# Learning Latent Concepts for Detecting Out-of-Distribution Objects

Supplementary Material

## 6. OOD-OD Results on Faster R-CNN

In Table 7, we demonstrate the effectiveness of our approach based on the Faster R-CNN [37] detection model. Compared to the previous best method, such as DFDD [54], UNO-Adapter significantly reduces the FPR95 by 2.61% and 8.35% on MS-COCO and OpenImages OOD datasets. It is noteworthy that the performance gap between different baseline methods is less significant compared to the DETR-based results. The justification is that the two-stage nature of Faster R-CNN limits its scalability in incorporating different types of OOD-OD methods. Despite the challenge, our approach can still achieve substantial improvements, revealing the superiority and potential of our unknown injection framework. By contrast, those methods that rely on task-related training objectives may fail to generalize well across distinct detection models. On the other hand, the proposed UNO-adapter can be simplified into a purely post-hoc method by fixing the slot attention component, eliminating the need for additional fine-tuning.

| Method | MS-COCO | | OpenImages | |
|---|---|---|---|---|
| | AUROC↑ | FPR95↓ | AUROC↑ | FPR95↓ |
| CSI [44] | 82.95 | 57.41 | 81.83 | 59.91 |
| GAN-Synthesis [29] | 82.67 | 59.97 | 83.67 | 60.93 |
| VOS [12] | 85.23 | 51.33 | 88.70 | 47.53 |
| SIREN [10] | 85.36 | 64.68 | 82.78 | 68.53 |
| TIB [50] | 90.36 | 41.55 | 88.09 | 47.19 |
| DFDD [53] | 90.79 | 41.34 | 88.65 | 44.52 |
| WFS [52] | 89.01 | 40.05 | 90.35 | 39.17 |
| **UNO-Adapter** | 91.25 | 38.73 | 92.40 | 35.74 |

Table 7. The performance (%) of OOD-OD based on Faster R-CNN. ID: Pascal-VOC; OOD: MS-COCO and OpenImages.

## 7. OOD Image Classification

In this section, we provide the datasets and implementation details for the OOD image classification task (see Tab. 6 in the main paper).

### 7.1. Datasets

We begin the OOD classification test with the ImageNet-1k [8] as the ID dataset. We use the standard split with 1,280,000 training images and 50,000 validation images. We evaluate all methods on the near and far OOD datasets: SSB-hard [47], NINCO [3], iNaturelist [46], and Texture [7]. All input images will be resized to 224 $\times$ 224.

### 7.2. Implementation Details

Following [40], we leverage the self-supervised pre-trained DINO as the feature extractor and optimize the feature reconstruction loss to obtain a set of slot vectors. Instead of repeatedly training this component on all ID datasets, we train it once on a specific ID dataset, such as the Pascal-VOC [14] dataset, and keep it fixed across all OOD-OD tasks. Building on this, we perform fine-tuning and utilize the proposed slot refinement module to refine each slot. Notably, object-centric methods retain their pre-trained parameters, eliminating the need for retraining. Once the slot attention mechanism is trained, each slot gains the capability to abstract objects, allowing for direct testing on the BDD-100K dataset without requiring further retraining of the slot attention component. The detailed training configurations and hyperparameters are presented in Tab. 8.

In Tab. 6, the OOD classification baseline model is constructed using a self-supervised DINO-ViT-B16 model with an additional MLP classification head. The model is fine-tuned for 20 epochs with an Adam optimizer, starting with an initial learning rate of $4 \times 10^{-4}$, and employing a cosine learning rate decay schedule that gradually reduces the learning rate to $5 \times 10^{-5}$. Our method enhances this baseline by integrating a Slot Attention module, which decomposes the global feature into multiple slots. The classifier processes each slot, and the resulting slot-wise predictions are aggregated to produce the final global logits.

