Towards Open-Set Object Detection and Discovery: Supplementary Material

Jiyang Zheng[⋆†] Weihao Li[†] Jie Hong[⋆†] Lars Petersson[†] Nick Barnes[⋆]
[⋆]The Australian National University [†]Data61-CSIRO
firstname.lastname@{[⋆]anu.edu.au, [†]data61.csiro.au}

1. Implementation

For the object detection part, we rebuild ORE [4] with ResNet-50 [3] as the backbone network. The initial learning rate is set to 0.01. The weight decay is $1e^{-4}$. The momentum η, the margin parameter Δ and the temperature parameter $\mathcal{T}$ are empirically set to 0.9, 15 and 1 respectively. The unknown object and non-maximum suppression threshold are set to 0.5 and 0.4. We train the model on three successive tasks (*i.e.*, T_1, T_2, T_3 or Task-1, Task-2, Task-3), with 8 epochs on each task. The experiments are conducted on NVIDIA Tesla P100 4 GPUs with a batch size of 128. For object category discovery, we select ResNet-50 [3] as the backbone with no pre-trained weights. The network encodes the object instance into $256-$dimension using a linear projection head in the last layer. The learning rate is 0.015 [2] with 200 epochs for each training circle.

1.1. Semantic Split

The detailed known classes and unknown classes split of each task in MS-COCO [5], and Pascal VOC [1] are shown in Tab. 2. The intersected classes between COCO and VOC are treated as the known classes for the first task, Task-1.

2. Ablation study on Memory Module

This section provides an ablation study on the memory module to show the effects of the known memory in representation learning. The results are reported in Tab. 1. The performance of our method with only working memory is shown in Case I where the detected known objects at the training phase will not be included in the representation learning. Compared to Case II where our model is using both working and known memory, we can see that the performance of I is worse over three tasks. It shows the design of our memory module is important for class discovery as it allows the model to learn more generalised embedding representations.

3. Mutual Information and Entropy

We have introduced the normalised mutual information for clustering performance evaluation. Here, we provide

	Memory Module		Task-1			Task-2			Task-3		
	Known	Unknown	NMI	ACC	Purity	NMI	ACC	Purity	NMI	ACC	Purity
I	✗	✓	8.8	5.6	9.9	5.2	6.3	12.4	5.5	11.7	28.2
II	✓	✓	**11.0**	**6.3**	**12.6**	**5.8**	**6.9**	**13.3**	**6.5**	**16.4**	**29.3**

Table 1. Ablation Study on Memory Module.

the formulation for the two major components in the normalised mutual information, which are the mutual information (MI) and the entropy (H). Let Cl be the set of ground truth classes, and $\widehat{Cl}$ be the set of predicted clusters. The MI and entropy are formulated as:

$$
\begin{aligned}
I(Cl,\widehat{Cl}) &= \sum_k \sum_j P(Cl_k \cap \widehat{Cl}_j) \log \frac{P(Cl_k \cap \widehat{Cl}_j)}{P(Cl_k)P(\widehat{Cl}_j)} \\
H(Cl) &= -\sum_k P(Cl_k) \log P(Cl_k) \\
H(\widehat{Cl}) &= -\sum_j P(\widehat{Cl}_j) \log P(\widehat{Cl}_j)
\end{aligned}
\tag{1}
$$

where $P(Cl_k)$, $P(\widehat{Cl}_j)$ and $P(Cl_k \cap \widehat{Cl}_j)$ are the probabilities of a object being in cluster Cl_k, $\widehat{Cl}_j$ and $Cl_k \cap \widehat{Cl}_j$ respectively. The probability is calculated as the number of corresponding objects divided by the total number of instances.

