Towards Open-Set Object Detection and Discovery

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

Abstract

With the human pursuit of knowledge, open-set object detection (OSOD) has been designed to identify unknown objects in a dynamic world. However, an issue with the current setting is that all the predicted unknown objects share the same category as "unknown", which require incremental learning via a human-in-the-loop approach to label novel classes. In order to address this problem, we present a new task, namely Open-Set Object Detection and Discovery (OSODD). This new task aims to extend the ability of open-set object detectors to further discover the categories of unknown objects based on their visual appearance without human effort. We propose a two-stage method that first uses an open-set object detector to predict both known and unknown objects. Then, we study the representation of predicted objects in an unsupervised manner and discover new categories from the set of unknown objects. With this method, a detector is able to detect objects belonging to known classes and define novel categories for objects of unknown classes with minimal supervision. We show the performance of our model on the MS-COCO dataset under a thorough evaluation protocol. We hope that our work will promote further research towards a more robust real-world detection system.

1. Introduction

Object detection is the task of localising and classifying objects in an image. In recent years, deep learning approaches have advanced the detection models [3, 4, 15, 20, 37, 38, 45] and achieved remarkable progress. However, these methods work under a strong assumption that all object classes are known at the training phase. As a result of this assumption, object detectors would incorrectly treat objects of unknown classes as background or classify them as belonging to the set of known classes [11] (see Fig. 1(a)).

To relax the above closed-set condition, open-set object detection (OSOD) [11, 24, 32] considers a realistic scenario where test images might contain novel classes that did not appear during training. OSOD aims at jointly detecting objects from the set of known classes and localising objects that belong to an unknown class. Although OSOD has improved the practicality of object detection by enabling detection of instances of unknown classes, there is still the issue that all identified objects of an unknown class share the same category as "unknown" (see Fig. 1(b)). Additional human annotation is required to incrementally learn novel object categories [24].

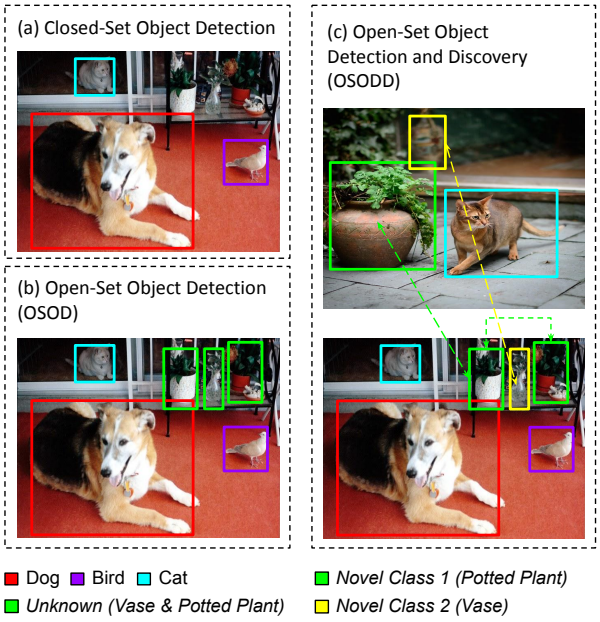

Figure 1. A visual comparison of object detection tasks. In closed-set detection, objects from unseen classes are ignored or incorrectly classified into the set of known classes. While in open-set object detection, unknown objects are localised but share the same category. Our task aims to detect objects of known classes and discover novel visual categories for the identified objects of unknown classes, which provides better scene understanding and a scalable learning paradigm.

Consider a child who is visiting a zoo for the first time. The child can recognise some animals that are seen and learned before, for example, 'rabbit' or 'bird', while the child might not recognise the species of many other rarely seen animals, like 'zebra' and 'giraffe'. After observing, the

child's perception system will learn from these previously unseen animals' appearances and cluster them into different categories even without being told what species they are.

In this work, we consider a new task, where we aim to localise objects of both known and unknown classes, assign pre-defined category labels for known objects, and discover new categories for objects of unknown classes (see Fig. 1(c)). We term this task *Open-Set Object Detection and Discovery* (OSODD). We motivate our proposed task, OSODD, by suggesting that it is better suited to extracting information from images. New category discovery provides additional knowledge of data belonging to classes not seen before, helping intelligent vision-based systems to handle more realistic use cases.

We propose a two-stage framework to tackle the problem of OSODD. First, we leverage the ability of an open set object detector to detect objects of known classes and identify objects of unknown classes. The predicted proposals of objects of known and unknown classes are saved to a memory buffer; Second, we explore the recurring pattern of all objects and discover new categories from objects of unknown classes. Specifically, we develop a self-supervised contrastive learning approach with domain-agnostic data augmentation and semi-supervised k-means clustering for category discovery.

Our contributions:

- We formalise the task Open-Set Object Detection and Discovery (OSODD), which enables a richer understanding within real-world detection systems.
- We propose a two-stage framework to tackle this problem, and we present a comprehensive protocol to evaluate the object detection and category discovery performance.
- We propose a category discovery method in our framework using domain-agnostic augmentation, contrastive learning and semi-supervised clustering. The novel method outperforms other baseline methods in experiments.