## 8. Additional Empirical Analysis

### 8.1. Sensitivity analysis *w.r.t.* OOD score

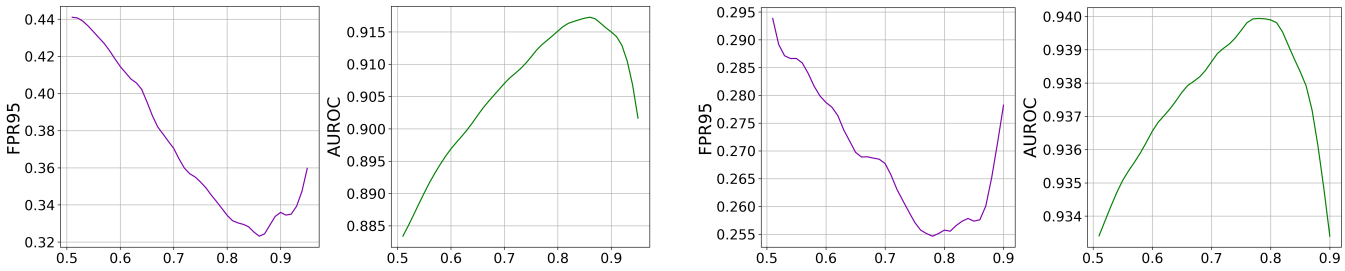
The experimental results with varying $\tau$ values are presented in Fig. 6. The experiments begin with $\tau = 0.5$ and increment by 0.1 until performance degradation is observed. As $\tau$ increases, the performance steadily improves. However, when $\tau$ approaches 1 (equivalent to MaxLogit), the performance gradually declines. This decline reveals that traditional post-processing methods are not robust enough for OOD-OD tasks, primarily due to excessive overlap and noise among candidate objects. By contrast, using quantile logits effectively addresses this challenge, enhancing the robustness of our method.

### 8.2. Visualization results

In Figure 7, we provide the model's activation maps and segmentation results with SA and our method. In Figure 8 we visualize the object predictions of SIREN [10], SAFE [48], and our approach on a wild test, using PASCAL-VOC as the ID

| | PASCAL-VOC | Berkeley DeepDrive-100K | ImageNet-1k |
|---|---|---|---|
| Fine-tune epochs | 30 | 30 | 30 |
| Batch Size | 64 | 64 | 64 |
| Image Size | 320 | 320 | 320 |
| Slot Dim. | 256 | 256 | 256 |
| $U$ in Eq. (4) | $256 \rightarrow 256$ | $256 \rightarrow 256$ | $256 \rightarrow 256$ |
| $V$ in Eq. (4) | $256 \rightarrow 256$ | $256 \rightarrow 256$ | $256 \rightarrow 256$ |
| Importance Score Net $h_\theta$ in Eq. (3) | $256 \rightarrow 128$<br>$128 \rightarrow 1$ | $256 \rightarrow 128$<br>$128 \rightarrow 1$ | $256 \rightarrow 128$<br>$128 \rightarrow 1$ |
| Information Bottleneck Encoder $h$ in Eq. (5) | $256 \rightarrow 128$<br>$128 \rightarrow 64$ | $256 \rightarrow 128$<br>$128 \rightarrow 64$ | $256 \rightarrow 128$<br>$128 \rightarrow 64$ |
| Information Bottleneck Decoder $g$ in Eq. (9) | $64 \rightarrow 128$<br>$128 \rightarrow 256$ | $64 \rightarrow 128$<br>$128 \rightarrow 256$ | $64 \rightarrow 128$<br>$128 \rightarrow 256$ |
| Slots number $S$ | 6 | 6 | 6 |
| $\tau$ in Eq. (14) | 0.8 | 0.7 | 0.98 |
| $\beta$ in Eq. (9) | 0.5 | 0.5 | 0.5 |

Table 8. Configurations of OOD-OD and OOD classification tasks.




(a) ID: Pascal-VOC; OOD: MS-COCO

(b) ID: Pascal-VOC; OOD: OpenImages

Figure 6. Sensitivity analysis with respect to OOD score.

