*, vol. 37, no. 4-5, pp. 405–420, 2018.

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

[14] X. Du, Z. Wang, M. Cai, and Y. Li, "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*, 2022.

[15] A. Wu and C. Deng, "Tib: Detecting unknown objects via two-stream information bottleneck," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[16] S. Wilson, T. Fischer, F. Dayoub, D. Miller, and N. Sünderhauf, "Safe: Sensitivity-aware features for out-of-distribution object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23565–23576, 2023.

[17] 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*, vol. 31, 2018.

[18] H. Wang, Z. Li, L. Feng, and W. Zhang, "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.

[19] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-distribution detection with deep nearest neighbors," in *International Conference on Machine Learning*, pp. 20827–20840, PMLR, 2022.

[20] 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*, vol. 33, pp. 11839–11852, 2020.

[21] A. Radford, J. W. 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*, pp. 8748–8763, PMLR, 2021.

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

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

[24] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, *et al.*, "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*, pp. 24185–24198, 2024.

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

[26] S. Esmaeilpour, B. Liu, E. Robertson, and L. Shu, "Zero-shot out-of-distribution detection based on the pre-trained model clip," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, pp. 6568–6576, 2022.

[27] 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*, pp. 1802–1812, 2023.

[28] D. Hendrycks and K. Gimpel, "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*, 2017.

[29] S. Liang, Y. Li, and R. Srikant, "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*, 2018.

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

[31] C. S. Sastry and S. Oore, "Detecting out-of-distribution examples with gram matrices," in *International Conference on Machine Learning*, pp. 8491–8501, PMLR, 2020.

[32] Y.-C. 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*, pp. 10951–10960, 2020.

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

[34] 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*, pp. 2636–2645, 2020.

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

[36] 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*, vol. 128, no. 7, pp. 1956–1981, 2020.

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

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.