Open-Set Object Detection By Aligning Known Class Representations

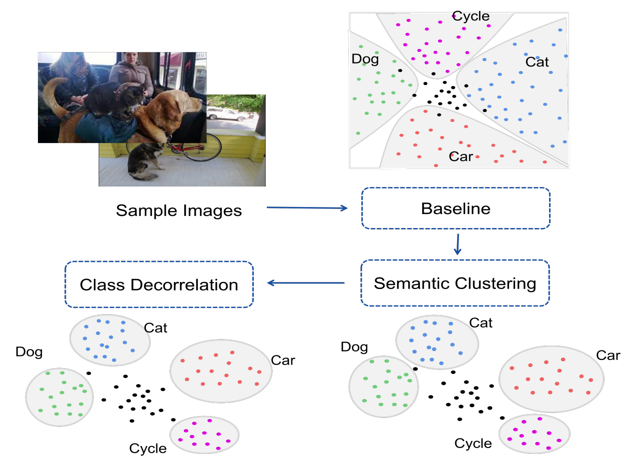
Hiran Sarkar[1] Vishal Chudasama[1] Naoyuki Onoe[1]
Pankaj Wasnik[1*] Vineeth N Balasubramanian[2]
[1]Sony Research India [2]Indian Institute of Technology Hyderabad
{hiran.sarkar,vishal.chudasama1,naoyuki.onoe,pankaj.wasnik}@sony.com, vineethnb@cse.iith.ac.in

Abstract

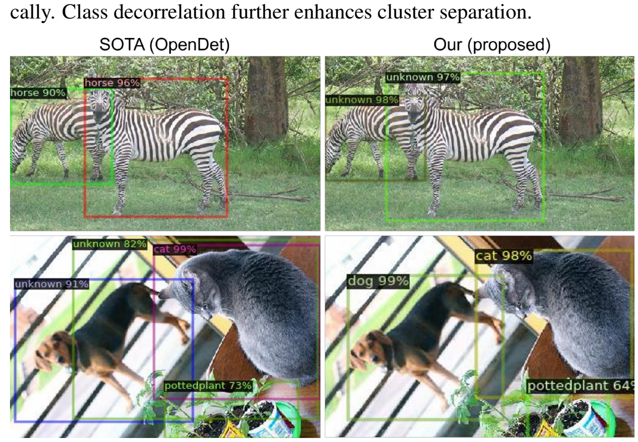
Open-Set Object Detection (OSOD) has emerged as a contemporary research direction to address the detection of unknown objects. Recently, few works have achieved remarkable performance in the OSOD task by employing contrastive clustering to separate unknown classes. In contrast, we propose a new semantic clustering-based approach to facilitate a meaningful alignment of clusters in semantic space and introduce a class decorrelation module to enhance inter-cluster separation. Our approach further incorporates an object focus module to predict objectness scores, which enhances the detection of unknown objects. Further, we employ i) an evaluation technique that penalizes low-confidence outputs to mitigate the risk of misclassification of the unknown objects and ii) a new metric called HMP that combines known and unknown precision using harmonic mean. Our extensive experiments demonstrate that the proposed model achieves significant improvement on the MS-COCO & PASCAL VOC dataset for the OSOD task.

1. Introduction

Object detection task has seen significant advancements in the past decade. However, many of the current object detectors often fail to localize or classify objects of novel or unseen classes. To address this, Open-Set Object Detection (OSOD) has been introduced, which aims to detect new or unidentified objects as "unknown" class along with the known objects with their respective categories. One of the key challenges in OSOD is the issue of misclassification of unknown class with high confidence. This is especially for the case of unknown objects that exhibit semantic closeness to a known class. For instance, an open-set detector that is trained on VOC classes [3] might misclassify an unknown class, e.g., "zebra" to a close known class like "horse". Such a misclassification has been observed in contrastive clustering-based previous OSOD works [6, 8, 29] due to the



(a) Illustration of cluster evolution through incremental component inclusion. Colored dots denote known classes, black dots indicate unknown classes. With semantic clustering, known class clusters align semantically. Class decorrelation further enhances cluster separation.



(b) Visual comparison between the proposed and OpenDet [6] methods.

Figure 1. Effectiveness of semantic clustering, class decorrelation, and visual comparison with OpenDet [6].

semantic proximity of known and unknown classes. Moreover, the methods utilizing classification-based Region Proposal Network (RPN) to predict objectness, obtain lower performance on objects with different geometric attributes than the known classes.