2. Related Work

Open-Set Recognition. Compared with closed-set learning, which assumes that only previously known classes are present during testing, open-set learning assumes the co-existence of known and unknown classes. Scheirer *et al*. [40] first introduce the problem of open-set recognition with incomplete knowledge at training time, *i.e*., unknown classes can appear during testing. They developed a classifier in a one-vs-rest setting, which enables the rejection of unknown samples. [22, 41] extend the framework in [40] to a multi-class classifier using probabilistic models with the extreme value theory to minimise fading confidence of the classifier. Recently, Liu *et al*. [31] proposed a deep metric learning method to identify unseen classes for imbalanced datasets. Self-supervised learning [14, 35, 43] approaches have been explored to minimise external supervision.

Task	Dataset	Known classes	Unknown classes
ODL	Open-Set	Non-Action	Loc/Cat
OSOD	Open-Set	Detect	Loc
OSODD (Ours)	Open-Set	Detect	Loc/Cat

Table 1. Comparisons of different Object Detection and Discovery tasks. OSOD: open-set object detection; ODL: Object discovery and localization. *Loc* means localise the objects of interest; *Cat* means discover novel categories.

Miller *et al*. [32] first investigate the utility of label uncertainty in object detection under open-set conditions using dropout sampling. Dhamija *et al*. [11] define the problem of open-set object detection (OSOD) and conducted a study on traditional object detectors for their abilities in avoiding classifying objects of unknown classes into one of the known classes. An evaluation metric is also provided to assess the performance of the object detector under the open-set condition.

Open-World Recognition. The open-world setting introduced a continual learning paradigm that extends the open-set condition by assuming new semantic classes are introduced gradually at each incremental time step. Bendale *et al*. [2] first formalise the open-world setting for image recognition and propose an open-set classifier using the nearest non-outlier algorithm. The model evolves when new labels for the unknown are provided by re-calibrating the class probabilities.

Joseph *et al*. [24] transfer the open-world setting to an object detection system and propose the task of open-world object detection (OWOD). The model uses example replay to make the open-set detector learn new classes incrementally without forgetting the previous ones. The OWOD or OSOD model cannot explore the semantics of the identified unknown objects, and extra human annotation is required to learn novel classes incrementally. In contrast, our OSODD model can discover novel category labels for objects of unknown classes without human effort.

Novel Category Discovery. The novel category discovery task aims to identify similar recurring patterns in the unlabelled dataset. In image recognition, it was earlier viewed as an unsupervised clustering problem. Xie *et al*. [46] proposed the deep embedding network that can cluster data and at the same time learn a data representation. Han *et al*. [18] formulated the task of novel class discovery (NCD), which clusters the unlabelled images into novel categories using deep transfer clustering. The NCD setting assumes that the training set contains both labelled and unlabelled

data, the knowledge learned on labelled data could be transferred to targeted unlabelled data for category discovery [13, 17, 23, 48, 52].

Object discovery and localisation (ODL) [6,9,27–29,36] aims to jointly discover and localise dominant objects from an image collection with multiple object classes in an unsupervised manner. Lee and Grauman [27] used object-graph and appearance features for unsupervised discovery. Rambhat *et al*. [36] assumed partial knowledge of class labels and conducted the discovery leveraging a dual memory module. Compared to ODL, our OSODD both performs detection on previously known classes and discovers novel categories for unknown objects, which provide a comprehensive scene understanding.

Please refer to Tab. 1 to see the summarised differences between our setting and other similar settings in the object detection problem.

3. Task Format

In this section, we formulate the task of Open-Set Detection and Discovery (OSODD). We have a set of known object classes $C_k = \{C_1, C_2, \cdots, C_m\}$, and there exists a set of unknown visual categories $C_u = \{C_{m+1}, C_{m+2}, \cdots, C_{m+n}\}$, where $C_k \cap C_u = \varnothing$. The training dataset contains objects from C_k, and the testing dataset contains objects from $C_k \cup C_u$. An object instance I is represented by $I = [c, x, y, w, h]$, denoting the class label ($c \in C_k$ or C_u), the top-left x, y coordinates, and the width and height from the centre of the object bounding box respectively. A model is trained to localise all objects of interest. Then, it classifies objects of a known class as one of C_k^t and clusters objects of an unknown class into novel visual categories C_u^t.

4. Our Approach

This section describes our approach for tackling OSODD, beginning with an overview of our framework. We propose a generic framework consisting of two main modules, *Object Detection and Retrieval (ODR)* and *Object Category Discovery (OCD)* (see Fig. 2).

The ODR module uses an open-set object detector with a dual memory buffer for object instances detection and retrieval. The detector predicts objects of known classes with their semantic labels from C_k and the location information, where the unknown objects are localised but with no semantic information available. We store the predicted objects in the memory buffer [36], which is used to explore novel categories. The buffer is divided into two parts: *known memory* and *working memory*. The known memory contains predicted objects of known classes with semantic labels; the working memory stores all current identified objects of unknown classes without categorical information. The model studies the recurring pattern of the objects from the memory buffer and discovers novel categories in the working memory. We assign the predicted objects of unknown classes from the detector with novel category labels using the discovered categories. The visualisation is shown in Fig. 4.

