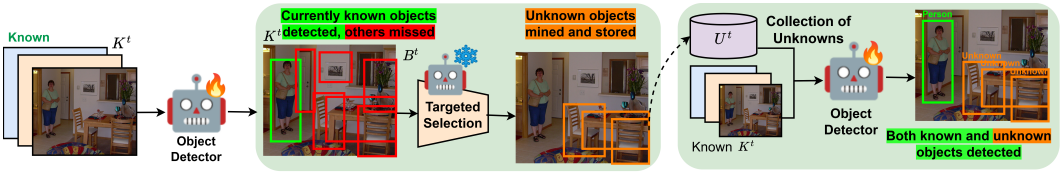
Looking Beyond the Known: Towards a Data Discovery Guided Open-World Object Detection

Anay Majee **Amitesh Gangrade*** **Rishabh Iyer**
The University of Texas at Dallas
firstname.lastname@utdallas.edu

Abstract

Open-World Object Detection (OWOD) enriches traditional object detectors by enabling continual discovery and integration of unknown objects via human guidance. However, existing OWOD approaches frequently suffer from semantic confusion between known and unknown classes, alongside catastrophic forgetting, leading to diminished unknown recall and degraded known-class accuracy. To overcome these challenges, we propose ***C**ombinato**r**ial **O**pen-**W**orld **D**etection* (**CROWD**[2]), a unified framework reformulating unknown object discovery and adaptation as an interwoven combinatorial (set-based) data-discovery (CROWD-Discover) and representation learning (CROWD-Learn) task. CROWD-Discover strategically mines unknown instances by maximizing Submodular Conditional Gain (SCG) functions, selecting representative examples distinctly dissimilar from known objects. Subsequently, CROWD-Learn employs novel combinatorial objectives that jointly disentangle known and unknown representations while maintaining discriminative coherence among known classes, thus mitigating confusion and forgetting. Extensive evaluations on OWOD benchmarks illustrate that CROWD achieves improvements of 2.83% and 2.05% in known-class accuracy on M-OWODB and S-OWODB, respectively, and nearly 2.4× unknown recall compared to leading baselines.



(a) CROWD-Discover (CROWD-D) mines unknown exemplars from low confidence background objects.

(b) CROWD-Learn (CROWD-L) learns distinct embeddings of known and unknown objects

Figure 1: **Overall Architecture of CROWD** showing our novel combinatorial data-discovery guided representation learning approach to (a) identify unknown objects[3] and (b) learn distinguishable representations of both known and unknown objects.

1 Introduction

Object Detection (OD) is central to numerous vision applications [35, 47, 30], but as shown in Figure 1 (left) conventional OD models operate under a closed-world assumption, where the object vocabulary remains fixed throughout deployment. This limitation hinders the model's ability to generalize to novel object categories. *Open-World Object Detection* (OWOD), introduced by Joseph et al. [21], addresses this by combining open-set recognition [63, 74, 65] with incremental learning [37, 12], enabling models to detect unknown objects and subsequently recognize them with minimal supervision, thus supporting continual self-improvement. However, recent efforts [64, 77, 61] reveal

*Work done as a graduate student at UTDallas.

[2]Project Page at https://anaymajee.me/assets/project_pages/crowd.html

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

two enduring challenges: (1) confusion between known and unknown objects [21, 15, 44], and (2) catastrophic forgetting of previously learned classes [61, 64, 45]. Confusion arises due to visual similarity between unknown and known classes (e.g., truck vs. car headlamps), while forgetting stems from the lack of supervision for unknowns, causing them to be misclassified as background. Existing methods struggle to address both issues, as evidenced by low unknown recall and high Wilderness Impact scores [21] (Table 3). These limitations motivate the need for a framework that can effectively discover unknown Region-of-Interests (RoIs) and learn representations that remain distinct from known categories, thereby mitigating confusion and preserving prior knowledge.

We cast Open-World Object Detection (OWOD) as a set-based discovery and learning problem. For each task t, we view the *known* object classes as a collection of sets K^t, group all candidate *unknowns* into a single pseudo-labeled set U^t [21], and treat everything else as *background* B^t. This formulation (Section 3.1) facilitates the incorporation of submodular functions into OWOD and gives rise to our **C**ombinato**r**ial **O**pen-**W**orld **D**etection (**CROWD**) framework. As illustrated in Figure 1, CROWD tackles OWOD as an interleaved process of data discovery (CROWD-D) and representation learning (CROWD-L), directly targeting the dual challenges of confusion and forgetting.

Starting from a OD trained only on known instances K^t (Figure 1[3]), CROWD-Discover (CROWD-D) identifies representative unknowns, formulated as a combinatorial targeted selection problem. Specifically, CROWD-D maximizes the Submodular Conditional Gain (SCG, Section 3.2) between K^t and candidate subsets, encouraging dissimilarity with known and background objects. Authors in [34] provide theoretical guarantees that greedy maximization [52] of SCG results in selection of samples in U^t which are *dissimilar* to K^t as well as B^t, constituting informative pseudo-labeled unknowns.

CROWD-Learn (CROWD-L) then fine-tunes the OD model on both known and mined unknowns as shown in Figure 1(b) via a novel combinatorial joint objective (Equation (1)), rooted in two families of submodular functions: SCG [34] and Total Submodular Information [11] (SIM). Maximizing SCG increases diversity between known and unknown objects, reducing feature overlap and confusion, as corroborated by Table 5. Conversely, minimizing SIM encourages intra-class cohesion within each known object, preserving discriminative features and mitigating forgetting. This formulation closely follows the observation in [48, 49] that submodular functions model cooperation [20] and diversity [38] when minimized and maximized respectively. Finally, we instantiate a family of submodular-based loss functions within CROWD-L that jointly reduce confusion and forgetting, achieving notable gains in unknown recall and known-class accuracy (Table 5). We validate our approach on two standard OWOD benchmarks, M-OWOD [21] and S-OWOD [15], demonstrating its effectiveness across diverse open-world settings. Our main contributions are -

- CROWD introduces a novel combinatorial viewpoint in OWOD by **modeling the identification of unknown instances of a given task as a data discovery problem** (CROWD-D), selecting unknown RoIs which maximize the SCG between and the known object instances.
- CROWD also introduces a novel **set-based learning paradigm CROWD-L, based on SCG functions which minimizes the cluster overlap between embeddings of known and unknown objects** while retaining the discriminative feature information from the known ones.
- Finally, CROWD demonstrates $\sim$2.4$\times$ increase in unknown recall per task alongside up to 2.8% improvement on M-OWODB and 2.1% improvement on S-OWODB in known class performance (measured as mAP) over several existing OWOD baselines.

2 Related Work

Open-World Object Detection (OWOD) first introduced in Joseph et al. [21] augmented a Faster R-CNN [58] model with contrastive clustering and an Energy-Based Unknown Classifier relying on a objectness threshold based pseudo labeling strategy. Subsequent work such as OW-DETR [15] adapted deformable DETR [76] and proposed an attention-based pseudo-labeling scheme that identifies high-activation regions as unknowns without requiring extra supervision. Further, CAT [43] improves transformer-based models by decoupling localization and classification, while introducing dual pseudo-labeling strategies, namely - model-driven and input-driven—to robustly mine unknowns. PROB [77] advanced the state of the art by modeling objectness probabilistically in the embedding

[3]Figure 1(a) shows a subset of background and unknown RoIs for clarity. The total number of RoIs in the original experiment is set to 512 (as in [61]) while the number of mined unknowns is set to 10 (per image).

space using a Gaussian likelihood, allowing better separation of unknowns from background without explicit negative examples. Other notable works include 2B-OCD [68], which integrates a localization-based objectness head [28], OCPL [73] enforces class-prototype separation to reduce known-unknown confusion, and UC-OWOD [69] employs feature-space regularization to suppress background misclassification. Some recent methods leverage external supervision, e.g., MViTs [46], or multimodal cues such as text for class-agnostic detection, these often fall outside strict OWOD assumptions but highlight promising directions for future research. Complementary to these, recent approaches such as RandBox [64] sidesteps detection bias via random bounding box sampling and dynamic-k filtering, while OrthogonalDet [61] enforces angular decorrelation in object features to disentangle objectness and class semantics. Interestingly, Randbox and OrthogonalDet outperforms larger models like OW-DETR, UC-OWOD etc. while using a simpler Faster-RCNN [58] based architecture. Despite substantial progress, OWOD methods continue to grapple with confusion between known and unknown objects and catastrophic forgetting during incremental adaptation recently evidenced in Xi et al. [70], motivating the development of our CROWD framework.In general our work is also related to standard object detection while CROWD-Discover (Section 3.3.1) is related to combinatorial subset selection, the related work for which is provided in Appendix A.2.

3 Method

3.1 Problem Definition: OWOD

We largely adopt the problem formulation of OWOD from Joseph et al. [21] with modifications towards a combinatorial (set-based) formulation. Given an incoming task T_t where $t \in [1, n]$, an object detector $h^t(.;\theta)$ recognizes a set of known classes $K^t = \{K_1^t, K_2^t, \ldots, K_{C^k}^t\}$, $|K^t| = C^k$ while also accounting for unknown classes U^t that may appear during inference (classes in U^t are not labeled during training). Here, K_i^t, $i \in [1, C^k]$ indicates examples for each known class in T_t. The dataset $D^t = \{(x_i^t, y_i^t)\}_{i=1}^M$ for each task T_t, where each label y_i^t contains K object instances (K can very for each image) defined by bounding box parameters $y_k^t = [c_k, x_k, y_k, w_k, h_k]$, with $c_k \in [1, C^k]$ representing the class label. The object detection model $h^t(.;\theta)$ is trained to learn newly introduced instances from labeled examples in K^t while identifying unknown objects U^t by assigning them a placeholder label (0). Examples in U^t can be reviewed by a human expert who identifies C^u new classes, allowing the model to update incrementally and produce $h^{t+1}(.;\theta)$ without retraining on the entire dataset. If $\hat{U}^t$ indicate the newly labeled set of unknown classes s.t. $|\hat{U}^t| = C^u$, then $K^{t+1} = K^t \cup \hat{U}^t$, enabling continual adaptation to new object categories over time.

3.2 Preliminaries: Submodularity

Adopting a set-based formulation allows us to explore combinatorial functions for OWOD. In particular we explore *Submodular functions* which are set functions exhibiting a unique diminishing returns property. Formally, a function $f : 2^{\mathcal{V}} \rightarrow \mathbb{R}$ defined on a ground set $\mathcal{V}$ is submodular if for any subset $A_i, A_j \subseteq \mathcal{V}$, it holds that $f(A_i) + f(A_j) \geq f(A_i \cup A_j) + f(A_i \cap A_j)$ [11]. These functions have been widely studied for applications such as data subset selection [27, 34, 19], active learning [67, 33, 2, 25], and video summarization [24, 26]. Typically, these tasks involve formulating subset selection or summarization as submodular maximization subject to a knapsack constraint. A classic result guarantees a $(1 - e^{-1})$ approximation factor [53] using a greedy algorithm, which can be implemented more efficiently via improved greedy strategies [53, 52]. Within this framework, **Submodular Information Functions** (SIMs) [18, 17, 3], such as Facility-Location or Graph-Cut, promote diversity when maximizing $f(A)$. On the other hand, **Submodular Conditional Gain** (SCG) [17], $H_f(A_i \mid A_j)$, captures elements in A_i most dissimilar to A_j. Extrapolating this, Kothawade et al. [32] defines discovery of unseen, rare examples as a targeted selection problem. Further works [48, 49] have demonstrated the utility of these combinatorial functions in continuous optimization. Majee et al. [48] introduces losses inspired by SIMs to enforce intra-group compactness (when minimized) and inter-group separation (when maximized), while [49] uses SMI-based losses to account for interactions between abundant and rare samples in few-shot learning. *Motivated by these insights, CROWD proposes a novel data-discovery guided representation learning framework to dynamically identify and incrementally adapt to unknown objects*.