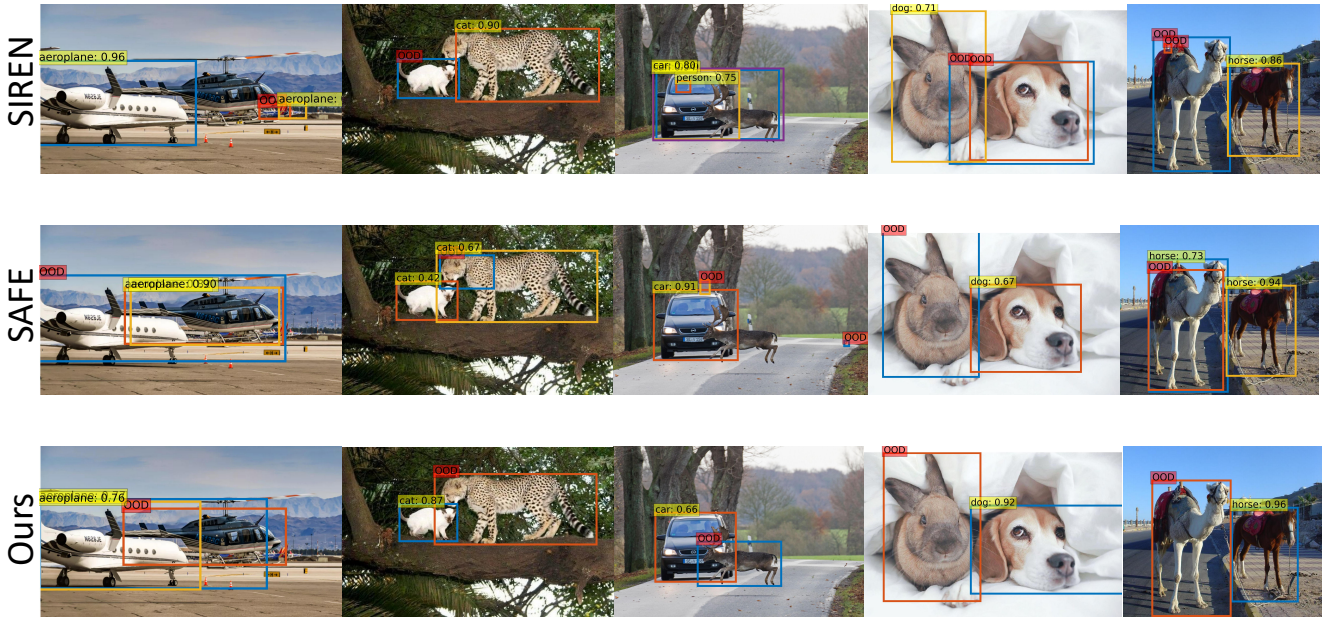



Figure 7. Additional qualitative visualization of detection results. We compare the proposed method with SIREN [10] and SAFE [48].