The OCD module explores the *working memory* to discover new visual categories. It consists of an encoder component as the feature extractor and a discriminator which clusters the object representations. To train the encoder, we first retrieve the predicted objects from known classes saved in the known memory and the identified objects of unknown classes saved in working memory. Then, these instance samples are transformed using class-agnostic augmentation to create a generalised view over the data [10, 26, 51]. We use unsupervised contrastive learning where the predicted labels for the objects of known classes are ignored, the pairwise contrastive loss [33] penalises dissimilarity of the same object in different views regardless of the semantic information. The contrastive learning enables the encoder to learn a more discriminating feature representation in the latent space [7, 19]. Lastly, with the learned feature space from the encoder, the discriminator clusters the object embedding into novel categories by using the constrained k-means clustering algorithm [44].

4.1. Object Detection and Retrieval

Open-Set Object Detector. An open-set object detector predicts the location of all objects of interest. Then it classifies the objects into semantic classes and identifies the unseen objects as unknown (See 'OSOD' in Fig. 3).

We use the Faster RCNN architecture [38] as the baseline model, following ORE [24]. Leveraging the class-agnostic property of the region proposal network, we utilise an unknown-aware RPN to identify unknown objects. The unknown-aware RPN labels the proposals that have high scores but do not overlap with any ground-truth bounding box as the potential unknown objects. To learn a more discriminative representation for each class, we use a prototype based constrictive loss on the feature vectors f_c generated by an intermediate layer in the ROI pooling head. A class prototype p_i is computed by the moving average of the class instance representations, and the features f_c of objects will keep approaching their class prototype in the latent space. The objective is formulated as:

$$\begin{aligned} \ell_{pcl}(f_c) &= \sum_{i=0}^{c} \ell(f_c, p_i) \\ \ell(f_c, p_i) &= \begin{cases} \|f_c, p_i\| & \text{if } i = c \\ \max\ (0, \Delta - \|f_c, p_i\|) & \text{otherwise} \end{cases} \end{aligned} \tag{1}$$

where f_c is the feature vector of class c, p_i is the prototype of class i, $\|f, p\|$ measures the distance between feature vectors and Δ is a fixed value that defines the maximum dis-

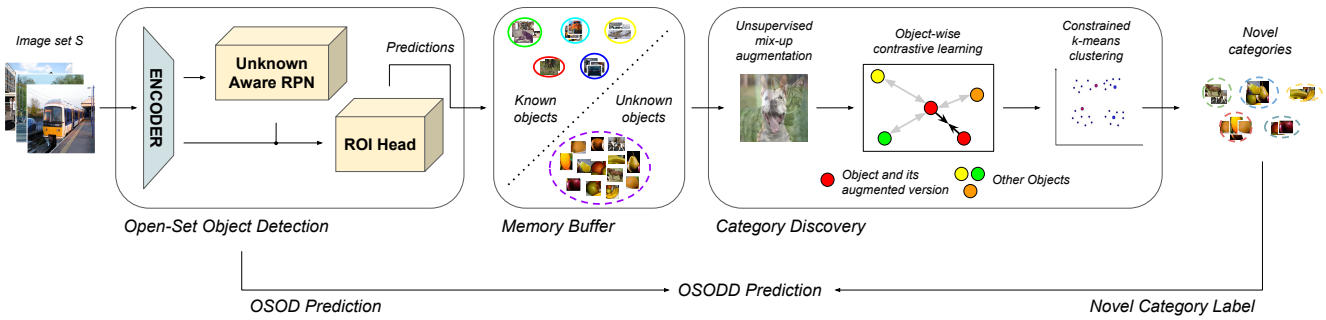

Figure 2. Illustration of the two-stage method for Open-Set Object Detection and Discovery (OSODD). The first stage includes detecting objects of known classes and identifying objects of unknown classes using an open-set object detector. The instances of unknown classes are saved into the *working memory* for category discovery. The instances of known classes are saved into the *known memory* with their predicted semantic categories to assist the representation learning and clustering. The second stage pre-processes the objects from the memory buffer in an unsupervised manner. The representations of these saved objects are first learned in the latent space by contrastive learning, followed by a constrained k-means clustering used to find the novel categories beyond the known classes. Lastly, we update the open-set detection predictions with the novel category labels to generate the final OSODD prediction (See visualisations in Figs. 3 and 4).

tance for dissimilar pairs. The total loss for the region of interest pooling is defined as:

$$\ell_{roi} = \alpha_{pcl} \cdot \ell_{pcl} + \alpha_{cls} \cdot \ell_{cls} + \alpha_{reg} \cdot \ell_{reg} \tag{2}$$

where α_{pcl}, α_{cls} and α_{reg} are positive adjustment ratios. ℓ_{cls}, ℓ_{reg} are the regular classification and regression loss.

Given the encoded feature f_c, we use an open-set classifier with an energy-based model [25] to distinguish the objects of known and unknown classes. The trained model is able to assign low energy values to known data and thus creates dissimilar representations of distribution for the objects of known and unknown classes. When new known class annotations are made available, we utilise the example replay to alleviate forgetting the previous classes.