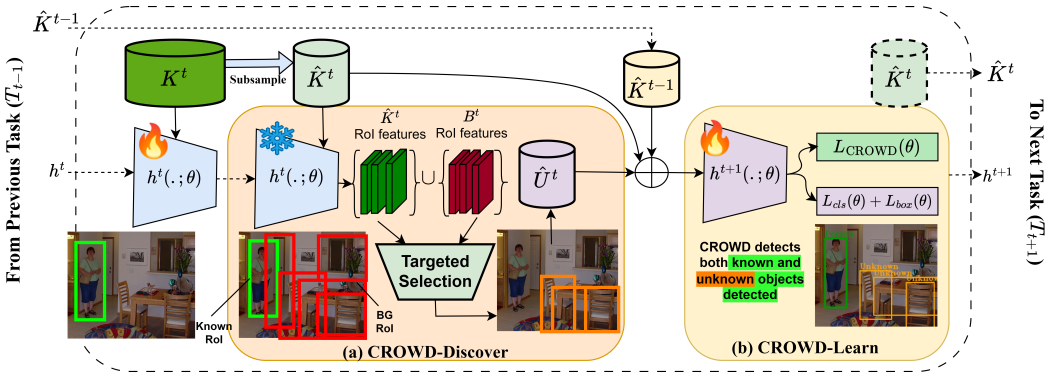


Figure 2: **Interleaved Data-Discovery and Representation Learning in CROWD** on an incoming task T_t. CROWD takes as input the model weights from T_{t-1} and a small replay buffer of previously known classes $\hat{K}^{t-1}$, applies (a) CROWD-Learn to discover unknown RoIs and (b) CROWD-L to learn discriminative features of both known and unknown instances to return an updated model h^{t+1} and the current task replay buffer $\hat{K}^t$.

3.3 The CROWD Framework

The problem formulation in Section 3.1 surfaces two unique challenges in the domain of OWOD - **(1)** How to *identify instances of unlabeled unknown objects* U^t given labeled examples of only known ones in K^t ? **(2)** How to effectively learn representations of currently known objects *without forgetting the previously known* (classes introduced in T_i, where $i < t$) ones ?

To this end, we introduce **C**ombinato**r**ial **O**pen-**W**orld **D**etection (**CROWD**) framework, which models the OWOD task as a interleaved set-based data discovery [32] and representation learning [48] problem. CROWD achieves this in two stages - namely **CROWD-Discover** (a.k.a. CROWD-D) and **CROWD-Learn** (a.k.a. CROWD-L) as shown in Figure 2. Given an incoming task T_t we first train $h^t(.;\theta)$ on currently known classes in D^t. At this point, CROWD-D utilizes the frozen model weights of h^t and uses a small replay buffer (typically containing examples from both previously known and currently introduced objects) to *discover* highly representative proposals of unknown classes U^t. We elucidate this in Section 3.3.1. Subsequently, CROWD-L introduces a novel combinatorial learning strategy to rapidly finetune h^t on this replay buffer (we adopt the predefined buffer in Joseph et al. [21]) to distinguish between known classes K^t and unknown U^t while preserving distinguishable features from the previously known classes. We discuss this in detail in Section 3.3.2.

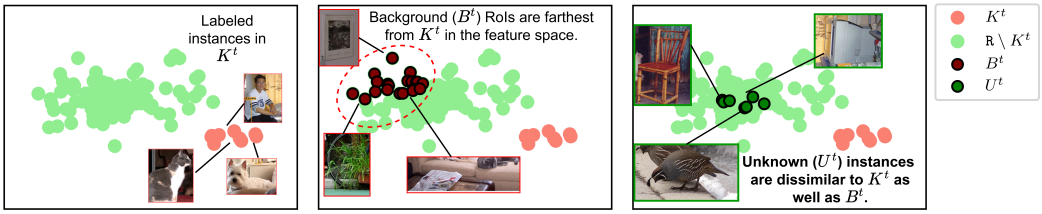
Algorithm 1 Discovering Unknown RoIs in CROWD-D

Require: A task t, set of RoI feature vectors $\mathtt{R} \in \mathbb{R}^{N \times d}$, Objectness scores $\mathbb{S}(.) \in \mathbb{R}^N$, Task specific Labels y_{K^t}, budget $\mathtt{k}$

1: */** Identify and Exclude outliers **/*
2: $\mathtt{R} \leftarrow \{r \in \mathtt{R} | \mathbb{S}(r) \geq \tau_e\}$
3: $K^t \leftarrow \mathtt{HUNGARIAN\text{-}MATCHING}(\mathtt{R}, y_{K^t})$ ▷ *Known class RoI features*
4: $\mathcal{V} \leftarrow \mathtt{R} \setminus K^t$
5: */** Select Background Samples **/*
6: $B^t \leftarrow \underset{\substack{B^t \subseteq \mathcal{V} \\ |B^t| \leq \tau_b \% |\mathcal{V}|}}{\arg\max} H_f(B^t \mid K^t)$ ▷ *Large feature separation from* K^t
7: */** Select Unknown samples from* $\mathtt{R} \setminus K^t$ ***/*
8: $U^t \leftarrow \underset{U^t \subseteq \mathcal{V}, |U^t| \leq \mathtt{k}}{\arg\max} H_f(U^t \mid K^t \cup B^t)$ ▷ *Unknowns are different from* $K^t \cup B^t$
9: **return** U^t

3.3.1 CROWD-Discover

During training of $h^t(.;\theta)$, label information is available only for the currently known classes K^t with no labels of the previously known K^{t-1} and the unknown classes U^t. CROWD-D tackles the challenge of **identifying potentially unknown instances** from Region-of-Interest (RoI) proposals produced by the Region-Proposal-Network (RPN) in $h^t(.;\theta)$. Unlike existing OWOD methods employ pseudo labeling [21], feature orthogonalization [61] etc. rely on the objectness score (probability of an RoI proposal to contain a foreground object), whereas CROWD-D achieves this by **modeling this task as a combinatorial data discovery problem** [32]. *Given a set of RoI proposals*



(a) RoI Feature Space $\mathtt{R}$ (b) Background B^t (c) Unknowns U^t

Figure 3: **Illustration of the data-discovery pipeline in CROWD-D** on a synthetic dataset with $|\mathtt{R}| = 500$ and budget $\mathtt{k} = 10$ and the underlying submodular function as Graph-Cut. CROWD-D selects U^t which are both dissimilar to background B^t and known K^t instances.

$\mathtt{R}$ and a submodular function f we define the data discovery task (Algorithm 1) as a targeted selection problem which selects a set of unknown instances U^t from $\mathcal{V} = \mathtt{R} \setminus K^t$ that maximizes the SCG H_f given a query set comprising of known K^t and the background B^t instances (line 8 in Algorithm 1).

Here, $\mathtt{k}$ denotes a budget for the number of unknown samples mined per image. From the definition of SCG in Section 3.2 **selected examples in U^t are largely dissimilar to examples in $K^t \cup B^t$** indicating that they are neither background objects nor visually similar to known objects as shown in Figure 3(c). Interestingly, the number of known RoIs K^t in $\mathtt{R}$ are significantly fewer than the background RoIs B^t (typically $|\mathtt{R}| = 500$ (total number of RoIs from the RPN) whereas $|K^t| \sim 10$ in MS-COCO [40] which leaves $|B^t| \sim 490$) in most OD models. To minimize the computation costs while selecting U^t from this large RoI pool we exclude all RoIs with low objectness scores $\mathbb{S} < \tau_e$ (line 2 in Algorithm 1) and likely background objects B^t predicted with high confidence (line 6 of Algorithm 1). Instances in B^t are selected which maximize the SCG between themselves and known RoIs K^t under a budget constraint of $\tau_b\%|\mathcal{V}|$ (samples that are significantly different from K^t). Known instances are identified by following the hungarian matching technique applied in Wang et al. [64] as shown in line 3 of Algorithm 1. The exclusion thresholds τ_e and τ_b are empirically determined to be 0.2 and 30% respectively and the underlying submodular function in our experiments is chosen to be Graph-Cut which has been evidenced in Kothawade et al. [34] to model both representation and diversity among selected examples.

3.3.2 CROWD-Learn

Including unknown examples U^t can potentially inject noisy labels into the training data with detrimental effects. We show in Table 5 that CROWD-D alone does not handle the knowledge retention from previously known classes K^{t-1}, despite significant improvements on unknown class recall, causing forgetting. CROWD-L overcomes the aforementioned challenges by introducing a novel **Combinatorial representation learning** strategy inspired from recent developments [48, 49], that ensures orthogonality (separation) between embeddings in K^t and U^t while minimizing the effect of forgetting of $\hat{K}^{t-1}$. Here, $\hat{K}^{t-1}$ denotes a replay buffer of previously known classes.

$$
\begin{aligned}
L^{self}_{\text{CROWD}}(\theta) &= \sum_{i=1}^{C^t} f(K^t_i; \theta)\ ; \\
L^{cross}_{\text{CROWD}}(\theta) &= \sum_{i=1}^{C^t} H_f(K^t_i | U^t; \theta) = \sum_{i=1}^{C^t} f(K^t_i \cup U^t) - f(U^t) \\
L_{\text{CROWD}}(\theta) &= L^{self}_{\text{CROWD}}(\theta) - \eta L^{cross}_{\text{CROWD}}(\theta)
\end{aligned}
\tag{1}
$$

Given a set of known $K^t \cup \hat{K}^{t-1}$, unknown U^t classes alongside a submodular function f we define a learning objective $L_{CROWD}(\theta)$ as shown in Equation (1) *which jointly minimizes the Submodular Total Information (L^{self}_{CROWD}) over each known class $K^t_i \in \{K^t \cup \hat{K}^{t-1}\}$ and the SCG (H_f as defined in Section 3.2) between known class K^t_i and the unknown set U^t (L^{cross}_{CROWD}).* Note that CROWD-L is applied during training of task T_t as a finetuning step as shown in Figure 2(b).

Note that f relies on the pairwise interaction between examples in a batch which we represent using cosine similarity $s_{ku}(\theta) = \frac{h^t(x_k,\theta)^{\mathsf{T}} \cdot h^t(x_u,\theta)}{||h^t(x_k,\theta)|| \cdot ||h^t(x_u,\theta)||}$ and can be different from the one used in CROWD-D decided through ablations in Section 4.2. Our loss formulation in L_{CROWD} follows the observation in [48] which entails that submodular functions model cooperation [20] and diversity [38] when

Table 1: **Summary of various instantiations of CROWD-L** by varying the submodular function f in L^{cross}_{CROWD} and L^{self}_{CROWD}. Here, $\mathcal{T}$ denotes a batch with instances from $K^t_i \cup U^t$.

Objective Name	Instances of L^{cross}_{CROWD}	Instances of L^{self}_{CROWD}
CROWD-GC	$\sum_{i=1}^{C^t} \frac{1}{\lvert\mathcal{T}\rvert}[f(K^t_i;\theta) - 2\lambda\nu \sum_{k\in K^t_i, u\in U^t_i} s_{ku}(\theta)]$	$\sum_{i=1}^{C^t} \frac{1}{\lvert K^t_i\rvert}[\sum_{i\in K^t_i}\sum_{j\in\mathcal{T}\setminus U^t} s_{ij}(\theta) - \lambda \sum_{i,j\in K^t_i} s_{ij}(\theta)]$
CROWD-FL	$\sum_{i=1}^{C^t} \frac{1}{\lvert\mathcal{T}\rvert} \sum_{n\in\mathcal{T}} \max(\max_{k\in K^t_i} s_{nk}(\theta) - \nu \max_{u\in U^t} s_{nu}(\theta), 0)$	$\sum_{i=1}^{C^t} \frac{1}{\lvert K^t_i\rvert} \sum_{i\in\mathcal{T}\setminus K^t_i} \max_{j\in K^t_i} s_{ij}(\theta)$
CROWD-LogDet	$\sum_{i=1}^{C^t} \frac{1}{\lvert\mathcal{T}\rvert} \log\det(s_{K^t_i}(\theta) - \nu^2 s_{K^t_i,U^t}(\theta) s^{-1}_{U^t}(\theta) s_{K^t_i,U^t}(\theta)^{\mathcal{T}})$	$\sum_{i=1}^{C^t} \frac{1}{\lvert K^t_i\rvert} \log\det(s_{K^t_i}(\theta) + \lambda\mathbb{I}_{\lvert K^t_i\rvert})$

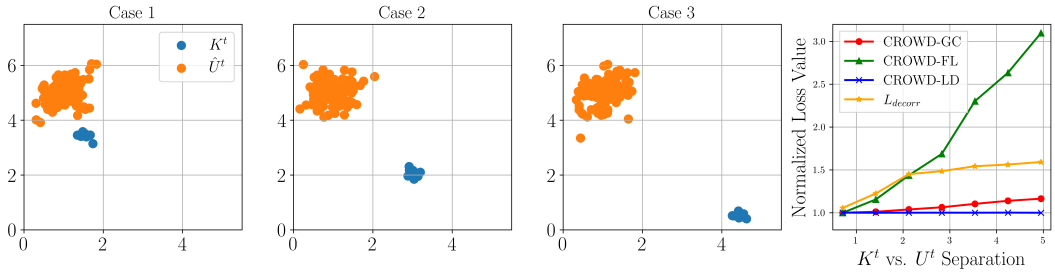


Figure 4: **Characterization of losses in CROWD-L** (L^{cross}_{CROWD}) on a synthetic two-cluster imbalanced dataset by increasing known vs. unknown class separation (cases 1 through 3) similar to the RoI embedding space of $h^t(.;\theta)$. The synthetic dataset generation is performed under the same seed. Here, L^{cross}_{CROWD} is maximized as in Equation (1).

minimized and maximized respectively. By varying the choice of f between popular submodular functions - Facility-Location (FL), Graph-Cut (GC) and Log-Determinant (LogDet) we introduce a family of loss functions summarized in Table 1 and derivations in Appendix A.3. L_{CROWD} is applied to the classification head of $h^t(.;\theta)$ model during all training stages described in Sec. 4. Our novel formulation entails some interesting properties -

(1) L^{self}_{CROWD} retains informative known class features : Following the insights in Jegelka and Bilmes [20] L^{self}_{CROWD} which minimizes the total information contained in K^t_i *encourages intra-class compactness retaining the most discriminative features from the known classes alleviating forgetting* [49].