4. Qualitative Analysis

4.1. Open-Set Detection and Discovery Results

In Fig. 1, we visualise the OSODD predictions under two tasks, Task-1 and Task-2. The left figure shows the prediction in Task-1, where the *zebra* and *giraffe* class are not introduced. The model successfully distinguishes two unknown animals. The right figure shows the prediction in Task-2 where the annotations of *zebra* and *giraffe* are made available. More results are shown in Fig. 2. We have also encountered failure cases, as shown in Fig. 3, where the model incorrectly assigns the *piazzas* to two novel categories. Additionally, in the second row, there is a false detection on *person* class. However, after the *piazza* class get introduced in Task-3, the model makes the correct pre-

Task-1									
Airplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow
Dining table	Dog	Horse	Motorcycle	Person	Potted plant	Sheep	Couch	Train	Tv
Truck	Traffic light	Fire hydrant	Stop sign	Parking meter	Bench	Elephant	Bear	Zebra	Giraffe
Backpack	Umbrella	Handbag	Tie	Suitcase	Microwave	Oven	Toaster	Sink	Refrigerator
Frisbee	Skis	Snowboard	Sports ball	Kite	Baseball bat	Baseball glove	Skateboard	Surfboard	Tennis racket
Banana	Apple	Sandwich	Orange	Broccoli	Carrot	Hot dog	Pizza	Donut	Cake
Bed	Toilet	Laptop	Mouse	Remote	Keyboard	Cell phone	Book	Clock	Vase
Scissors	Teddy bear	Hair drier	Toothbrush	Wine glass	Cup	Fork	Knife	Spoon	Bowl

Task-2									
Airplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow
Dining table	Dog	Horse	Motorcycle	Person	Potted plant	Sheep	Couch	Train	Tv
Truck	Traffic light	Fire hydrant	Stop sign	Parking meter	Bench	Elephant	Bear	Zebra	Giraffe
Backpack	Umbrella	Handbag	Tie	Suitcase	Microwave	Oven	Toaster	Sink	Refrigerator
Frisbee	Skis	Snowboard	Sports ball	Kite	Baseball bat	Baseball glove	Skateboard	Surfboard	Tennis racket
Banana	Apple	Sandwich	Orange	Broccoli	Carrot	Hot dog	Pizza	Donut	Cake
Bed	Toilet	Laptop	Mouse	Remote	Keyboard	Cell phone	Book	Clock	Vase
Scissors	Teddy bear	Hair drier	Toothbrush	Wine glass	Cup	Fork	Knife	Spoon	Bowl

Task-3									
Airplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow
Dining table	Dog	Horse	Motorcycle	Person	Potted plant	Sheep	Couch	Train	Tv
Truck	Traffic light	Fire hydrant	Stop sign	Parking meter	Bench	Elephant	Bear	Zebra	Giraffe
Backpack	Umbrella	Handbag	Tie	Suitcase	Microwave	Oven	Toaster	Sink	Refrigerator
Frisbee	Skis	Snowboard	Sports ball	Kite	Baseball bat	Baseball glove	Skateboard	Surfboard	Tennis racket
Banana	Apple	Sandwich	Orange	Broccoli	Carrot	Hot dog	Pizza	Donut	Cake
Bed	Toilet	Laptop	Mouse	Remote	Keyboard	Cell phone	Book	Clock	Vase
Scissors	Teddy bear	Hair drier	Toothbrush	Wine glass	Cup	Fork	Knife	Spoon	Bowl

Table 2. Semantic splits for Task-1, Task-2 and Task-3. Known classes are highlighted in blue. Unknown classes are highlighted in yellow.

dictions on all *piazza* instances and eliminates the ambiguity of the novel categories. Two different failure cases are shown in Fig. 4 and Fig. 5. In the first case (See Fig. 4), the detector incorrectly classifies the unknown objects as the known classes. In the second case (See Fig. 5), the detector correctly finds the novel category for the unknown objects when the labels are not available, but it does not detect the objects when the actual class is introduced. This suggests there is still a large space to be improved.