Memory Module. As described above, we propose to use a dual memory module to store predicted instances for category discovery. The open-set detector detects the objects of interest with their locations and the predicted label. The objects of a known class I_k are saved into known memory M_k with their semantic labels $c \in C_k$. These objects are treated as a labelled dataset for the following category discovery. The identified objects of an unknown class I_u are stored in the working memory M_w. We perform the category discovery on M_w, which aims to assign all instances in M_w with a novel category label $c \in C_u$. We update the open-set object detector's prediction using the novel category labels and produce our final OSODD predictions.

4.2. Object Category Discovery

Category Number Estimation. Our category discovery approach requires an estimation of the number of potential classes. We use the class estimation method from [18], one of the most commonly used techniques for image-level novel category discovery. The model uses a k-means clustering method to estimate the category number in the target dataset without any parametric learning. The generalisation ability of the method towards our problem has been evaluated in Sec. 6.2.1.

Figure 3. Comparison between OSOD and OSODD prediction. OSODD (Right figure) has extended the OSOD (Left figure) prediction by assigning novel category labels to instances of an unknown class.

Representation Learning. Representation learning aims to learn more discriminative features for input samples. We adapt contrastive learning [33] and utilise objects from both known and working memory to help the network to learn an informative embedding space. The learning is conducted in an unsupervised manner. Following [19], we build a dynamic dictionary to store samples. The network is trained to maximise similarities for positive pairs (an object and its augmented version) while minimising similarity for negative pairs (different object instances) in the embedding space. For an object representation, the contrastive loss is formulated as [8]:

$$\ell_{q,\{k\}} = -\log \frac{\exp(q \cdot k^+/\tau)}{\exp(q \cdot k^+/\tau) + \sum_{k^-} \exp(q \cdot k^-/\tau)} \tag{3}$$

where q is a query object representation, $\{\mathrm{k}\}$ is the queue of key object samples, k^+ is an augmented version of q, known

as the positive key, and k^- is the representations of other samples, known as the negative key. τ is a temperature parameter. On top of the contrastive learning head, we adopt an unsupervised augmentation strategy [26] which replaces all samples with mixed samples. It minimises the vicinal risk [5] which discriminates classes with very different pattern distributions and create more training samples [47]. For each sample in the queue $\{k\}$, we combine it with the query object representation q via linear interpolation and generate a new view $k_{m,i}$. Correspondingly, a new virtual label v_i for the ith mix sample $x_{m,i}$ is defined as:

$$v_i = \begin{cases} 1, & \text{if } q \text{ and } k^+ \text{ are chosen;} \\ 0, & \text{otherwise;} \end{cases} \tag{4}$$

where q and k^+ are the positive sample pairs, the virtual label is assigned to 1 if the mixing pair are from the same object instance.

Novel Category Labelling. Using the encoded representation of the objects, we perform the label assignments using constrained k-means clustering [44], a non-parametric semi-supervised clustering method. The constrained k-means clustering takes object encoding from both known and working memory as its input. It converts the standard k-means clustering into a constraint algorithm by forcing the labelled object representation to be hard-assigned to their ground-truth class. In particular, we treat the object instances from the known memory M_k as the labelled samples. We manually calculate the centroid for each labelled class. These centroids from M_k serve as the first group of initial centroids for the k-means algorithm. We then randomly initialise the rest of the centroids for novel categories using the k-mean++ algorithm [1]. For each iteration, the labelled object instances are assigned to the pre-defined clusters while the unknown object instances from M_w are assigned to the cluster with the minimal distance between the cluster centroids and the object embedding. By doing this, we effectively avoid falsely predicted objects (*i.e.* objects that belong to one of the semantic classes being predicted as unknown) from influencing the centroid update. We run the last cluster assignment step using only the novel centroids to ensure that all unknown objects from working memory are assigned to a discovered visual category in the final prediction. The novel centroids from the algorithm represent the discovered novel categories.

5. Experimental Setup

We provide a comprehensive evaluation protocol for studying the performance of our model in detecting objects from known classes and discovery of new novel categories for objects of unknown classes in our target dataset.

	Task-1	Task-2	Task-3
Semantic Split	VOC Classes	Outdoor,Accessory, Appliance, Truck Wild Animal	Sports, Food
Known/Unknown Class	20/60	40/40	60/20
Training Set	16551	45520	39402
Validation Set	1000		
Test Set	4952		

Table 2. Details of class split for the Benchmark. Task-1, Task-2 and Task-3 have different dataset splits of known and unknown classes.

5.1. Benchmark Dataset

Pascal VOC 2007 [12] contains 10k images with 20 labelled classes. MS-COCO [30] contains around 80k training and 5k validation images with 80 labelled classes. These two object detection datasets are used to build our benchmark. Following the setting of open-world object detection [49], the classes are separated into known and unknown for three tasks $\mathcal{T} = \{T_1, T_2, T_3\}$. For task $T_t \in \mathcal{T}$, all known classes from $\{T_i \mid i < t\}$ are treated as known classes for T_t while the remaining classes are treated as unknown. For the first task T_1, we consider 20 VOC classes as known classes, and the remaining non-overlapping 60 classes in MS-COCO are treated as the unknown classes. New classes are added to the known set in the successive tasks,*i.e.*, T_2 and T_3. For evaluation, we use the validation set from MS-COCO except for 48 images that are incompletely labelled [49]. We summarise the benchmark details in Tab. 2.