(2) L^{cross}_{CROWD} models known vs. unknown separation : As shown in Equation (1) L^{cross}_{CROWD} minimizes the SCG (H_f) between the embeddings of the known and unknown classes. As observed in Kothawade et al. [34] maximizing SCG models *dissimilarity* between two sets. In OWOD, L^{cross}_{CROWD} *promotes a large inter-class boundary between K^t_i and U^t, minimizing their cluster overlap resulting in reduced confusion*. Further, a hyper-parameter η controls the trade-off between intra-class compactness modeled by L^{self}_{CROWD} and the inter-class separation in L_{CROWD}.

(3) Sensitivity to unknown classes with varying f : Table 1 highlights the instances of L_{CROWD} by varying f. Such variations injects domain specific properties into CROWD-L critical for the OWOD tasks. As depicted in Figure 4 for L^{cross}_{CROWD}, under varying known vs. unknown class separation, CROWD-FL explicitly models *representation* [34] by adopting the FL based submodular function while CROWD-LogDet injects *diversity* by modeling cluster volume through Log-Determinant [18]. CROWD-GC models both representation and diversity but is not resilient to imbalance between K^t and U^t as CROWD-FL [48]. Thus, CROWD-FL emerges to be a suitable choice for OWOD modeling both representation and resilience to imbalance.

(4) Generalization to Incremental Object Detection (IOD) : As observed in several related works [21, 15, 61, 77], learning from unknown pseudo labels in OWOD benefits Incremental Object Detection (IOD) tasks. However, its important to note that CROWD relies on mined unknowns (from CROWD-D) while no unknown objects are provided in the IOD setting. This requires us to slightly modify our learning objective L^{cross}_{CROWD} to model dissimilarity between currently known K^t and the previously known objects $\hat{K}^{t-1}$ (a replay buffer as in Joseph et al. [21]) instead of unknowns, s.t. $L^{cross}_{\text{CROWD}}(\theta) = -\sum_{i=1}^{C^t} H_f(K^t_i|\hat{K}^{t-1};\theta)$.

Table 2: **Open-world object detection results across incremental tasks.** U-Recall and mAP (%) are reported for various baselines on M-OWOD and S-OWOD benchmarks. Best results are in **bold**.

Method	Task 1		Task 2				Task 3				Task 4		
	U-Recall	mAP	U-Recall	mAP			U-Recall	mAP			mAP		
		Curr.		Prev.	Curr.	Both		Prev.	Curr.	Both	Prev.	Curr.	Both
M-OWOD Benchmark Results													
ORE [21]	4.9	56.0	2.9	52.7	26.0	39.4	3.9	38.2	12.7	29.7	29.6	12.4	25.3
OST [72]	-	56.2	-	53.4	26.5	39.9	-	38.0	12.8	29.6	30.1	13.3	25.9
OW-DETR [15]	7.5	59.2	6.2	53.6	33.5	42.9	5.7	38.3	15.8	30.8	31.4	17.1	27.8
UC-OWOD [69]	-	50.7	-	33.1	30.5	31.8	-	28.8	16.3	24.6	25.6	15.9	23.2
ALLOW [45]	13.6	59.3	10.0	53.2	34.0	45.6	14.3	42.6	26.7	38.0	33.5	21.8	30.6
PROB [77]	19.4	59.5	17.4	55.7	32.2	44.0	19.6	43.0	22.2	36.0	35.7	18.9	31.5
CAT [44]	23.7	60.0	19.1	55.5	32.7	44.1	24.4	42.8	18.7	34.8	34.4	16.6	29.9
RandBox [64]	10.6	**61.8**	6.3	-	-	45.3	7.8	-	-	39.4	-	-	35.4
OrthogonalDet [61]	24.6	61.3	26.3	55.5	38.5	47.0	29.1	46.7	30.6	41.3	42.4	24.3	37.9
CROWD (ours)	**57.9**	61.7	**53.6**	**56.7**	**38.9**	**47.8**	**69.6**	**48.0**	**31.4**	**42.5**	**42.9**	**25.4**	**38.5**
S-OWOD Benchmark Results													
ORE [21]	1.5	61.4	3.9	56.5	26.1	40.6	3.6	38.7	23.7	33.7	33.6	26.3	31.8
OW-DETR [15]	5.7	71.5	6.2	62.8	27.5	43.8	6.9	45.2	24.9	38.5	38.2	28.1	33.1
PROB [77]	17.6	73.4	22.3	66.3	36.0	50.4	24.8	47.8	30.4	42.0	42.6	31.7	39.9
CAT [44]	24.0	**74.2**	23.0	**67.6**	35.5	50.7	24.6	51.2	32.6	45.0	45.4	35.1	42.8
OrthogonalDet [61]	24.6	71.6	27.9	64.0	39.9	51.3	31.9	52.1	**42.2**	**48.8**	48.7	38.8	46.2
CROWD (ours)	**50.5**	**73.5**	**41.7**	64.9	**41.2**	**53.1**	**49.6**	**54.7**	42.1	48.4	**49.8**	**43.0**	**46.4**

Table 3: **Unknown Class Metrics on M-OWODB.** Comparison of U-Recall, WI, and A-OSE across tasks (excluding Task 4 where all classes are known $U^t = \phi$). Best results are in **bold**.

Method	Task 1			Task 2			Task 3		
	U-Recall (↑)	WI (↓)	A-OSE (↓)	U-Recall (↑)	WI (↓)	A-OSE (↓)	U-Recall (↑)	WI (↓)	A-OSE (↓)
ORE [21]	4.9	0.0621	10459	2.9	0.0282	10445	3.9	0.0211	7990
OST [72]	-	0.0417	4889	-	0.0213	2546	-	0.0146	2120
OW-DETR [15]	7.5	0.0571	10240	6.2	0.0278	8441	5.7	0.0156	6803
PROB [77]	19.4	0.0569	5195	17.4	0.0344	6452	19.6	0.0151	2641
RandBox [64]	10.6	0.0240	4498	6.3	0.0078	1880	7.8	0.0054	1452
OrthogonalDet [61]	24.6	**0.0299**	4148	26.3	**0.0099**	1791	29.1	0.0077	1345
CROWD (Ours)	**57.6**	0.0380	**3823**	**53.6**	0.0101	**1508**	**69.6**	**0.0066**	**1266**

4 Experiments

Datasets : We evaluate our approach on two well established benchmarks - **M-OWOD** [21] and **S-OWOD**[15]. M-OWOD, (*Superclass-Mixed OWOD Benchmark*) consists of images from both MS-COCO [40] and PASCAL-VOC [10] depicting 80 classes grouped into 4 tasks (20 classes per task). On the other hand, S-OWOD (*Superclass-Separated OWOD Benchmark*) consists of images from only MS-COCO dataset. Both benchmarks split the underlying data points into four distinct (non-overlapping) tasks T_t, where $t \in [1, 4]$. During training on a task T_t the model in provided labeled examples from T_t alone while at inference the model is expected to identify objects in tasks leading up to T_t, s.t $t \in [1, t]$. No prior knowledge of subsequent tasks $t \in [t + 1, n]$ (n refers to maximum number of tasks in an experiment) are available during training and inference on T_t. In contrast to M-OWODB, S-OWODB introduces a distinct separation between super-categories (eg. animals, vehicles etc.) and distributes these super-categories between tasks (each task will have examples from one or more unique super-categories).

Experimental Setup : Following Sun et al. [61] we adopt a Faster-RCNN [58] based model with a pretrained ResNet-50 [16] backbone. Our model is trained incrementally on 4 tasks as described above with a batch size of 12, an AdamW optimizer, a base learning rate to 2.5×10^{-5} and weight decay of 1×10^{-4}. CROWD-D utilizes the RoI features ($|\mathtt{R}| = 500$) to mine $\mathtt{k} = 10$ unknown instances (determined through ablation study in Appendix A.4.1) per image. The CROWD-L loss is applied across tasks as an additional head and operates on RoI features projected to a 256-dimensional feature space. We train our model on 4 NVIDIA V100 GPUs, provide additional experimental details in Appendix A.4 and release our code at `https://github.com/amajee11us/CROWD.git`.

Metrics : We use mean average precision (mAP) to evaluate known classes, partitioned into previously seen and newly introduced categories. For unknown object class, we follow OWOD conventions [21, 15] and report unknown object recall (U-Recall), as mAP is inapplicable due to incomplete annotations. To measure confusion between known and unknown classes, we report Wilderness Impact (WI) [7] and Absolute Open-Set Error (A-OSE) [51].

Table 5: **Ablation Experiments on the M-OWOD benchmark.** We report the U-Recall and mAP (all known classes) by varying the choice of selection strategies in CROWD-D and learning objectives in CROWD-L. We show that a joint (data discovery + combinatorial loss) strategy provides the best overall performance (denoted as CROWD (joint)).

Method	Baseline	CROWD -D	CROWD -L	Task 1 U-Recall	Task 1 mAP	Task 2 U-Recall	Task 2 mAP	Task 3 U-Recall	Task 3 mAP	Task 4 mAP
OrthogonalDet [61]	✓			$24.6_{\pm 0.04}$	$61.3_{\pm 0.11}$	$26.3_{\pm 0.01}$	$47.0_{\pm 0.06}$	$29.1_{\pm 0.01}$	$41.3_{\pm 0.10}$	$37.9_{\pm 0.09}$
CROWD-D (w/ FLCG)	✓	✓		$50.7_{\pm 0.23}$	$60.3_{\pm 0.07}$	$52.2_{\pm 0.33}$	$45.7_{\pm 0.04}$	$60.1_{\pm 0.18}$	$40.6_{\pm 0.03}$	$38.3_{\pm 0.11}$
CROWD-D (w/ GCCG)	✓	✓		$57.0_{\pm 0.17}$	$61.2_{\pm 0.05}$	$54.1_{\pm 0.72}$	$45.2_{\pm 0.02}$	$69.6_{\pm 0.11}$	$40.8_{\pm 0.01}$	$38.1_{\pm 0.09}$
CROWD-D (w/ LogDetCG)	✓	✓		$56.4_{\pm 0.46}$	$61.2_{\pm 0.10}$	$54.1_{\pm 0.65}$	$44.1_{\pm 0.07}$	$69.1_{\pm 0.26}$	$39.7_{\pm 0.10}$	$37.6_{\pm 0.08}$
CROWD-L (w/ FLCG)	✓		✓	$25.0_{\pm 0.01}$	$61.7_{\pm 0.02}$	$26.8_{\pm 0.03}$	$47.7_{\pm 0.16}$	$28.8_{\pm 0.30}$	$42.4_{\pm 0.11}$	$38.5_{\pm 0.06}$
CROWD-L (w/ GCCG)	✓		✓	$24.3_{\pm 0.03}$	$61.3_{\pm 0.12}$	$27.1_{\pm 0.10}$	$47.4_{\pm 0.26}$	$31.0_{\pm 0.44}$	$40.2_{\pm 0.11}$	$38.2_{\pm 0.10}$
CROWD-L (w/ LogDetCG)	✓		✓	$22.7_{\pm 0.01}$	$59.5_{\pm 0.09}$	$27.0_{\pm 0.14}$	$44.6_{\pm 0.22}$	$27.2_{\pm 0.21}$	$38.3_{\pm 0.14}$	$36.0_{\pm 0.27}$
CROWD (joint)	✓	✓	✓	$57.9_{\pm 0.33}$	$61.7_{\pm 0.02}$	$53.6_{\pm 0.41}$	$47.8_{\pm 0.02}$	$69.6_{\pm 0.26}$	$31.4_{\pm 0.03}$	$38.5_{\pm 0.07}$