4.2. Object Category Discovery Results

In Fig. 6, we visualise some discovered object clusters from the first task, Task-1. We assume that 20 classes are known and the rest 60 classes are unknown (See 'Task-1' in Tab. 2). Most clusters can be quantitatively evaluated by the ground-truth labels in the validation step. Some objects or categories of interest are not annotated by human in the original dataset (*e.g.* plate). Surprisingly, our model can identify those un-annotated objects and cluster them to find new novel categories (See 'PLATE' in Fig. 6). It is noticed that some objects from the known class have been falsely predicted as unknown and clustered into novel categories (*e.g.* potted plant).

References

[1] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1

[2] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[4] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5830–5840, 2021. 1

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

Figure 1. OSODD predictions in Task-1 (Left) and Task-2 (Right). In Task-1, the model successfully localises the unknown objects and recognise them as two different categories. In Task-2, the wild animal classes including *zebra* and *giraffe* are introduced to model, it correctly classifies the objects into their corresponding classes.

Figure 2. The left column shows the results in Task-1 where only 20 classes are available. The right column shows the results in Task-2 where 20 new classes, like *stop sign* and *fire hydrant*, have been introduced to the model.

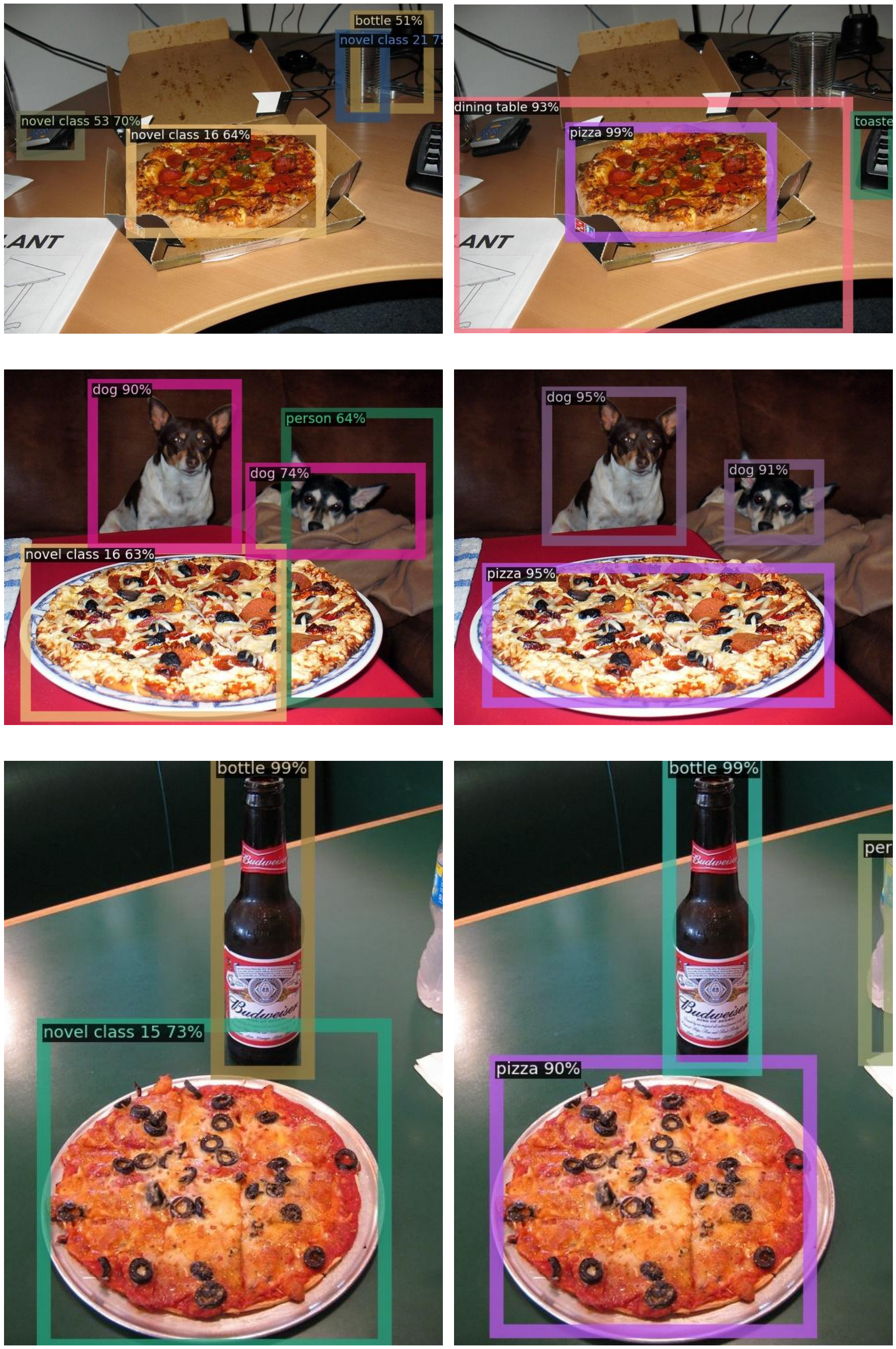

Figure 3. The left column shows the predictions in Task-1 where the *piazzas* are clustering into two novel categories. The right column shows the predictions in Task-3 where all the *piazza* objects are correctly classified.

Figure 4. The left column shows the predictions in Task-1. The banana has not been introduced, the model has correctly predicted the novel categories. The right column shows the predictions in Task-2 where the food classes are not available for the task. The model should predict the banana into one of the novel categories, but it incorrectly classifies the unknown object into one of the known classes.

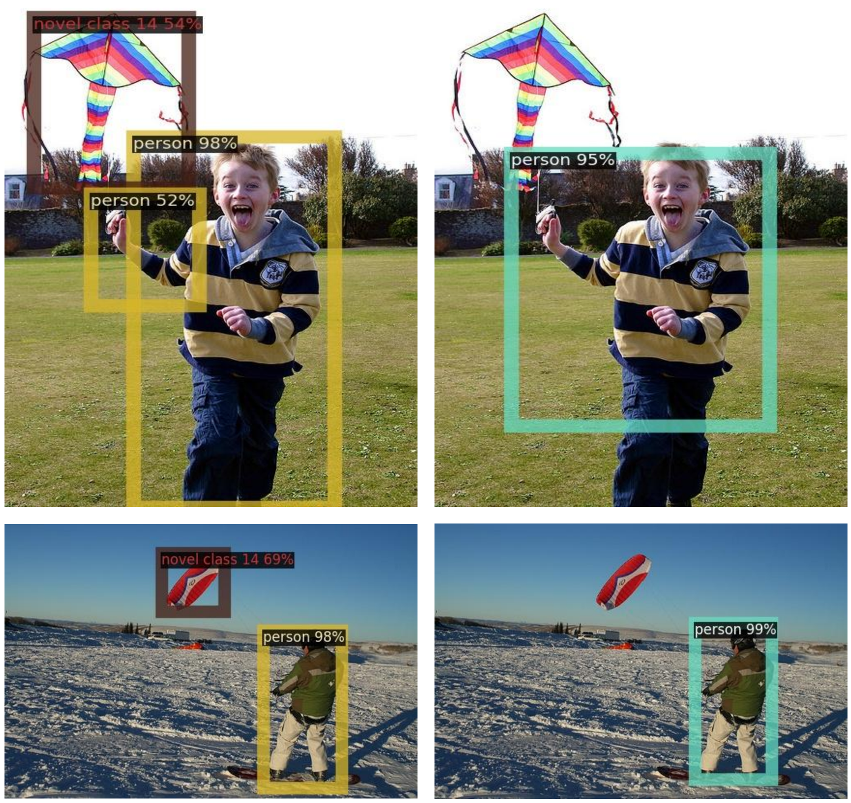

Figure 5. A failure case in open-set learning. The model successfully discovers the novel category for the *kite* class in Task-1 (Left). However, after more semantic classes of labels are provided in the following task, Task-2, the model fails to localise the *kites* in the image.