5.2. Evaluation Metrics.

Object Detection Metrics. A qualified open-set object detector needs to accurately distinguish unknown objects [11]. UDR (Unknown Detection Recall) [49] is defined as the localisation rate of unknown objects, and UDR (Unknown Detection Precision) [49] is defined as the rate of correct rejection of objects of an unknown class. Let true-positives (TP_u) be the predicted unknown object proposals that have intersection over union IoU > 0.5 with ground truth unknown objects. Half false-negatives (FN_u^*) be the predicted known object proposals that have IoU > 0.5 with ground truth unknown objects. False-negatives(FN_u) is the missed ground truth unknown objects. UDR and UDP are calculated as follow:

$$\begin{aligned} \text{UDR} &= \frac{\text{TP}_\text{u} + \text{FN}_\text{u}^*}{\text{TP}_\text{u} + \text{FN}_\text{u}} \\ \text{UDP} &= \frac{\text{TP}_\text{u}}{\text{TP}_\text{u} + \text{FN}_\text{u}^*} \end{aligned} \tag{5}$$

In our task, the other important aspect is to localise and classify objects of interest from the known classes. We evaluate the closed-set detection performance using the

standard mean average precision (mAP) at IoU threshold of 0.5 [38]. To show the incremental learning ability, we provide the mAP measurement for the newly introduced known classes and previously known classes separately [24, 34].

Category Discovery Metrics. Category discovery can be evaluated using clustering metrics [18, 21, 27, 36, 44, 50]. We adopt the three most commonly used clustering metrics for our object-based category discovery performance. Suppose a predicted proposal of an object of an unknown class has matched to a ground truth unknown object. Let the predicted category label of the object proposal be $\hat{y}_i$, the ground truth label for the object is denoted as y_i. We calculate the clustering accuracy (ACC) [18] by:

$$\text{ACC} = \max_{p \in P_y} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{y_i = p(\hat{y}_i)\} \tag{6}$$

where N is the number of clusters, and P_y is the set of all permutations of the unknown class labels.

Mutual Information $I(X, Y)$ quantifies the correlation between two random variables X and Y. The range of $I(X, Y)$ is from 0 (Independent) to $+\infty$. Normalised mutual information (NMI) [42] is bounded in the range $[0, 1]$. Let Cl be the set of ground truth classes, and $\widehat{Cl}$ be the set of predicted clusters. The NMI is formulated as:

$$\text{NMI} = \frac{I(Cl, \widehat{Cl})}{[H(Cl) + H(\widehat{Cl})]/2} \tag{7}$$

where $I(Cl, \widehat{Cl})$ is the sum of mutual information between each class-cluster pair. $H(Cl)$ and $H(\widehat{Cl})$ compute the entropy using maximum likelihood estimation. The Purity of the clusters is defined as:

$$\text{Purity} = \frac{1}{N} \sum_{i=1}^{N} \max_{k} |Cl_k \cap \widehat{Cl}_i| \tag{8}$$

Here, N is the number of clusters and $\max$ is the highest count of objects for a single class within each cluster.

6. Results and Analysis

6.1. Baselines

Object Detection Baselines. Our framework uses an open-set object detector for known and unknown instance detection. We compare two recent approaches: Faster-RCNN+ [24] and ORE [24]. The Faster-RCNN+ is a popular two-stage object detection method, which is modified from Faster RCNN [38] to localise objects of unknown classes by additionally adapting an unknown-aware regional proposal. ORE uses contrastive clustering and an energy-based classifier to discriminate the representations of known and unknown data. Our generic framework could cooperate with any open-set object detector, hence it is highly flexible.

Category Discovery Baselines. We compare our novel method with three baseline methods, including k-means, FINCH [39] and a modified approach from DTC [18].

K-means clustering is a non-parametric clustering method that minimises within-cluster variances. In every iteration, the algorithm first assigns the data points to the cluster with the minimum pairwise squared deviations between samples and centroids; then, it updates cluster centroids with the current data points belonging to the cluster.

FINCH [39] is a parameter-free clustering method that discovers linking chains in the data by using the first nearest neighbour. The method directly develops the grouping of data without any external parameters. To make a fair comparison, we set the number of clusters to the same as the other baseline methods. We discuss the performance of FINCH in estimating the number of novel classes in Sec. 6.2.1.

DTC+, the DTC method [18] is proposed for NCD problems [16], where the setting assumes the availability of unlabelled data at the training phase. The algorithm modifies deep embedded clustering [46] to learn knowledge of the labelled subset and transfer it to the unlabelled subset. This setting requires the unlabelled data in the training and testing set to be from the same classes. However, no unknown instances are available in training under the open-set detection setting. Hence, the NCD-based approaches, such as DTC cannot be directly applied to our problem. To facilitate the method in our settings, we modify it by transferring a portion of the classes from the known memory to the working memory during training and treating them as additional unknown classes. We evaluate DTC's generalisation performance on our problem in Sec. 6.2.3.

6.2. Experimental Results

We report the quantitative results of the novel category number estimation, object detection and novel category discovery performance in Secs. 6.2.1 to 6.2.3. We show and discuss the qualitative results in Fig. 4 and in the supplementary material.