4.1 Results on Benchmark OWOD and IOD tasks

OWOD: We compare the performance of CROWD against several existing baselines on M-OWOD and S-OWOD benchmarks as shown in Table 2. Note, that we follow Sun et al. [61] and report our results on the same seed and compute settings for fair comparisons. CROWD surpasses the latest baseline OrthogonalDet [61] by up to 2.8% and 2.1% on M-OWOD and S-OWOD benchmarks while achieving up to 2.4× gains in U-recall. For approaches like PROB [77], CAT [44] which adopt selection strategies to mine unknowns our combinatorial approach achieves up to 8.4% (on M-OWOD) improvements. This can be attributed to the contributions of CROWD-D which mines representative unknown examples effectively increasing the coverage on such objects. Also, we observe $\sim 3\%$ increase in mAP for previously known classes indicating a reduction in forgetting. The competitive results on the currently known classes (Curr. in Table 2) indicates that $h^t(.;\theta)$ enforces a stronger decision boundary between K^t and U^t through L^{cross}_{CROWD} while retaining performance on K^t through L^{self}_{CROWD}. Additionally, in Table 3 we show that CROWD achieves lesser confusion over existing baselines while boosting U-Recall establishing the importance of modeling OWOD as a combinatorial data-discovery problem. This is further highlighted qualitatively in Figure 5.

Table 4: **Results of CROWD on PASCAL VOC for three IOD tasks** shown in terms of Prev., Curr., and overall mAP.

	Prev.	Curr.	mAP
10 + 10 setting			
ILOD [60]	63.2	63.2	63.2
Faster ILOD [55]	69.8	54.5	62.1
PROB [77]	66.0	67.2	66.5
OrthogonalDet [61]	69.4	71.8	67.0
CROWD (ours)	**73.5**	**75.1**	**72.0**
15 + 5 setting			
ILOD [60]	68.3	58.4	65.8
Faster ILOD [55]	71.6	56.9	67.9
PROB [77]	73.2	60.8	70.1
OrthogonalDet [61]	74.5	66.9	72.6
CROWD (ours)	**76.2**	**68.9**	**74.4**
19 + 1 setting			
ILOD [60]	68.5	62.7	68.2
Faster ILOD [55]	68.9	61.1	68.5
PROB [77]	73.9	48.5	72.6
OrthogonalDet [61]	73.5	74.5	73.6
CROWD (ours)	**74.2**	**75.3**	**74.2**

IOD: Our novel loss formulation described in Section 3.3.2 (point 4) is applied to the finetuning stage of IOD across three popular task splits from the PASCAL-VOC [10] dataset. Note, that for IOD we do not apply CROWD-D due to absence of unknown examples. Our results summarized in Table 4 and detailed in Table 9 (Appendix) shows up to 5.9% boost in overall mAP showing *better generalization to IOD* tasks while *minimizing the impact of forgetting via stronger retention of previously known classes*, a very common pitfall in IOD.

4.2 Ablations

We conduct ablations on the M-OWOD benchmark to analyze the contributions of individual components of CROWD. On top of the baseline method OrthogonalDet [61] we first introduce instances of CROWD-D to assess the impact of data-discovery under a fixed budget $\mathtt{k} = 10$. Next, we decouple CROWD-D and introduce our novel learning objectives in CROWD-L to assess their impact on forgetting and confusion as discussed in Section 3.3.2. For each of the above steps we ablate among instances of f - Graph-Cut (GC), Log-Determinant (LogDet) and Facility-Location (FL). Finally, we combine the best performing instances from CROWD-D and CROWD-L into a *joint* formulation (referred to as CROWD (joint)) as shown in Table 5 which *achieves the best*

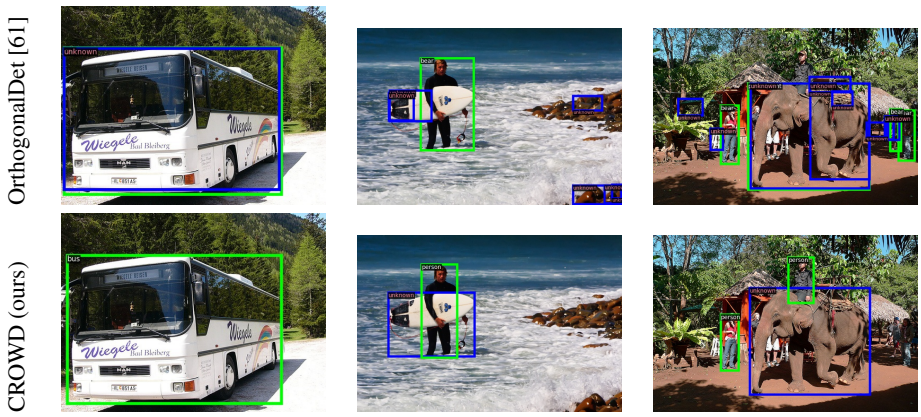


Figure 5: **Qualitative results from CROWD** contrasted against OrthogonalDet [61] showing that our approach mitigates (a) confusion (b) generalizes to unknowns and (c) reduces forgetting.

Table 6: **Ablation Experiments on the exclusion criterion τ_e and background budget τ_b in CROWD-D.** The submodular function is kept constant as Graph-Cut in CROWD-D and a CROWD-FL based learning objective is chosen in CROWD-L under constant machine seed.

Method	Value	Task 1		Task 2		Task 3		Task 4
		U-Recall	mAP	U-Recall	mAP	U-Recall	mAP	mAP
	0.05	57.5	61.7	52.7	47.1	68.2	41.5	38.0
CROWD (τ_e)	**0.2**	57.9	61.7	53.6	47.8	69.6	42.4	38.5
$\tau_b = 30\%$	0.5	53.3	59.5	49.4	45.9	65.0	40.1	37.8
	10%	57.9	61.7	53.1	46.8	69.2	41.3	37.4
CROWD (τ_b)	**30%**	57.9	61.7	53.6	47.8	69.6	42.4	38.5
$\tau_e = 0.2$	50%	55.4	60.0	50.0	42.1	63.7	38.9	34.5

overall performance, balancing the tradeoff between boosting currently known class performance and retaining performance on previously learnt ones.

Impact of Data-Discovery in CROWD-D : As shown in Table 5, irrespective of the choice of f, CROWD-D boosts the U-Recall over the baseline by introducing additional information in the form of pseudo labeled unknowns. We observe that CROWD-D (w/GCCG) (f here is Graph-Cut) provides the best gains in U-Recall up to $2\times$ over the latest baseline OrthogonalDet. This follows the observation in Kothawade et al. [34] which shows that greedy maximization of GCCG models relevance (examples which are dissimilar to both K^t and U^t) while others model diversity (CROWD-D w/LogDet) and representation (CROWD-D w/Facility-Location). Thus, we adopt GCCG based selection strategy in Algorithm 1 for our experiments in Table 2.

Impact of `k` : As stated in Section 3.3.1, `k` controls the number of potential unknown RoIs identified by CROWD-D per image. We ablate among several plausible values of $\texttt{k} \in [0, 100]$ and summarize the results in Table 10 of the Appendix. Increasing the number of identified unknowns from 0 (OrthogonalDet) to 10 shows an increase in performance of the underlying model (U-Recall) while the performance does not increase beyond 20. The increase in U-Recall can be attributed to inclusion of informative RoIs in the training loop. In fact, the mAP on known classes slightly drops below existing baselines for $\texttt{k} = 100$ due to inclusion of spurious background RoIs in the training pipeline.

Impact of Combinatorial Objectives in CROWD-L : Similar to CROWD-D we ablate on variations of f to contrast between formulations summarized in Table 1. As shown in Table 5 our learning formulation, particularly CROWD-FL (based on Facility-Location) demonstrates better retention of previously known class performance while achieving competitive results on latest baseline OrthogonalDet. This follows the observation in Majee et al. [48] which demonstrates that FL based objectives model representation, retaining the most discriminative features through L^{self}_{CROWD} while enforcing sufficient inter-cluster boundary between known and unknown RoI features (L^{cross}_{CROWD}). This also re-establishes the properties described in Figure 4 wherein CROWD-FL shows larger sensitivity to inter-cluster separation as compared to CROWD-GC, CROWD-LogDet and L_{decorr} introduced in OrthogonalDet.

Table 7: **Ablation Experiments on the variation in η in CROWD-L.** Given the Graph-Cut based selection strategy in CROWD-D we vary η between [0.5, 1.0, 1.5] and adopt the best performing value for our pipeline in CROWD-L. The selection budget k in CROWD-D was set to 10 for all experiments and a fixed seed value.

Method	η	Task 1		Task 2				Task 3				Task 4		
		U-Recall	mAP	U-Recall	mAP			U-Recall	mAP			mAP		
			Curr.		Prev.	Curr.	Both		Prev.	Curr.	Both	Prev.	Curr.	Both
OrthogonalDet [61]	-	24.6	61.3	26.3	55.5	38.5	47.0	29.1	46.7	30.6	41.3	42.4	24.3	37.9
CROWD (ours)	0.5	57.8	58.8	53.4	57.1	32.8	44.9	65.3	50.2	25.9	42.1	44.0	21.1	38.3
	1.0	**57.9**	**61.7**	**53.6**	**56.7**	**38.9**	**47.8**	**69.6**	**48.0**	**31.4**	**42.5**	**42.9**	**25.4**	**38.5**
	1.5	57.9	61.7	53.6	55.6	39.1	47.4	69.5	44.0	34.6	40.9	44.0	21.1	38.3

Ablation on Exclusion Criterion τ_e and τ_b in CROWD-D - At first, τ_e is an exclusion threshold which reduces the search space of CROWD-D by eliminating RoIs which have a low confidence threshold. As shown in Table 6, increasing τ_e from 0 to 1 increases performance until $\tau_e = 0.2$ and then reduces. A lower value of τ_e allows for a large search space but includes a lot of noisy background objects leading to reduced selection performance. On the other hand a large value of τ_e can potentially earmark unknown foregrounds as unknowns resulting in reduced performance.

Keeping τ_e fixed at 0.2 we ablate τ_b which controls the selection budget for backgrounds (higher the value more are the number of background RoIs identified). Increasing τ_b (percentage here) increases the fraction of RoIs treated as backgrounds. This widens the search space for the combinatorial function causing a small drop in performance due to confusions between true backgrounds and foreground unknowns. On the other hand, very large values of τ_b shrink the search space oftentimes considering unknown foregrounds as background objects showing a steep drop in performance.

Ablation on Trade-off between $L^{self}_{\textbf{CROWD}}$ and $L^{cross}_{\textbf{CROWD}}$ in CROWD-L - The hyper-parameter η controls the trade-off between known-unknown class separation and known class cluster compactness discussed in Table 7. A lower value of η does not enforce separation between currently known and unknown exemplars but enforces intra-class compactness. This results in better retention of previously known objects but a drop in currently known objects due to increased confusion with unknown exemplars. On the other hand for a large value of η the model enforces large separation between currently known and unknown objects boosting performance on the currently knowns but suffers from catastrophic forgetting of the previously known classes.