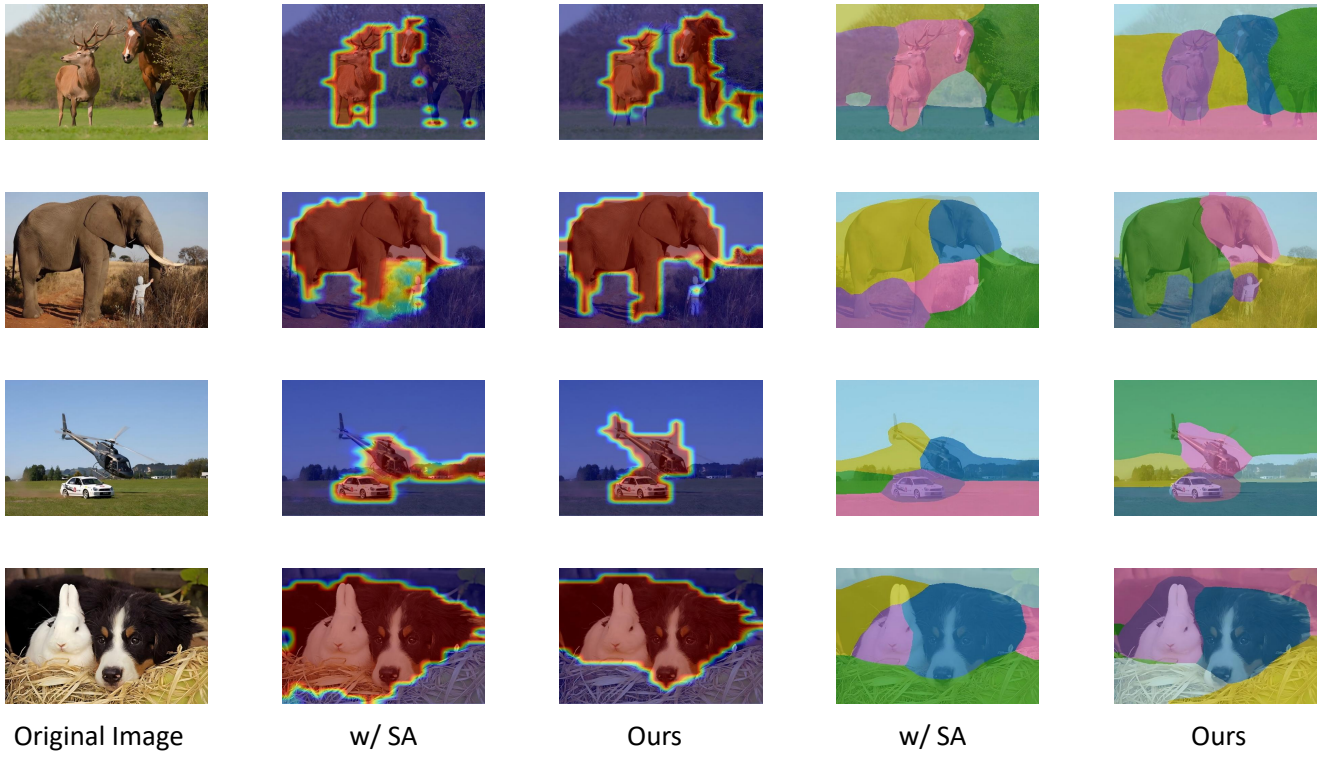
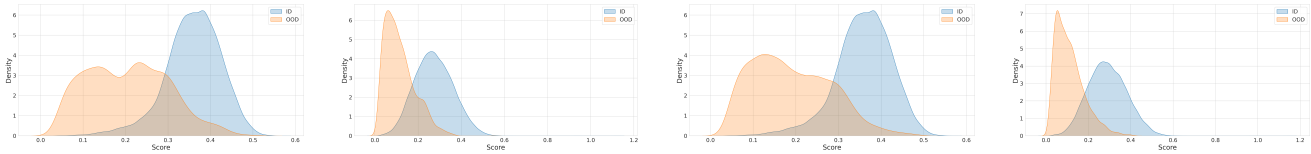



Figure 8. The visualization results of the activation map and segmentation mask. SA stands for original slot attention [35].



(a) w/o NCD&NCB (b) w/ NCD&NCB (c) w/o NCD&NCB (d) w/ NCD&NCB

Figure 9. (a)&(b): ID: Pascal VOC; OOD: MS-COCO; (c)&(d): ID: Pascal VOC; OOD: OpenImages.

| Method | MS-COCO AUROC / FPR95 | OpenImages AUROC / FPR95 |
|---|---|---|
| ResNet-34 | 90.83 / 35.61 | 94.62 / 22.50 |
| Random | 88.48 / 37.21 | 92.18 / 25.30 |
| Ours | **91.68 / 32.61** | **95.40 / 19.90** |

Table 9. OOD-OD results on PASCAL-VOC (ID).

training data. In Figure 9, we provide the distribution of our OOD score w/o and w/ the proposed unsupervised concept discovery (UCD), neural concept binder (NCB). As can be seen, the proposed NCD and NCB effectively make the distributions more concentrated and create a clearer separation between ID and OOD samples. This further validates that the proposed unknown injection can significantly enhance the model's ability to perceive and recognize unknowns.