# Unknown object detection by leveraging Segment Anything Model for labeling the OOD objects

Seher Kanwal
*Dept of Artificial Intelligence*
*University of Science & Technology*
Daejeon, South Korea
seherkanwal223@gmail.com

Assefa Seyoum Wahd
*Dept of Radiology & Diagnostic Imaging*
*University of Alberta*
Edmonton, Canada
assefas221@gmail.com

Seungsik Lee
*Dept of automobile & IT conversion*
*Kookmin University*
Seoul, South Korea
sungsik090@gmail.com

Minsu Jang
*Dept of Artificial Intelligence*
*Electronics & Telecommunications Research Institute*
*University of Science & Technology*
Daejeon, South Korea
minsu@etri.re.kr

Seung-Ik Lee*
*Dept of Artificial Intelligence*
*Electronics & Telecommunications Research Institute*
*University of Science & Technology*
Daejeon, South Korea
the_silee@etri.kr.re

***Abstract*—This paper proposes the use of class-agnostic foundation models, such as the segment anything model (SAM), for a realistic object detection scenarios where the model should detect bounding boxes for known and unknown objects. We utilize unknown bounding boxes proposed by SAM to train a faster R-CNN model-based OOD detector. Our method outperforms the previous best approaches with a 9.01% decrease in FPR95 and an 4.69% increase in AUROC. However, our approach comes with a drop of 8.2% in the known class mAP. Our findings highlight the potential of leveraging class-agnostic object proposal models for real-time OOD detection. Our proposed method adds OOD detection ability to the faster R-CNN model without adding any computational overhead.**

***Index Terms*—Out of Distribution Object Detection**

## I. Introduction

Standard object detection (OD) models classify object proposals that do not overlap with a labeled object as background and have been observed to assign high posterior probabilities to out-of-distribution (OOD) test inputs [1]. These OOD samples should not be classified by the model as they originate from unknown categories. This naturally leads us to turn our attention towards Open-world Object Detection (OWOD) [2] which aims to discover these unknown objects. Therefore, simply applying OD methods to OWOD fails as unknown objects would be predicted as background.

One of the main challenges in OOD detection lies in the scarcity of diverse and high-quality OOD datasets. Some prior works attempted to tackle this limitation by generating synthetic unknown instances or by collecting or reusing existing separate and additional large-scale datasets for outlier exposure. Existing methods to get OOD objects using unknown discovery strategies, such as selecting areas with high activation maps, tends to choose the background or parts of the known objects as unknowns [3].

*Corresponding author

ORE [5] incorporates contrastive clustering, an unknown-aware proposal network, and energy-based unknown identification for open-world object detection (OWOD). OW-DETR [6] generates pseudo-unknowns by identifying regions with the highest activation maps that do not overlap with any ground truth bounding box. PROB [7] fits a Gaussian distribution on the query embeddings of the DETR framework. Queries of unknown objects are assumed to fall in the low-density regions of the Gaussian distribution. Virtual outlier synthesis (VOS) [8] proposed to generate virtual outliers from low-likelihood region of the Gaussian distribution formed from the empirical means and standard deviations of the RoI-pooled features. SAFE [9] use residual layers to extract distinguish in-distribution features and trains a network to classify adversarially perturbed ID example as OOD.

Although these approaches have been shown to be effective and successful in detecting OOD samples, they pose significant issues including the potential introduction of noise in case of acquiring auxiliary OOD datasets, or computational overhead and additional neural network modules for synthetic OOD data generation. To address these issues, we propose to utilize bounding boxes proposed by the segment anything model (SAM) [4] to train OD models as the proposals are more reliable than manually designing unknown discovery strategies.

## II. THE PROPOSED METHOD

In this work, we propose a 2-stage training method for an out-of-distribution (OOD) detector using bounding boxes proposed by the class-agnostic segment-anything model (SAM). SAM is segmentation model that can be prompted with points, bounding boxes, or text to segment objects of interest. To segment all objects in a given image, we prompt SAM with a regular grid of points ($32 \times 32$) and predict a segmentation mask for each point.

Overlapping predictions are removed with non-maximum suppression (NMS). To obtain bounding boxes for unknown objects, we remove proposals having an IoU greater than 0.5 with any ground-truth bounding box of known categories, and label the rest of the proposals as unknown. We then train a faster R-CNN [10] object detector with a $K+1-$way classification branch, where the $K+1-th$ class represents the unknown category.

Our approach can be considered as a real-data outlier exposure since the bounding boxes proposed by SAM contain reliable foreground objects. It is important to note that foundation models like SAM cannot be used for real-time applications due to their computational complexity. On the other hand, faster R-CNN is a lightweight model and can be used in real-time. Therefore, our work can be considered as a form of knowledge distillation, where SAM distills its knowledge to the region proposal network (RPN) of a faster R-CNN model.

## III. EXPERIMENTS

We use a faster version of SAM called Fast SAM [11] as the bounding box proposal network. Bounding boxes that do not overlap with any labeled bounding boxes are labeled as unknowns. We use Faster R-CNN with ResNet-50 as our backbone architecture, and trained with a $K+1$ classifier, where last class represents the unknown class. Unknown labeling is done offline prior to training. We use PASCAL-VOC dataset as the ID dataset and COCO dataset as OOD. Images of overlapping classes with the ID dataset are filtered out from COCO dataset.