5 Conclusion, Limitations and Future Work

We introduced CROWD, a novel combinatorial framework in OWOD, which reformulates OWOD as interleaved set-based discovery (CROWD-D) and representation learning (CROWD-L) tasks. Leveraging Submodular Conditional Gain (SCG) functions, CROWD-D strategically selects representative unknown instances distinctly dissimilar from known objects while CROWD-L consumes mined unknowns to preserve discriminative coherence over known classes. Our evaluations confirm that CROWD effectively addresses known vs. unknown class confusion and forgetting, achieving significant improvements in unknown recall and known-class accuracy on standard OWOD and IOD benchmarks. Despite improvements in U-Recall, operating under a fixed budget CROWD-D injects some spurious exemplars into the selected unknown pool (particularly in images with no known unknown objects). We aim to address this in future works by exploring alternative combinatorial formulations beyond SCG, and introducing stricter constraints in CROWD-D.

Acknowledgements

We gratefully thank anonymous reviewers for their valuable comments. We would also like to extend our gratitude to our fellow researchers from the CARAML lab at UT Dallas for their suggestions. This work is supported by the National Science Foundation under Grant Numbers IIS-2106937, a gift from Google Research, an Amazon Research Award, and the Adobe Data Science Research award. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, Google or Adobe.

References

[1] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[2] Nathan Beck, Durga Sivasubramanian, Apurva Dani, Ganesh Ramakrishnan, and Rishabh K. Iyer. Effective evaluation of deep active learning on image classification tasks. *ArXiv*, abs/2106.15324, 2021.

[3] Jeff Bilmes. Submodularity in machine learning and artificial intelligence. *arXiv preprint arXiv:2202.00132*, 2022.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, 2020.

[5] Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations*, 2020.

[6] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1601–1610, 2021.

[7] 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 (WACV)*, March 2020.

[8] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6568–6577, 2019.

[9] Melanie Ducoffe and Frederic Precioso. Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841*, 2018.

[10] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, January 2015.

[11] Satoru Fujishige. *Submodular Functions and Optimization*, volume 58. Elsevier, 2005.

[12] Takuma Fukuda, Hiroshi Kera, and Kazuhiko Kawamoto. Adapter merging with centroid prototype mapping for scalable class-incremental learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 4884–4893, 2025.

[13] Ross Girshick. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR '14, page 580–587, 2014.

[15] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. OW-DETR: Open-world detection transformer. In *CVPR*, 2022.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[17] Rishabh Iyer, Ninad Khargoankar, Jeff Bilmes, and Himanshu Asanani. Submodular combinatorial information measures with applications in machine learning. In *Algorithmic Learning Theory*, pages 722–754. PMLR, 2021.

[18] Rishabh Iyer, Ninad Khargonkar, Jeff Bilmes, and Himanshu Asnani. Generalized submodular information measures: Theoretical properties, examples, optimization algorithms, and applications. *IEEE Transactions on Information Theory*, 68(2):752–781, 2022.

[19] Eeshaan Jain, Tushar Nandy, Gaurav Aggarwal, Ashish V. Tendulkar, Rishabh K Iyer, and Abir De. Efficient data subset selection to generalize training across models: Transductive and inductive networks. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[20] Stefanie Jegelka and Jeff Bilmes. Submodularity beyond submodular energies: Coupling edges in graph cuts. In *CVPR 2011*, 2011.

[21] K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 5826–5836, jun 2021.

[22] K. J. Joseph, Jathushan Rajasegaran, Salman Khan, Fahad Shahbaz Khan, and Vineeth N. Balasubramanian. Incremental object detection via meta-learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9209–9216, 2022.

[23] Athresh Karanam, Krishnateja Killamsetty, Harsha Kokel, and Rishabh Iyer. ORIENT: Submodular mutual information measures for data subset selection under distribution shift. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[24] V. Kaushal, R. Iyer, K. Doctor, A. Sahoo, P. Dubal, S. Kothawade, R. Mahadev, K. Dargan, and G. Ramakrishnan. Demystifying multi-faceted video summarization: Tradeoff between diversity, representation, coverage and importance. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 452–461, 2019.

[25] Vishal Kaushal, Rishabh Iyer, Suraj Kothawade, Rohan Mahadev, Khoshrav Doctor, and Ganesh Ramakrishnan. Learning from less data: A unified data subset selection and active learning framework for computer vision. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[26] Vishal Kaushal, Sandeep Subramanian, Suraj Kothawade, Rishabh Iyer, and Ganesh Ramakrishnan. A framework towards domain specific video summarization. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 666–675. IEEE, 2019.

[27] Krishnateja Killamsetty, Guttu Sai Abhishek, Aakriti, Ganesh Ramakrishnan, Alexandre V. Evfimievski, Lucian Popa, and Rishabh Iyer. AUTOMATA: gradient based data subset selection for compute-efficient hyper-parameter tuning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2024.

[28] Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *IEEE Robotics and Automation Letters (RA-L)*, 2022.

[29] Jeremias Knoblauch, Hisham Husain, and Tom Diethe. Optimal continual learning has perfect memory and is NP-Hard. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, 2020.

[30] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[31] Suraj Kothawade, Nathan Beck, Krishnateja Killamsetty, and Rishabh Iyer. SIMILAR: Submodular information measures based active learning in realistic scenarios. *Advances in Neural Information Processing Systems*, 34, 2021.

[32] Suraj Kothawade, Shivang Chopra, Saikat Ghosh, and Rishabh Iyer. Active data discovery: Mining unknown data using submodular information measures, 2022.

[33] Suraj Kothawade, Saikat Ghosh, Sumit Shekhar, Yu Xiang, and Rishabh K. Iyer. Talisman: Targeted active learning for object detection with rare classes and slices using submodular mutual information. In *Computer Vision - ECCV 2022 - 17th European Conference*, 2022.

[34] Suraj Kothawade, Vishal Kaushal, Ganesh Ramakrishnan, Jeff A. Bilmes, and Rishabh K. Iyer. PRISM: A rich class of parameterized submodular information measures for guided data subset selection. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI*, pages 10238–10246, 2022.

[35] Suraj Kothawade, Atharv Savarkar, Venkat Iyer, Ganesh Ramakrishnan, and Rishabh Iyer. Clinical: Targeted active learning for imbalanced medical image classification. In *Medical Image Learning with Limited and Noisy Data: First International Workshop, MILLanD 2022, Held in Conjunction with MICCAI*, 2022.

[36] Changbin Li, Suraj Kothawade, Feng Chen, and Rishabh Iyer. PLATINUM: Semi-supervised model agnostic meta-learning using submodular mutual information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12826–12842. PMLR, 17–23 Jul 2022.

[37] Linhao Li, Yongzhang Tan, Siyuan Yang, Hao Cheng, Yongfeng Dong, and Liang Yang. Adaptive decision boundary for few-shot class-incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18359–18367, 2025.

[38] Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

[39] Hui Lin and Jeff Bilmes. Learning mixtures of submodular shells with application to document summarization. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, 2012.

[40] 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 *Computer Vision – ECCV 2014*, Cham, 2014. Springer International Publishing.

[41] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017.

[42] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016.

[43] Shuailei Ma, Yuefeng Wang, Jiaqi Fan, Ying yu Wei, Thomas H. Li, Hongli Liu, and Fanbing Lv. Cat: Localization and identification cascade detection transformer for open-world object detection. In *CVPR*, 2023.

[44] Shuailei Ma, Yuefeng Wang, Ying Wei, Jiaqi Fan, Thomas H. Li, Hongli Liu, and Fanbing Lv. Cat: Localization and identification cascade detection transformer for open-world object detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19681–19690, 2023.

[45] Yuqing Ma, Hainan Li, Zhange Zhang, Jinyang Guo, Shanghang Zhang, Ruihao Gong, and Xianglong Liu. Annealing-based label-transfer learning for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11454–11463, 06 2023.

[46] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *European Conference on Computer Vision (ECCV)*. Springer, 2022.

[47] Anay Majee, Kshitij Agrawal, and Anbumani Subramanian. Few-Shot Learning For Road Object Detection. In *AAAI Workshop on Meta-Learning and MetaDL Challenge*, volume 140, pages 115–126, 2021.

[48] Anay Majee, Suraj Nandkishor Kothawade, Krishnateja Killamsetty, and Rishabh K Iyer. SCoRe: Submodular combinatorial representation learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 34327–34349, 2024.

[49] Anay Majee, Ryan Sharp, and Rishabh Iyer. Smile: Leveraging submodular mutual information for robust few-shot object detection. In *European Conference on Computer Vision (ECCV)*, 2024.

[50] Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[51] 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)*, 2018.

[52] Baharan Mirzasoleiman, Ashwinkumar Badanidiyuru, Amin Karbasi, Jan Vondrak, and Andreas Krause. Lazier than lazy greedy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), 2015.

[53] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294, 1978.

[54] Patrik Okanovic, Roger Waleffe, Vasilis Mageirakos, Konstantinos Nikolakakis, Amin Karbasi, Dionysios Kalogerias, Nezihe Merve Gürel, and Theodoros Rekatsinas. Repeated random sampling for minimizing the time-to-accuracy of learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[55] Can Peng, Kun Zhao, and Brian C Lovell. Faster ILOD: Incremental learning for object detectors based on faster rcnn. *Pattern Recognition Letters*, 2020.

[56] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2016.

[57] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2015.

[58] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[59] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018.

[60] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *ICCV*, 2017.

[61] Zhicheng Sun, Jinghan Li, and Yadong Mu. Exploring orthogonality in open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17302–17312, 2024.

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[63] Hongjun Wang, Sagar Vaze, and Kai Han. Dissecting out-of-distribution detection and open-set recognition: A critical analysis of methods and benchmarks. *International Journal of Computer Vision (IJCV)*, 2024.

[64] Yanghao Wang, Zhongqi Yue, Xian-Sheng Hua, and Hanwang Zhang. Random boxes are open-world object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6233–6243, 2023.

[65] Yu Wang, Junxian Mu, Pengfei Zhu, and Qinghua Hu. Exploring diverse representations for open set recognition. AAAI, 2024.

[66] Kai Wei, Yuzong Liu, Katrin Kirchhoff, Chris Bartels, and Jeff Bilmes. Submodular subset selection for large-scale speech training data. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[67] Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *ICML*, 2015.

[68] Yan Wu, Xiaowei Zhao, Yuqing Ma, Duorui Wang, and Xianglong Liu. Two-branch objectness-centric open world detection. In *Proceedings of the 3rd International Workshop on Human-Centric Multimedia Analysis*, 2022.

[69] Zhiheng Wu, Yue Lu, Xingyu Chen, Zhengxing Wu, Liwen Kang, and Junzhi Yu. Uc-owod: Unknown-classified open world object detection. In *Computer Vision – ECCV 2022*, page 193–210, 2022.

[70] Xing Xi, Yangyang Huang, Zhijie Zhong, and Ronghua Luo. UMB: Understanding model behavior for open-world object detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[71] Binbin Yang, Xi Deng, Han Shi, Changlin Li, Gengwei Zhang, Hang Xu, Shen Zhao, Liang Lin, and Xiaodan Liang. Continual object detection via prototypical task correlation guided gating mechanism. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9245–9254, 2022.

[72] Shuo Yang, Peize Sun, Yi Jiang, Xiaobo Xia, Ruiheng Zhang, Zehuan Yuan, Changhu Wang, Ping Luo, and Min Xu. Objects in semantic topology. In *The Tenth International Conference on Learning Representations, ICLR*, 2022.

[73] Jinan Yu, Liyan Ma, Zhenglin Li, Yan Peng, and Shaorong Xie. Open-world object detection via discriminative class prototype learning. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 626–630. IEEE, 2022.

[74] Zhenyu Zhang, Guangyao Chen, Yixiong Zou, Yuhua Li, and Ruixuan Li. Learning unknowns from unknowns: Diversified negative prototypes generator for few-shot open-set recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6053–6062, 2024.