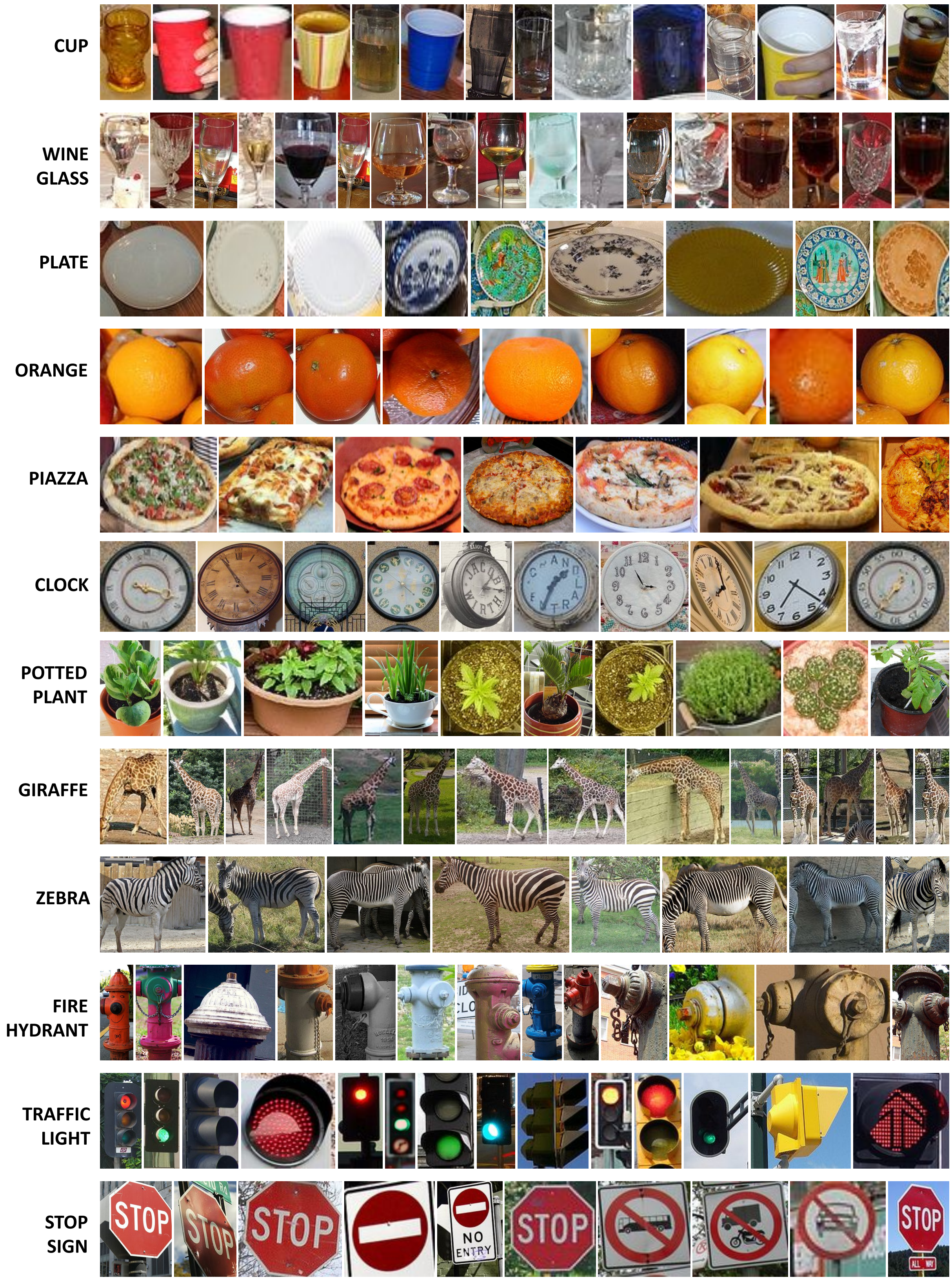

Figure 6. Some discovered results from object category discovery.