## IV. ANOMALY SCORE

At inference time, a threshold-based approach is used to detect OOD samples as shown in (1). Usually, the threshold $\tau$ is chosen to ensure that about 95% of ID samples are accurately detected. For a given input $x$, a score function $s_\theta(x) \in \mathbb{R}$ is used to assign a higher value to ID inputs and a lower value to OOD inputs. The score function measures how likely $x$ is to come from the training data distribution. If the score $s_\theta(x)$ is low, then the input $x$ is likely to be OOD data.

$$Prediction(\mathbf{x};\tau) = \begin{cases} \text{OOD}, & \text{if } s_\theta(\mathbf{x}) \leq \tau, \\ \text{ID}, & \text{if } s_\theta(\mathbf{x}) > \tau, \end{cases} \tag{1}$$

The scoring function $s_\theta(x) \in \mathbb{R}$ is defined as the maximum softmax probability of the $K$ classes only (excluding the K+1 class). This enables us to obtain lower scores for the OOD data as the logit values over the K classes tend to be lower and more uniform, resulting in the lower maximum softmax probability. In contrast, ID inputs yield higher scores due to stronger activation on the logit of one or more known classes. This provides a straightforward measure for how confident the model is that the input belongs to any known class.

## V. RESULT EVALUATION

Following the standard OOD detection evaluation, we report two metrics: FPR95 and AUROC. FPR95 measures the false positive rate (FPR) of the OOD samples when the true positive rate (TPR) of ID samples is fixed at 95%. The performances of the proposed approach outperform previous approaches by a large margin. All baseline approaches mentioned in Table. I, were trained solely on in-distribution datasets.

VOS [10] leverages synthetic outliers generated in the feature space during training and SAFE [9] use adversarially perturbed ID example as OOD. Our method takes advantage of SAM [6] to include real outlier examples from the in-distribution dataset during training, leading to better performance in OOD detection with a 9.01% decrease in FPR95 and an 4.69% increase in AUROC. We also evaluate the mAP of the known classes. However, our approach comes with a drop of 8.2% in known class mAP. Table. I illustrates that our proposed approach has better OOD detection performance.

TABLE I
OOD DETECTION AND IN-DISTRIBUTION DETECTION RESULTS ON MS-COCO (OOD) AND PASCAL-VOC (ID)

| **Method** | **MS-COCO** | | **PASCAL-VOC** |
|---|---|---|---|
| | ***FPR95*** ↓ | ***AUROC*** ↑ | ***mAP*** ↑ |
| MSP [12] | 70.99 | 83.45 | 48.7 |
| ODIN [13] | 59.82 | 82.20 | 48.7 |
| Mahalanobis [14] | 96.46 | 59.25 | 48.7 |
| Energy [15] | 56.89 | 83.69 | 48.7 |
| Gram [16] | 62.75 | 79.88 | 48.7 |
| Gen. ODIN [17] | 59.57 | 83.12 | 48.1 |
| CSI [18] | 59.91 | 81.83 | 48.1 |
| GAN-syn. [19] | 60.93 | 83.67 | 48.5 |
| VOS-ResNet50 [8] | 47.53 | 88.70 | 48.9 |
| VOS-RegX4.0 [8] | 47.77 | 89.00 | **51.6** |
| SAFE-ResNet50 [9] | 47.40 | 80.30 | - |
| **SAM-ResNet50** | **38.39** | **93.69** | 43.4 |

## VI. CONCLUSION

Our work demonstrates that class-agnostic foundation models such as SAM can be effectively used in realistic object detection settings, where the model should detect bounding boxes for known and unknown objects. By utilizing the bounding boxes proposed by SAM, we were able to train a faster R-CNN model that outperforms previous best approaches. Our findings highlight the potential of leveraging class-agnostic object proposal models for real-time OOD detection.

## VII. ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00951, Development of Uncertainty-Aware Agents Learning by Asking Questions).

## REFERENCES


[1] A. T. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 427–436, 2015.

[2] K. J. Joseph, S. Khan, F. S. Khan, and V. N. Balasubramanian, "Towards open world object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5830–5840, 2021.

[3] Y. Ma et al., "Annealing-based Label-Transfer Learning for Open World Object Detection," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 11454-11463, doi: 10.1109/CVPR52729.2023.01102.

[4] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv:2304.02643*, 2023.

[5] K. J. Joseph, S. Khan, F. S. Khan, and V. N. Balasubramanian, "Towards open world object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5830–5840, 2021.

[6] A. Gupta, S. Narayan, K. J. Joseph, S. Khan, F. S. Khan, and M. Shah, "OW-DETR: Open-world detection transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9235–9244, 2022.

[7] O. Zohar, K.-C. 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*, pp. 11444–11453, 2023.

[8] X. Du, Z. Wang, M. Cai, and Y. Li, "VOS: Learning what you don't know by virtual outlier synthesis," *arXiv preprint arXiv:2202.01197*, 2022.

[9] Wilson, Samuel et al. "SAFE: Sensitivity-Aware Features for Out-of-Distribution Object Detection." 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (2022): 23508-23519.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, pp. 91–99, 2015.

[11] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang, "Fast segment anything," *arXiv preprint arXiv:2306.12156*, 2023.

[12] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[13] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *International Conference on Learning Representations (ICLR)*, 2018.

[14] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Advances in Neural Information Processing Systems*, pp. 7167–7177, 2018b.

[15] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Advances in Neural Information Processing Systems*, 2020.

[16] C. S. Sastry and S. Oore, "Detecting out-of-distribution examples with gram matrices," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, vol. 119, pp. 8491–8501, 2020.

[17] 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.

[18] J. Tack, S. Mo, J. Jeong, and J. Shin, "CSI: Novelty detection via contrastive learning on distributionally shifted instances," in *Advances in Neural Information Processing Systems*, 2020.

[19] K. Lee, H. Lee, K. Lee, and J. Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," in *International Conference on Learning Representations (ICLR)*, 2018a.