[75] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable {detr}: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

[76] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

[77] Orr Zohar, Kuan-Chieh Wang, and Serena Yeung. PROB: Probabilistic objectness for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11444–11453, June 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We call out our contributions in brief in the abstract and enlist them on page 2 at the end of the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Section 5 of the main paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We define CROWD-L and CROWD-D in Section 3.3.1 and Section 3.3.2 respectively. Instances of CROWD-L (novel to CROWD) are derived based on Theorems A.3.1 through A.3.3 with proofs in Sec. A.3 of the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We release the anonymized version of our code at `https://github.com/amajee11us/CROWD.git` which houses the evaluation metrics following the popular Detectron2 framework (`https://github.com/facebookresearch/detectron2`).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We evaluate our work on MS-COCO [40] and PASCAL-VOC [10] which are publicly available.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We detail out the experimental settings in Section 4 with additional details in Appendix A.4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide statistical significance experiments on key components of CROWD contrasted against State-of-the-Art approach OrthogonalDet in Table 5. For every task T_t in the OWOD benchmark we run our experiments on three different randomly generated seed values (kept constant across methods) keeping other parameters constant and report results in Table 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We run all our experiments on 4 NVIDIA V100 GPUs with an approximate training time of 36 hours for all tasks. This is described in Section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: For our research and experiments, we use public datasets and models that have been used in prior published works. Our research does not involve direct interaction with human subjects.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Object detection is extensively used in real world scenario. As a broad impact, the proposed work would be meaningful in scenarios for such real-world scenario (where setting is often open-world) and model has to continually adapt to new tasks.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: Safeguards would fall outside the scope of our research and we mainly focus on subjects that are inanimate. No experiments explicitly use personally identifiable features of humans as a subject.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper relies on several existing benchmarks which are cited and credited in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The key asset as a result of this paper is the code base where we provide execution instructions.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not crowdsource images by ourselves; we use images from publicly available datasets MS-COCO and PASCAL-VOC which has been vetted through.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: There is no crowdsourcing involved in this paper.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

A Appendix

A.1 Notation

Following the problem definition in the main paper we introduce the notations used in Table 8 throughout the paper.

Table 8: Collection of notations used in the paper.

Symbol	Description
t	Task identifier for each OWOD task.
T_t	Each Task in OWOD.
D^t	Training dataset for each task.
$\mathcal{T}$	The Ground set, here refers to the mini-batch at each iteration.
K^t	Complete set of currently known classes in T_t.
K^{t-1}	Complete set of previously known classes in T_t.
U^t	Complete set of unknown classes in T_t.
$\hat{K}^t$	Predefined Replay buffer of currently known classes in T_t.
$\hat{K}^{t-1}$	Predefined Replay buffer of previously known classes in T_t.
$h^t(x, \theta)$	Task specific Object Detector used as feature extractor.
$Clf(.,.)$	Multi-Layer Perceptron as classifier. In our case a two layer network.
θ	Parameters of the feature extractor.
$s_{A,B}(\theta)$	Cross-Similarity between sets $A, B \in \mathcal{V}$.
$s_A(\theta)$	Self-Similarity between samples in set $A \in \mathcal{T}$.
$f(A)$	Submodular Information function over a set A.
$H_f(A\|Q)$	Submodular Conditional Gain function between sets A and Q.
$L_{\text{CROWD}}(\theta)$	Loss value computed over all known and unknown objects.
$L^{self}_{\text{CROWD}}(\theta)$	Combinatorial loss computed over all known classes $K^t_i \in \mathcal{T}$.
$L^{cross}_{\text{CROWD}}(\theta)$	Combinatorial loss computed between known classes K^t_i and unknown classes U^t.

A.2 Additional Related Work

Data subset selection aims at identifying a distinct set of examples from a large pool which accurately captures the properties of the data distribution. This has rendered subset selection to be a natural choice for data-efficient machine learning tasks like Active Learning [31, 50, 59, 9], Continual Learning [29, 1], Data Summarization [31, 34] etc. Traditionally subset selection has been defined as a subsampling technique based on similarity [23], uncertainty [5] etc. or random [54]. Orthogonally, a new line of work based on combinatorial functions, particularly submodular functions [11, 18] have emerged which effectively selects informative subsets by modeling the notions of cooperation, diversity and representation [34]. These functions formulate subset selection as a greedy maximization task [52] based on several information theoretic measures like Total Information, Mutual Information, Conditional Gain etc. (discussed in Section 3.2 of the main paper). Concurrent to their success in vision [34, 31], language [39], speech [66] etc. domains,

subset selection has been used in auxiliary learning mechanisms like meta-learning [36] and data-discovery [32] targeting identification of rare or unseen examples from an unlabeled example pool. CROWD exploits this line of investigation adopting a combinatorial subset selection technique (detailed in Section 3.3.1 to discover unknown objects in the open-world setting).

Object detection (OD) is a fundamental task in computer vision encapsulating both localization and recognition tasks under the same roof. OD methods are traditionally grouped into two principal paradigms: single-stage and two-stage detectors. Single-stage detectors, exemplified by SSD [42], RetinaNet [41], and YOLO [57, 56], CenterNet [8] unify the processes of object localization and classification into a single feed-forward network, enabling real-time performance with relatively low computational overhead. In contrast, two-stage detectors, such as Faster R-CNN [14, 13, 58], adopt a cascaded architecture wherein a Region Proposal Network (RPN) first hypothesizes candidate object regions, followed by a refinement stage that simultaneously predicts the class and precise bounding box of each proposal. CNN based architectures struggles with the long-range dependencies, which is important for understanding the complex spatial relationships between objects at varying scales (perspective views). Transformer based models [4, 75, 6] improve upon this vulnerability by introducing a self-attention [62] mechanism based on an encoder-decoder architecture [4]. While these models achieve impressive performance in closed-world settings (all object categories present during testing are known and predefined in the training data) they under-perform in open-world scenarios when encountering unknown objects unseen during training.

Preliminaries of Submodularity (continued from Section 3.2) As discussed in Section 3.2 of the main paper, submodular functions have been recognized to model notions of cooperation [20], diversity [38], representation [34] and coverage [24]. Following the combinatorial formulation in Section 3.1 of the main paper we define the ground set $\mathcal{V} = \{A_1, A_2, \cdots A_N\}$, s.t. $|\mathcal{V}| = N$ and explore four different categories of submodular information functions in our work, namely -

(1) *Submodular Total Information* (S_f) which measures the total information contained in each set [11], expressed as $S_f(A_1, A_2, \ldots, A_N)$ as in Equation (2). Maximizing S_f over a set A_i models diversity [38] while minimizing S_f models cooperation [20].

$$S_f(A_1, A_2, \ldots, A_N) = \sum_{i=1}^{N} f(A_i) \tag{2}$$

(2) *Submodular Conditional Gain* (H_f) which models the gain in information when a set A_j is added to A_i. H_f models the notion of *dissimilarity* between sets and can be expressed in Equation (3).

$$H_f(A_i|A_j) = f(A_i \cup A_j) - f(A_j)\,, \forall i, j \in |\mathcal{V}| \tag{3}$$

Given a submodular function f (can alternatively be H_f) tasks like selection [19, 27] and summarization [26, 24] have been modeled as a discrete optimization problem to identify a summarized set of examples $A \subseteq \mathcal{V}$ via submodular maximization under a cardinality constraint ($|A| \leq k$), i.e. $\max_{A \subseteq \mathcal{V}, |A| \leq k} f(A)$. This can be fairly approximated with a $(1 - e^{-1})$ constant factor guarantee [53] using greedy optimization techniques [52] as shown in Algorithm 2. Extending the definition of submodular functions to continuous optimization space Majee et al. [48] have proposed a set of novel family of learning objectives which minimize total information and total correlation among sets in D_{train} using continuous optimization techniques like SGD. These objectives have been shown to be significantly more robust to large imbalance demonstrated in real-world tasks like longtail recognition [48] and few-shot learning [49].

Algorithm 2 Greedy Submodular Maximization [53]

Require: Submodular function $f : 2^{\mathcal{V}} \rightarrow \mathbb{R}$, cardinality constraint $\mathtt{k}$
Ensure: Set $A \subseteq \mathcal{V}$ maximizing $f(A)$ under cardinality constraint $\mathtt{k}$
1: $A \leftarrow \emptyset$
2: **for** $j = 1$ to $\mathtt{k}$ **do**
3: $\quad e \leftarrow \arg\max_{v \in \mathcal{V} \setminus A} [f(A \cup \{v\}) - f(A)]$
4: $\quad A \leftarrow A \cup \{e\}$
5: **return** A

A.3 Derivations of Instances of L_{CROWD}

As discussed in Section 3.3.2 of the main paper, varying the choice of Submodular function f in Equation (1) results in several instances of L_{CROWD}. Based on three popular choices of f among

Facility-Location, Graph-Cut and Log-Determinant, we derive the respective formulations of L_{CROWD}. Note, that the derivations of L^{self}_{CROWD} are adapted from Majee et al. [48] and are thus not included below.

A.3.1 Derivation of CROWD-FL

Theorem A.1. *Given a set of known RoIs K^t_i, $i \in [1, C^t]$, a set of unknown RoIs U^t ($\mathcal{T} = K^t \cup U^t$) and the Facility-Location based submodular function f defined over any set A s.t. $f(A) = \sum_{i \in \mathcal{T}} \max_{j \in A} s_{ij}$, we define CROWD-FL learning objective to learn the parameters θ of the model h^t, containing two components L^{self}_{CROWD} and L^{cross}_{CROWD} as shown in Equation* (4)*. Here, s_ij resembles the similarity between samples i and j respectively.*

$$
\begin{aligned}
L^{self}_{CROWD} &= \sum_{i=1}^{C^t} \frac{1}{|K^t_i|} \sum_{i \in \mathcal{T} \setminus K^t_i} \max_{j \in K^t_i} s_{ij}(\theta) \\
L^{cross}_{CROWD}(\theta) &= \sum_{i=1}^{C^t} \frac{1}{|\mathcal{T}|} \sum_{n \in \mathcal{T}} \max(\max_{k \in K^t_i} s_{nk}(\theta) - \nu \max_{u \in U^t} s_{nu}(\theta), 0)
\end{aligned}
\tag{4}
$$

Proof. From the definition of L^{cross}_{CROWD} in Equation (1) we find,

$$
\begin{aligned}
L^{cross}_{\text{CROWD}} &= \sum_{i=1}^{C^t} H_f(K^t_i | U^t) \\
L^{cross}_{\text{CROWD}} &= \sum_{i=1}^{C^t} f(K^t_i \cup U^t) - f(U^t)
\end{aligned}
\tag{5}
$$

Substituting the definition of $f(A)$ over any set from the theorem in the above expression we get -

$$
\begin{aligned}
L^{cross}_{\text{CROWD}} &= \sum_{i=1}^{C^t} H_f(K^t_i | U^t) \\
L^{cross}_{\text{CROWD}} &= \sum_{i=1}^{C^t} \sum_{n \in \mathcal{T}} \max_{k \in K^t_i \cup U^t} s_{nk} - \sum_{n \in \mathcal{T}} \max_{u \in U^t} s_{nu} \\
L^{cross}_{\text{CROWD}} &= \sum_{i=1}^{C^t} \sum_{n \in \mathcal{T}} \max \left(\max_{k \in K^t_i} s_{nk}, \max_{kU^t} s_{nk} \right) - \sum_{n \in \mathcal{T}} \max_{u \in U^t} s_{nu} \\
L^{cross}_{\text{CROWD}} &= \sum_{i=1}^{C^t} \sum_{n \in \mathcal{T}} \max \left(\underbrace{\max_{k \in K^t_i} s_{nk}}_{\text{Term 1}} - \underbrace{\max_{u \in U^t} s_{nu}}_{\text{Term 2}}, 0 \right)
\end{aligned}
\tag{6}
$$

The Term 2 in the above equation controls the degree of separation between K^t_i and U^t. Due to this we introduce a hyper-parameter ν which we can control during model training. Sine ν is a constant it does not affect the submodular properties of L^{cross}_{CROWD}. The final loss formulation, normalized by the size of $\mathcal{T}$ thus becomes -

$$
L^{cross}_{\text{CROWD}} = \sum_{i=1}^{C^t} \frac{1}{|\mathcal{T}|} \sum_{n \in \mathcal{T}} \max \left(\max_{k \in K^t_i} s_{nk} - \nu \max_{u \in U^t} s_{nu}, 0 \right) \tag{7}
$$

Additionally, we do not provide proofs for L^{self}_{CROWD} since this function largely resembles the total information formulation in Majee et al. [48]. □