6.2.1 Novel Category Number Estimation

We show the results of estimating the number of novel categories in Tab. 3. The middle two columns show the automatically discovered grouping by the FINCH algorithm [39]. The numbers are under-estimated by a large margin of 30%, 32.5% and 40% respectively. The last two columns show the result using DTC [18]. It is found that

Task	GT	FINCH [39]	Error	Est. [18]	Error
1	60	42	30%	48	20%
2	40	27	32.5%	31	22.5%
3	20	12	40%	16	20%

Table 3. Result of novel Category estimation.

the estimated number was lower than the ground truth class number, with an average error rate of 21%. By exploring the ground-truth labels in the grouping, we found that both methods tend to ignore object classes with a small number of samples. Compared to the class estimation in the image recognition task [18, 44], the detection task faces more biased datasets as well as fewer available samples. Hence, it is still a challenging task for object category estimation.

6.2.2 Open-Set Object Detection

We compare two baseline models for the object detection part in our framework and show the result in Tab. 4. For each task, we record the mAP of all objects to evaluate the closed-world detection result. UDR and UDP reflect the unknown objectness performance and discrimination performance. The ORE outperforms the modified Faster-RCNN on known classes detection by a smaller margin, which are $-0.14\%, +1.14\%$ and $+1.01\%$ respectively. The mAP scores get lower when new semantic classes are being introduced. The UDR result shows that ORE performs better on unknown object localisation, with a $+0.95\%$ average unknown detection rate. As opposed to closed-set detection, the UDR scores improved when more classes are made available to the model. The Faster-RCNN baseline can only localise objects of an unknown class, but it does not identify them from known classes hence there is no UDP score.

6.2.3 Novel Category Discovery

Results of the object category discovery are shown in Tab. 5 and Tab. 6. The test condition is the same as the open-set detection. Our discovery method is able to accurately explore novel categories among the objects of unknown classes.

Using the estimated number of classes, the discovery results are reported in Tab. 5. We observe that our method outperformed other baseline methods in the first two tasks. In Task-3, where there are 60 known classes and 20 unknown classes, our accuracy and purity score is slightly lower than the FINCH algorithm by 0.8% and 0.1%. We suggest that Task-3 may contain more biased unknown object classes, therefore becoming challenging for self-supervised learning to learn generalised representations.

We report the results using the ground truth number of classes in Tab. 6. The results shown are similar to Tab. 5, where our method has the best-aggregated performance over three tasks. The method achieves respectable quantitative results considering the challenging level of the task.

	Task-1			Task-2			Task-3		
Method	mAP	UDR	UDP	mAP	UDR	UDP	mAP	UDR	UDP
F-RCNN +	-/ **56.16**	**20.14**	-	51.09/ 23.84	21.54	-	35.69/ 11.53	30.01	-
ORE [24]	-/ 56.02	20.10	**36.74**	**52.19/ 25.03**	**22.63**	**21.51**	**37.23/ 12.02**	**31.82**	**23.55**

Table 4. Baseline model comparison for open-set detectors. The mean average precision (mAP) is recorded for the previous/current known objects, there is no previous known for Task-1.

	Task-1			Task-2			Task-3		
Method	NMI	ACC	Purity	NMI	ACC	Purity	NMI	ACC	Purity
K-means	8.5	5.3	9.3	5.0	6.2	12.0	5.3	10.9	27.6
FINCH [39]	2.8	6.0	8.2	5.4	6.3	9.9	5.3	**17.2**	**29.4**
DTC+ [18]	7.5	4.6	5.2	4.0	4.2	7.5	3.9	5.0	25.4
Ours	**11.0**	**6.3**	**12.6**	**5.8**	**6.9**	**13.3**	**6.5**	16.4	29.3

Table 5. Results of discovery with estimated class number (48, 31, 16 for Task-1, Task-2 and Task-3 respectively). The highest score in each column is bold in black, and the second-highest score in each column is bold in grey. Our novel method has outperformed the proposed baseline models for all scores in Task-1 and Task-2. The cluster accuracy and purity scores are the second-highest in Task-3, with a marginal difference to the best-performed baseline.

	Task-1			Task-2			Task-3		
Method	NMI	ACC	Purity	NMI	ACC	Purity	NMI	ACC	Purity
K-means	11.9	6.0	12.4	5.9	6.1	12.8	6.0	11.6	27.9
FINCH [39]	10.3	6.1	12.5	4.8	7.5	13.4	5.5	**13.6**	28.3
DTC+ [18]	8.3	4.7	9.2	4.2	5.0	12.1	5.0	7.7	26.1
Ours	**13.1**	**6.5**	**13.1**	**7.0**	**7.5**	**13.8**	**6.1**	13.2	**29.1**

Table 6. Results of discovery with ground truth class number (60, 40, 20 for Task-1, Task-2 and Task-3 respectively). The highest score in each column is bold in black, and the second-highest score in each column is bold in grey. With the pre-defined number of classes, our method has achieved the highest scores for all three tasks, except for the accuracy in Task-3, which is behind the highest scoring baseline method by a small margin. The overall performance of our method is the best among all the proposed baselines.