A.3.2 Derivation of CROWD-GC

Theorem A.2. *Given a set of known RoIs K_i^t, $i \in [1, C^t]$, a set of unknown RoIs U^t ($\mathcal{T} = K^t \cup U^t$) and the Graph-Cut based submodular function f defined over any set A s.t. $f(A) = \sum_{i \in \mathcal{T}} \sum_{j \in A} s_{ij} - \lambda \sum_{i,j \in A} s_{ij}$, we define CROWD-GC learning objective to learn the parameters θ of the model h^t containing two components $L_{CROWD}^{self}(\theta)$ and $L_{CROWD}^{cross}(\theta)$ as shown in Equation* (8)*. Here, s_ij resembles the similarity between samples i and j respectively.*

$$
\begin{aligned}
L_{CROWD}^{self} &= \sum_{i=1}^{C^t} \frac{1}{|K_i^t|} [\sum_{i \in K_i^t} \sum_{j \in \mathcal{T} \setminus U^t} s_{ij}(\theta) - \lambda \sum_{i,j \in K_i^t} s_{ij}(\theta)] \\
L_{CROWD}^{cross}(\theta) &= \sum_{i=1}^{C^t} \frac{1}{|\mathcal{T}|} [f(K_i^t; \theta) - 2\lambda\nu \sum_{k \in K_i^t, u \in U_i^t} s_{ku}(\theta)]
\end{aligned}
\tag{8}
$$

Proof. From the definition of L_{CROWD}^{cross} in Equation (1) we find,

$$
\begin{aligned}
L_{\text{CROWD}}^{cross} &= \sum_{i=1}^{C^t} H_f(K_i^t | U^t) \\
L_{\text{CROWD}}^{cross} &= \sum_{i=1}^{C^t} f(K_i^t \cup U^t) - f(U^t)
\end{aligned}
\tag{9}
$$

Substituting the definition of $f(A)$ over any set from the theorem in the above expression of L_{CROWD}^{cross}we get -

$$
\begin{aligned}
L_{\text{CROWD}}^{cross} &= \sum_{i=1}^{C^t} H_f(K_i^t | U^t) = \sum_{i=1}^{C^t} f(K_i^t \cup U^t) - f(U^t) \\
L_{\text{CROWD}}^{cross} &= \sum_{i=1}^{C^t} \sum_{n \in \mathcal{T}} \sum_{k \in K_i^t \cup U^t} s_{nk} - \lambda \sum_{n,k \in K_i^t \cup U^t} s_{nk} - \sum_{n \in \mathcal{T}} \sum_{u \in U^t} s_{nu} + \lambda \sum_{n,u \in U^t} s_{nu} \\
L_{\text{CROWD}}^{cross} &= \sum_{i=1}^{C^t} \sum_{n \in \mathcal{T}} \sum_{k \in K_i^t} s_{nk} + \sum_{n \in \mathcal{T}} \sum_{u \in U^t} s_{nu} - \lambda \sum_{n,k \in K_i^t \cup U^t} s_{nk} \\
&\quad - \sum_{n \in \mathcal{T}} \sum_{u \in U^t} s_{nu} + \lambda \sum_{n,u \in U^t} s_{nu}
\end{aligned}
\tag{10}
$$

The second term and the fourth term cancels out (same value with opposite signs).

$$
\begin{aligned}
L_{\text{CROWD}}^{cross} &= \sum_{i=1}^{C^t} \sum_{n \in \mathcal{T}} \sum_{k \in K_i^t} s_{nk} - \lambda \left(\sum_{n,k \in K_i^t \cup U^t} s_{nk} + \sum_{n,u \in U^t} s_{nu} \right) \\
L_{\text{CROWD}}^{cross} &= \sum_{i=1}^{C^t} \sum_{n \in \mathcal{T}} \sum_{k \in K_i^t} s_{nk} - \lambda \left(\sum_{n,k \in K_i^t} s_{nk} + 2 \sum_{n,u \in U^t} s_{nu} \right)
\end{aligned}
\tag{11}
$$

Now, rearranging the terms of the equation we get -

$$
\begin{aligned}
L_{\text{CROWD}}^{cross} &= \sum_{i=1}^{C^t} \left(\sum_{n \in \mathcal{T}} \sum_{k \in K_i^t} s_{nk} - \lambda \sum_{n,k \in K_i^t} s_{nk} \right) + 2\lambda \sum_{n,u \in U^t} s_{nu} \\
L_{\text{CROWD}}^{cross} &= \sum_{i=1}^{C^t} \underbrace{f(K_i^t)}_{\text{Term 1}} + \underbrace{2\lambda \sum_{n,u \in U^t} s_{nu}}_{\text{Term 2}}
\end{aligned}
\tag{12}
$$

Similar to CROWD-FL the Term 2 in the above equation controls the degree of separation between K_i^t and U^t. Due to this we introduce a hyper-parameter ν which we can control during model training. Sine ν is a constant it does not affect the submodular properties of L_{CROWD}^{cross}. The final loss formulation, normalized by the size of $\mathcal{T}$ thus becomes -

$$L_{\text{CROWD}}^{cross} = \sum_{i=1}^{C^t} f(K_i^t) + 2\lambda\nu \sum_{n,u \in U^t} s_{nu} \tag{13}$$

Additionally, we do not provide proofs for L_{CROWD}^{self} since this function largely resembles the total information formulation in Majee et al. [48]. □

A.3.3 Derivation of CROWD-LogDet

Theorem A.3. *Given a set of known RoIs K_i^t, $i \in [1, C^t]$, a set of unknown RoIs U^t ($\mathcal{T} = K^t \cup U^t$) and the Log-Determinant based submodular function f defined over any set A s.t. $f(A) = \log\det(s_A)$, we define CROWD-LogDet learning objective which contains two components L_{CROWD}^{self} and L_{CROWD}^{cross} as shown in Equation* (14)*. Here, s_ij resembles the similarity between samples i and j respectively.*

$$\begin{aligned} L_{CROWD}^{self} &= \sum_{i=1}^{C^t} \frac{1}{|K_i^t|} \log\det(s_{K_i^t}(\theta) + \lambda \mathbb{I}_{|K_i^t|}) \\ L_{CROWD}^{cross}(\theta) &= \sum_{i=1}^{C^t} \frac{1}{|\mathcal{T}|} \log\det(s_{K_i^t}(\theta) - \nu^2 s_{K_i^t,U^t}(\theta) s_{U^t}^{-1}(\theta) s_{K_i^t,U^t}(\theta)^T) \end{aligned} \tag{14}$$

Proof. From the definition of L_{CROWD}^{cross} in Equation (1) we find,

$$\begin{aligned} L_{\text{CROWD}}^{cross} &= \sum_{i=1}^{C^t} H_f(K_i^t | U^t) \\ L_{\text{CROWD}}^{cross} &= \sum_{i=1}^{C^t} f(K_i^t \cup U^t) - f(U^t) \end{aligned} \tag{15}$$

Substituting the definition of $f(A)$ over any set from the theorem in the above expression of L_{CROWD}^{cross}we get -

$$\begin{aligned} L_{\text{CROWD}}^{cross} &= \sum_{i=1}^{C^t} H_f(K_i^t | U^t) = \sum_{i=1}^{C^t} f(K_i^t \cup U^t) - f(U^t) \\ L_{\text{CROWD}}^{cross} &= \sum_{i=1}^{C^t} \log\det(s_{K_i^t \cup U^t}) - \log\det(s_{U^t}) \\ &= \sum_{i=1}^{C^t} \log \frac{\det(s_{K_i^t \cup U^t})}{\det(s_{U^t})} \end{aligned} \tag{16}$$

From Schur's complement which states that given two sets A and B $\det(s_{A \cup B}) = \det(s_A).\det(s_{A \cup B} \setminus s_A)$. Replacing the term $\det(s_{K_i^t \cup U^t})$ with the above definition we get -

$$\begin{aligned} L_{\text{CROWD}}^{cross} &= \sum_{i=1}^{C^t} \log \frac{\det(s_{U^t}).\det(s_{K_i^t \cup U^t} \setminus s_{U^t})}{\det(s_{U^t})} \\ &= \sum_{i=1}^{C^t} \log\det(s_{K_i^t \cup U^t} \setminus s_{U^t}) \end{aligned} \tag{17}$$

Table 9: **Generalization performance on Incremental Object Detection (IOD)** where we show that our CROWD approach (here only CROWD-L) when applied to the fintetuning stage of IOD tasks show better generalizability. Best results are in **bold** while new classes introduced in the task are shaded gray.

10 + 10 setting	aero	cycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv	mAP
ILOD [60]	69.9	70.4	69.4	54.3	48	68.7	78.9	68.4	45.5	58.1	59.7	72.7	73.5	73.2	66.3	29.5	63.4	61.6	69.3	62.2	63.2
Faster ILOD [55]	72.8	75.7	71.2	60.5	61.7	70.4	83.3	76.6	53.1	72.3	36.7	70.9	66.8	67.6	66.1	24.7	63.1	48.1	57.1	43.6	62.1
ORE [21]	63.5	70.9	58.9	42.9	34.1	76.2	80.7	76.3	34.1	66.1	56.1	70.4	80.2	72.3	81.8	42.7	71.6	68.1	77.0	67.7	64.5
Meta-ILOD [22]	76.0	74.6	67.5	55.9	57.6	75.1	85.4	77.0	43.7	70.8	60.1	66.4	76.0	72.6	74.6	39.7	64.0	60.2	68.5	60.7	66.3
ROSETTA [71]	74.2	76.2	64.9	54.4	57.4	76.1	84.4	68.8	52.4	67.0	62.9	63.3	79.8	72.8	78.1	40.1	62.3	61.2	72.4	66.8	66.8
OW-DETR[15]	61.8	69.1	67.8	45.8	47.3	78.3	78.4	78.6	36.2	71.5	57.5	75.3	76.2	77.4	79.5	40.1	66.8	66.3	75.6	64.1	65.7
PROB [77]	70.4	75.4	67.3	48.1	55.9	73.5	78.5	75.4	42.8	72.2	64.2	73.8	76.0	74.8	75.3	40.2	66.2	73.3	64.4	64.0	66.5
CAT [44]	76.5	75.7	67.0	51.0	62.4	73.2	82.3	83.7	42.7	64.4	56.8	74.1	75.8	79.2	78.1	39.9	65.1	59.6	78.4	67.4	67.7
OrthogonalDet [61][1]	82.9	80.1	75.8	64.3	60.6	81.5	87.9	54.9	48	82.1	57.7	63.5	80.5	77.6	78.2	38.9	69.8	62.8	76.9	64.2	69.41
CROWD (ours)	84.1	84.5	73.9	60.0	65.1	80.1	89.3	82.7	53.3	77.4	63.4	78.5	80.9	83.4	83.9	46.5	72.6	60.9	77.9	71.5	**73.5**
15 + 5 setting	aero	cycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv	mAP
ILOD [60]	70.5	79.2	68.8	59.1	53.2	75.4	79.4	78.8	46.6	59.4	59.0	75.8	71.8	78.6	69.6	33.7	61.5	63.1	71.7	62.2	65.8
Faster ILOD [55]	66.5	78.1	71.8	54.6	61.4	68.4	82.6	82.7	52.1	74.3	63.1	78.6	80.5	78.4	80.4	36.7	61.7	59.3	67.9	59.1	67.9
ORE [21]	75.4	81.0	67.1	51.9	55.7	77.2	85.6	81.7	46.1	76.2	55.4	76.7	86.2	78.5	82.1	32.8	63.6	54.7	77.7	64.6	68.5
Meta-ILOD [22]	78.4	79.7	66.9	54.8	56.2	77.7	84.6	79.1	47.7	75.0	61.8	74.7	81.6	77.5	80.2	37.8	58.0	54.6	73.0	56.1	67.8
ROSETTA [71]	76.5	77.5	65.1	56.0	60.0	78.3	85.5	78.7	49.5	68.2	67.4	71.2	83.9	75.7	82.0	43.0	60.6	64.1	72.8	67.4	69.2
OW-DETR [15]	77.1	76.5	69.2	51.3	61.3	79.8	84.2	81.0	49.7	79.6	58.1	79.0	83.1	67.8	85.4	33.2	65.1	62.0	73.9	65.0	69.4
PROB [77]	77.9	77.0	77.5	56.7	63.9	75.0	85.5	82.3	50.0	78.5	63.1	75.8	80.0	78.3	77.2	38.4	69.8	57.1	73.7	64.9	70.1
CAT [44]	75.3	81.0	84.4	64.5	56.6	74.4	84.1	86.6	53.0	70.1	72.4	83.4	85.5	81.6	81.0	32.0	58.6	60.7	81.6	63.5	72.2
OrthogonalDet [61][1]	81.8	79.3	71.0	71.0	58.8	62.1	82.6	89.7	79.8	47.0	80.5	61.1	79.9	80.2	81.6	44.2	65.5	71.5	75.6	74.2	72.6
CROWD (ours)	82.8	80.6	72.5	59.6	61.3	83.1	89.3	83	49.2	86.1	62.2	83.7	86	80.3	82.8	46.1	80	63.7	79.5	75.6	**74.4**
19 + 1 setting	aero	cycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv	mAP
ILOD [60]	69.4	79.3	69.5	57.4	45.4	78.4	79.1	80.5	45.7	76.3	64.8	77.2	80.8	77.5	70.1	42.3	67.5	64.4	76.7	62.7	68.2
Faster ILOD [55]	64.2	74.7	73.2	55.5	53.7	70.8	82.9	82.6	51.6	79.7	58.7	78.8	81.8	75.3	77.4	43.1	73.8	61.7	69.8	61.1	68.5
ORE [21]	67.3	76.8	60	48.4	58.8	81.1	86.5	75.8	41.5	79.6	54.6	72.8	85.9	81.7	82.4	44.8	75.8	68.2	75.7	60.1	68.8
Meta-ILOD [22]	78.2	77.5	69.4	55.0	56.0	78.4	84.2	79.2	46.6	79.0	63.2	78.5	82.7	79.1	79.9	44.1	73.2	66.3	76.4	57.6	70.2
ROSETTA [71]	75.3	77.9	65.3	56.2	55.3	79.6	84.6	72.9	49.2	73.7	68.3	71.0	78.9	77.7	80.7	44.0	69.6	68.5	76.1	68.3	69.6
OW-DETR [15]	70.5	77.2	73.8	54.0	55.6	79.0	80.8	80.6	43.2	80.4	53.5	77.5	89.5	82.0	74.7	43.3	71.9	66.6	79.4	62.0	70.2
PROB [77]	80.3	78.9	77.6	59.7	63.7	75.2	86.0	83.9	53.7	82.8	66.5	82.7	80.6	83.8	77.9	48.9	74.5	69.9	77.6	48.5	72.6
CAT [44]	86.0	85.8	78.8	65.3	61.3	71.4	84.8	84.8	52.9	78.4	71.6	82.7	83.8	81.2	80.7	43.7	75.9	58.5	85.2	61.1	73.8
OrthogonalDet [61][1]	81.8	82.6	77.0	56.3	66.0	74.4	88.5	78.7	51.2	84.3	63.1	84.4	81.3	78.8	80.9	46.8	77.9	68.6	74.1	74.5	73.6
CROWD (ours)	81.7	80.3	77.4	57.2	66.8	80.7	87.1	67.9	49.4	87.3	65.6	84.2	85.4	79.9	81.6	48.6	77.0	69.0	82.2	75.3	**74.2**

Following Schur's complement yet again which states that $s_{A\cup B} \setminus s_A = s_B - s_{A,B}^T s_A^{-1} s_{A,B}$, where $s_{A,B}$ refers to the cross-similarities between sets A and B while s_A and s_B represent the corresponding self-similarities and substitute this definition into the aforementioned equation as -

$$L_{\text{CROWD}}^{cross} = \sum_{i=1}^{C^t} \log\det(s_{K_i^t} - s_{K_i^t,U^t} s_{U^t}^{-1} s_{K_i^t,U^t}^T) \tag{18}$$

Normalizing this term with the size of the ground set $|\mathcal{T}|$ and introducing the hyper-parameter ν which trades-off between inter-cluster separation and intra-cluster compactness, we derive the function for L_{CROWD}^{cross} as -

$$L_{\text{CROWD}}^{cross} = \sum_{i=1}^{C^t} \frac{1}{|\mathcal{T}|} \log\det(s_{K_i^t} - \nu^2 s_{K_i^t,U^t} s_{U^t}^{-1} s_{K_i^t,U^t}^T) \tag{19}$$

Similar to previously derived objectives, we do not provide proofs for L_{CROWD}^{self} since this function largely resembles the total information formulation in Majee et al. [48]. □

A.4 Additional Experimental Details

In this section we provide additional experimental details for training our CROWD approach on M-OWOD, S-OWOD and IOD benchmarks discussed in Section 4 of the main paper.

M-OWOD and S-OWOD benchmarks - M-OWOD and S-OWOD benchmarks are created from MS-COCO [40] and split into 4 tasks T_t, where $t \in [1, 4]$ detailed in the "Datasets" section in Section 4. For each task, the model in provided labeled examples from T_t alone while at inference the model is expected to identify objects in tasks leading up to T_t, s.t $t \in [1, t]$. We split the training

[1]This is a reproduction of the results from OrthogonalDet from their public repo.

Table 10: **Ablation Experiments on the variation in k in CROWD-D.** Given the Graph-Cut based selection strategy in Algorithm 1 of CROWD-D and the CROWD-FL based learning objective in CROWD-L we vary k in [0, 100] and adopt the best performing budget for our pipeline.

Method	Budget k	Task 1 U-Recall	Task 1 mAP	Task 2 U-Recall	Task 2 mAP	Task 3 U-Recall	Task 3 mAP	Task 4 mAP
OrthogonalDet [61]	-	24.6	61.3	26.3	47.0	29.1	41.3	37.9
CROWD (ours)	5	51.2	61.0	49.1	45.9	62.7	40.3	37.4
	10	57.9	61.7	53.6	47.8	69.6	42.4	38.5
	30	58.4	61.7	53.5	48.0	70.1	42.4	38.5
	100	57.5	59.3	53.7	44.3	70.9	38.8	32.0

into two splits. In the first stage the model is exposed only to the currently known classes K^t and the learnt model h^t biases on labeled examples in K^t. At the end of the first stage CROWD-D kicks in and selects representative unknowns as described in Section 3.3.1. Lets call in U^t. Next, we store a replay buffer of the currently known objects $\hat{K}^t$, s.t. $\hat{K}^t \subseteq K^t$. Following this, we combine $\hat{K}^t$, K^t and a replay buffer from the previous task $\hat{K}^{t-1}$ into a single dataset to finetune h^t using CROWD-L. As detailed in Section 3.3.2 this ensures known vs. unknown separation wile retaining discriminative features from known classes.

IOD benchmarks - In contrast to OWOD, IOD does not encounter unknowns during model training but experiences heavy catastrophic forgetting on previously known classes K^{t-1}. Following recent benchmarks like Sun et al. [61], Zohar et al. [77], Joseph et al. [21] we evaluate the IOD performance of CROWD on PASCAL-VOC benchmark on three settings produced by varying the number of newly added classes - 10 + 10, 15 + 5, 19 + 1 as shown in Table 9. In the absence of unknowns we do not apply CROWD-D and only rely on CROWD-L applied to the finetuning stage of IOD. Following latest works we adopt a replay based learning technique whichh stores a small subset of the previously known objects $\hat{K}^{t-1}$ in a buffer. $\hat{K}^{t-1}$ combined with the newly introduced classes K^t is used to finetune h^t. This also requires us to slightly modify the formulation of L^{cross}_{CROWD} as detailed in Section 3.3.2. For each setting h^t is trained on a batch size of 12 for 3000 iterations using an AdamW optimizer, a base learning rate to 2.5×10^{-5} and weight decay of 1×10^{-4}.

A.4.1 Ablation on Selection Budget k

As detailed in Section 3.3.1, the parameter k dictates how many candidate unknown RoIs CROWD-D selects per image. We conduct an ablation over several plausible settings of k within the interval $[0, 100]$, and present the outcome in Table 10. Fo this experiment we keep the choice of submodular function f in CROWD-D as Graph-Cut and Facility-Location (CROWD-FL) for CROWD-L following the results of the ablation experiments in Table 5 in the main paper. Notably, raising k from 0 (i.e., OrthogonalDet) to 10 yields a marked uplift in the model's unknown–recall (U-Recall), yet further increases beyond $\texttt{k} = 20$ confer no additional gains. This initial boost in U-Recall stems from the integration of truly informative RoIs into the training loop. However, when k reaches its upper bound of 100, the mean average precision (mAP) on known classes experiences a slight decline relative to existing baselines—a consequence of inadvertently incorporating spurious background proposals.

A.4.2 Results on Synthetic Datasets - CROWD-D

In addition to the illustrations provided in Figure 3 we contrast the selection performance of CROWD-D by varying the underlying submodular function f between Grap-Cut (GC), Facility-Location (FL) and Log-Determinant (LogDet) on synthetic datasets as shown in Figure 6. The use of synthetic datasets provide us with complete control over the embedding space allowing us to pathologically inject imbalance, inter-cluster separation etc. in a compute efficient fashion. Particularly in our experiments we use a two-cluster imbalanced setup mimicking the RoI embedding space in Faster-RCNN [58] model. Similar to Sun et al. [61] the number of known class and unknown class feature vectors are severely imbalanced with total number of RoIs $\texttt{R} = 500$ and the number of knowns $|K^t| = 10$. R and K^t are sampled from a normal distribution with fixed variance values. The LogDet based selection strategy enforces the notion of diversity in the selection mechanism which does not select representative unknowns negatively impacting OWOD performance as shown in Table 5. The

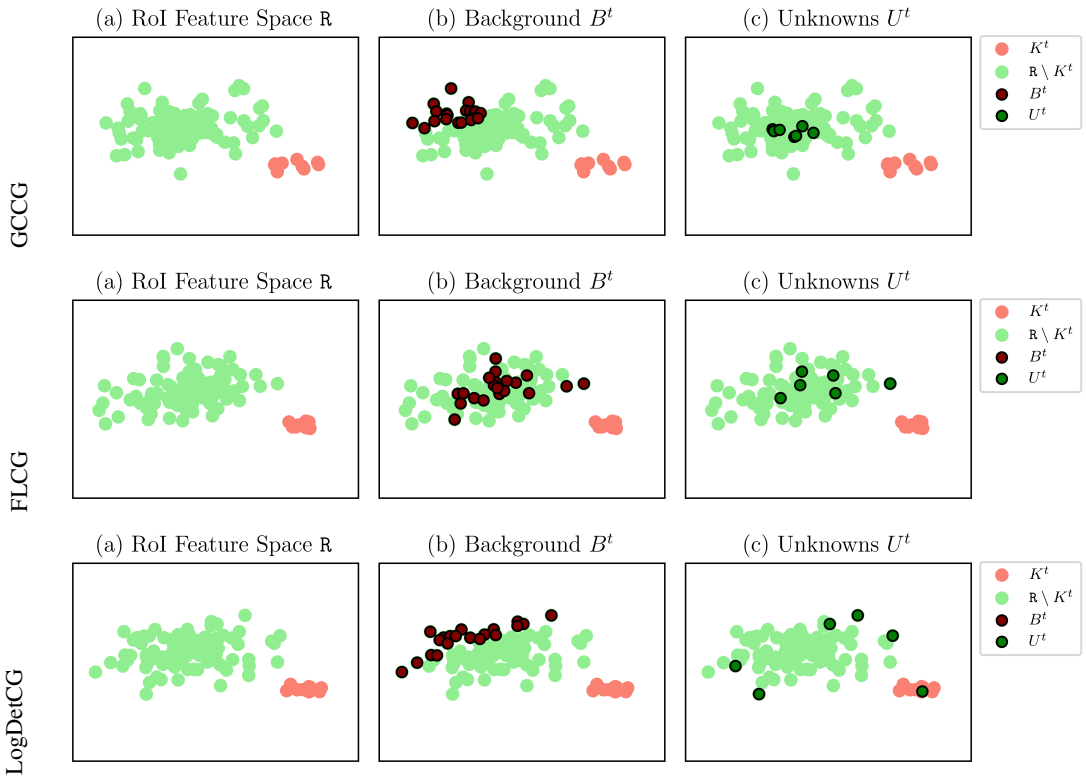


Figure 6: **CROWD-D results on synthetic dataset** contrasted against instances of popular submodular functions - Graph-Cut, Facility-Location and Log-Determinant. Graph-Cut based selection strategy models both representation and diversity resulting in the best possible choice of unknown instances in U^t.

FL based selection strategy models representation as shown in Figure 6 alone during selection resulting in erroneous selection of background instances negatively affecting OWOD performance. Lastly, GC based selection strategy shown in Figure 6 models notions of both diversity and representation selecting diverse backgrounds B^t farthest to K^t as well as representative unknowns U^t. This results in GC based selection strategy to produce the best overall results as shown in Table 2.