6.3. Ablation study

To study the contribution of each component in our proposed framework, we design ablation experiments and show the results in Tab. 7.

Representation Learning. The effects of the representation learning in discovering novel classes are shown in Cases I, II and IV. The clustering result without encoding is reported in Case I. The result with only contrastive learning is reported in Case II. We observe that the performance without encoding is around 10% lower compared to Case IV which is our method. Contrastive learning without the mix-up argumentation reflects higher scores compared to Case I, but it is still around 4% lower in the aggregated scores compared to Case IV. This suggests that representation learning is critical for constructing a strong baseline.

Category Discovery. We evaluate the effects of using

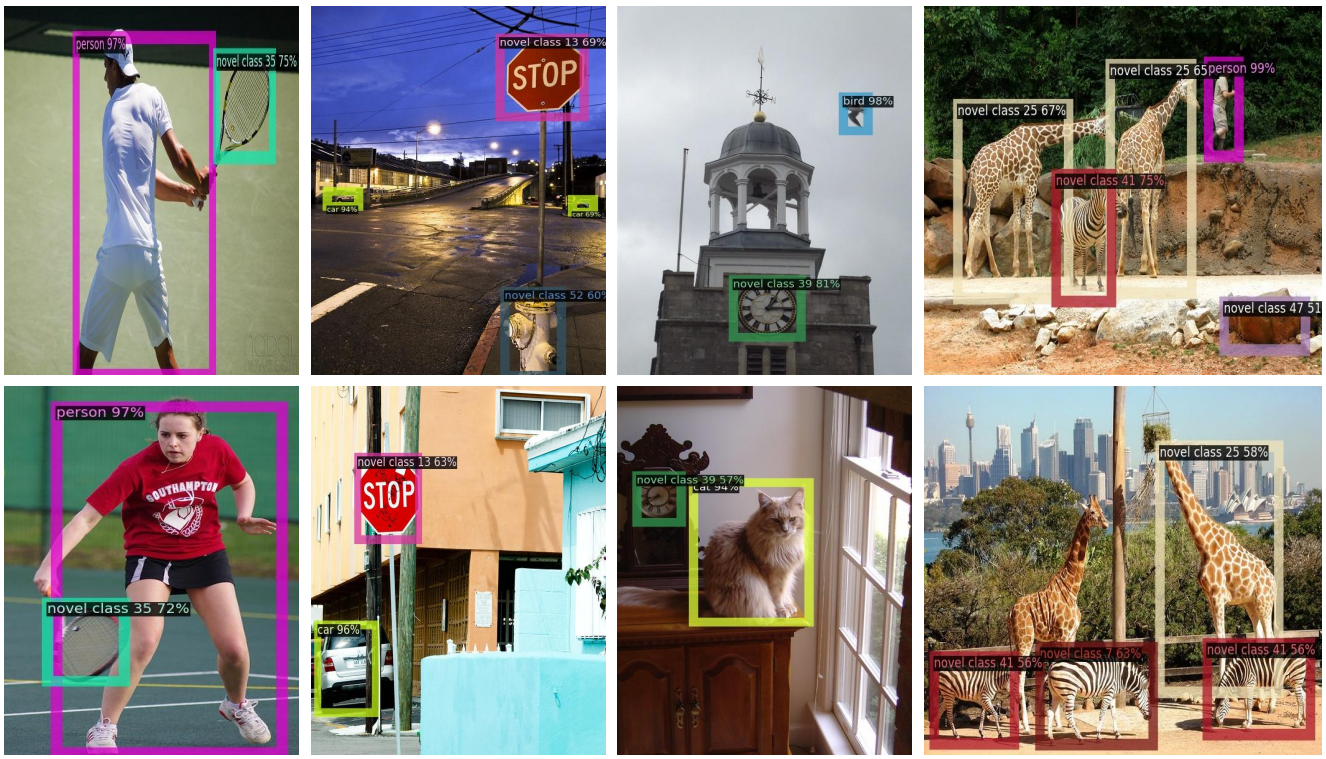

Figure 4. Visualisation of OSODD predictions for Task-1. The *tennis racket*, *stop sign*, *fire hydrant*, *clock*, *giraffe* and *zebra* are the novel classes that have not been introduced at this stage. The same bounding box colour indicates objects that belong to the same class or novel category. The last column demonstrates a failure case where a giraffe is not detected, and one of the zebras is assigned to the wrong visual category. More visualised results are provided in the supplementary material.

	Representation Learning		Category Discovery	Task-1			Task-2			Task-3		
	Mix-Up Augmentation	Contrastive Learning	Semi-supervised Clustering	NMI	ACC	Purity	NMI	ACC	Purity	NMI	ACC	Purity
I	✗	✗	✓	8.9	5.6	10.5	4.7	5.4	11.9	5.5	14.7	27.7
II	✗	✓	✓	10.5	6.3	12.0	5.6	5.4	13.2	6.1	15.5	28.6
III-1	✓	✓	✗	9.6	5.7	11.7	5.2	6.3	12.9	5.8	15.9	28.8
III-2	✓	✓	✗	7.4	6.3	12.3	5.4	6.4	13.1	6.0	**16.8**	28.7
IV	✓	✓	✓	**11.0**	**6.3**	**12.6**	**5.8**	**6.9**	**13.3**	**6.5**	16.4	**29.3**

Table 7. Ablation Study on components of our proposed category discovery method. The complete method with all the proposed modules achieves the best-aggregated performance in all tasks, which shows the importance of each component contributing to the method.

semi-supervised clustering in Case III-1, III-2 and IV. In Case III-1. We make the clustering algorithm fully unsupervised by removing the labelled centroids and instances. The results decrease by around 8% in all tasks. Since the FINCH algorithm [39] shows a competitive result in Tab. 5 and Tab. 6. In Case III-2, we replace the semi-supervised clustering with the FINCH algorithm. The results show that Case IV outperforms Case III-2 in the task aggregation scores, which indicates our model better clusters the samples with the same learned feature space.

Memory Module. To show the effects of the current memory design, we ablate the module by removing the known memory in representation learning. We report the results in the supplementary material.

7. Conclusion

In this work, we propose a framework to detect known objects and discover novel visual categories for unknown objects. We term this task Open-Set Object Detection and Discovery (OSODD), as a natural extension of open-set object detection tasks. We develop a two-stage framework and a novel method for label assignment, outperforming other popular baselines. Compared to detection and discovery tasks, OSODD can provide more comprehensive information for real-world practices. We hope our work will contribute to the object detection community and motivate further research in this area.

References

[1] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006. 5

[2] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1893–1902, 2015. 2

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 1

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[5] Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. *Advances in neural information processing systems*, 13, 2000. 5

[6] Jia Chen, Yasong Chen, Weihao Li, Guoqin Ning, Mingwen Tong, and Adrian Hilton. Channel and spatial attention based deep object co-segmentation. *Knowledge-Based Systems*, 211:106550, 2021. 3

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 4

[9] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1201–1210, 2015. 3

[10] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 3

[11] Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. The overlooked elephant of object detection: Open set. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1021–1030, 2020. 1, 2, 5

[12] 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. 5

[13] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9284–9292, 2021. 3

[14] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 2

[15] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[16] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. *arXiv preprint arXiv:2002.05714*, 2020. 6

[17] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[18] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2, 4, 6, 7

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3, 4

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[21] Jie Hong, Weihao Li, Junlin Han, Jiyang Zheng, Pengfei Fang, Mehrtash Harandi, and Lars Petersson. Goss: Towards generalized open-set semantic segmentation. *arXiv preprint arXiv:2203.12116*, 2022. 6

[22] Lalit P Jain, Walter J Scheirer, and Terrance E Boult. Multi-class open set recognition using probability of inclusion. In *European Conference on Computer Vision*, pages 393–409. Springer, 2014. 2

[23] Xuhui Jia, Kai Han, Yukun Zhu, and Bradley Green. Joint representation learning and novel category discovery on single-and multi-modal data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 610–619, 2021. 3

[24] 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, 2, 3, 6, 7

[25] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 4

[26] Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee. I-mix: A domain-agnostic strategy for contrastive representation learning. *arXiv preprint arXiv:2010.08887*, 2020. 3, 5

[27] Yong Jae Lee and Kristen Grauman. Object-graphs for context-aware category discovery. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2010. 3, 6

[28] Weihao Li, Omid Hosseini Jafari, and Carsten Rother. Deep object co-segmentation. In *Asian Conference on Computer Vision*, pages 638–653. Springer, 2018. 3

[29] Weihao Li, Omid Hosseini Jafari, and Carsten Rother. Localizing common objects using common component activation

map. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 3

[30] 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. 5

[31] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 2

[32] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3243–3249. IEEE, 2018. 1, 2

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3, 4

[34] Can Peng, Kun Zhao, and Brian C Lovell. Faster ilod: Incremental learning for object detectors based on faster rcnn. *Pattern Recognition Letters*, 140:109–115, 2020. 6

[35] Pramuditha Perera, Vlad I Morariu, Rajiv Jain, Varun Manjunatha, Curtis Wigington, Vicente Ordonez, and Vishal M Patel. Generative-discriminative feature representations for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11814–11823, 2020. 2

[36] Sai Saketh Rambhatla, Rama Chellappa, and Abhinav Shrivastava. The pursuit of knowledge: Discovering and localizing novel categories using dual memory. *arXiv preprint arXiv:2105.01652*, 2021. 3, 6

[37] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 1, 3, 6

[39] Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient parameter-free clustering using first neighbor relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2019. 6, 7, 8

[40] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012. 2

[41] Walter J Scheirer, Lalit P Jain, and Terrance E Boult. Probability models for open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2317–2324, 2014. 2

[42] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002. 6

[43] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *NeurIPS*, 2020. 2

[44] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. *arXiv preprint arXiv:2201.02609*, 2022. 3, 5, 6, 7

[45] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*, 2020. 1

[46] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016. 2, 6

[47] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 5

[48] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[49] Xiaowei Zhao, Xianglong Liu, Yifan Shen, Yuqing Ma, Yixuan Qiao, and Duorui Wang. Revisiting open world object detection. *arXiv preprint arXiv:2201.00471*, 2022. 5

[50] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10875, 2021. 6

[51] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020. 3

[52] Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9462–9470, 2021. 3