*Corresponding author

This study proposes a new framework to address the aforementioned concerns. To tackle the issue of unknown misclassification, we introduce a new semantic clustering module that aligns region proposal features with their respective semantic class embeddings. This enables the detector to establish meaningful class decision boundaries, thereby preventing unknown misclassification, as illustrated in Figure 1(a). Moreover, we impose an orthogonality constraint on the features to ensure a clear separation of the clusters. For this purpose, we introduce a new class decorrelation module inspired by [24] that utilizes the feature decorrelation-based softmax-formulated orthogonality constraint on the cluster features. This module facilitates an increase in inter-class cluster distance, yielding improved unknown separation, as demonstrated in Figure 1(a).

Previous approaches [6, 8] employ binary classification-based objectness prediction that tends to overfit on the training categories and constraints the effectiveness of RPN, especially in OSOD setting as discussed in [10, 29]. We alleviate this constraint with object focus loss, that learns objectness with the help of centerness and classification-based objectness loss. Here, the centerness helps the detector to predict how far a proposal is from a ground-truth bounding box. This enables a more robust learning of the RPN, as it prevents overfitting on the training categories by learning from object cues, such as location, geometry and other spatial relationships. This, in turn, facilitates easier and unconstrained detection of unknown objects.

The proposed model accurately identifies unseen objects as unknown class and enhances detection performance compared to previous state-of-the-art (SOTA) methods. In Figure 1(b), results obtained from the proposed method and OpenDet [6] are illustrated that demonstrates our method detects the unknown classes accurately as compared to OpenDet [6]. For instance, the proposed model accurately predicts the “zebra” as an “unknown” object, while OpenDet [6] fails and identifies it as a “horse”. Furthermore, in another example, OpenDet [6] predicts “dog” as an “unknown” class, where our model correctly identifies its class. We summarize the contributions of this paper as follows.

- We propose a new OSOD framework which aligns class representations effectively and detects the unknown objects accurately.
- We introduce i) a novel semantic clustering module to group the features in the semantic space that facilitates improved cluster boundary separation, especially between semantically similar objects, ii) a class decorrelation module to further encourage separation of the formed clusters, iii) a new loss known as object focus loss to enable a more resilient learning process of RPN that facilitates the unconstrained detection of unknown objects.
- A new evaluation technique, entropy thresholding, is employed to penalize low-confidence outputs, thereby mitigating the risk of misclassifying the unknown objects as known classes. In addition, a new evaluation metric, Harmonic Mean Precision (*HMP*), is employed to combine the precision scores of known and unknown objects.
- We performed extensive experiments on benchmark datasets, showing significant improvements over prior work. We also conducted various ablation studies to validate the usefulness of the proposed method.

2. Related Works

Open-Set Recognition (OSR) aims to identify the known objects along with unknown or novel objects that were not seen during the training phase. Bendale *et al.* [1] were the first to introduce a deep learning-based OSR method. Subsequently, several reconstruction-based methods have been proposed to enhance the performance of the OSR task. Few works [4, 13] have employed the generative adversarial network to generate potential open-set images to train an open-set classifier. While other approaches [14, 18, 25] utilized the auto-encoder to recover latent features and identify unknown class by reconstruction errors.

Open-Set Object Detection (OSOD) is an extension of OSR that aims to detect unseen object as unknown. Dhamija *et al.* [2] have first formalized OSOD and found that the performance of most detectors is exaggerated in open-set conditions. Joseph *et al.* [8] proposed ORE method by introducing an energy-based unknown identifier. Subsequently, several works have been proposed [5, 21, 23, 26, 31] to improvise the performance of the ORE model. Recently, Gupta *et al.* [5] adapted the Deformable DETR model [30] for the open world objective and introduced OW-DETR. Zohar *et al.* [31] proposed a method by integrating the probabilistic objectness into the deformable DETR model [30]. Miller *et al.* [12] have implemented the dropout sampling technique to estimate uncertainty in object detection, aiming to mitigate open-set errors. Recently, Han *et al.* [6] have introduced contrastive feature learner to encourage compact features of known classes and unknown probability learner to separate known and unknown classes. Zhou *et al.* [29] have enhanced the generalization ability for unknown object proposals using a classification-free RPN. These methods [6, 8, 29] adopt a contrastive-based clustering approach to distinguish the unknown objects from the known clusters. However, these methods underperform in cases where an unknown object is semantically closer to a known class. To address this issue, we propose semantic clustering and class decorrelation modules that aims to learn a clear cluster boundary between semantically similar cluster and encourage separation among them.

Classification-free object detection (CFOD) has emerged recently which focuses on object detection by learning general objectness features rather than relying on class information. This allows to detect previously unseen

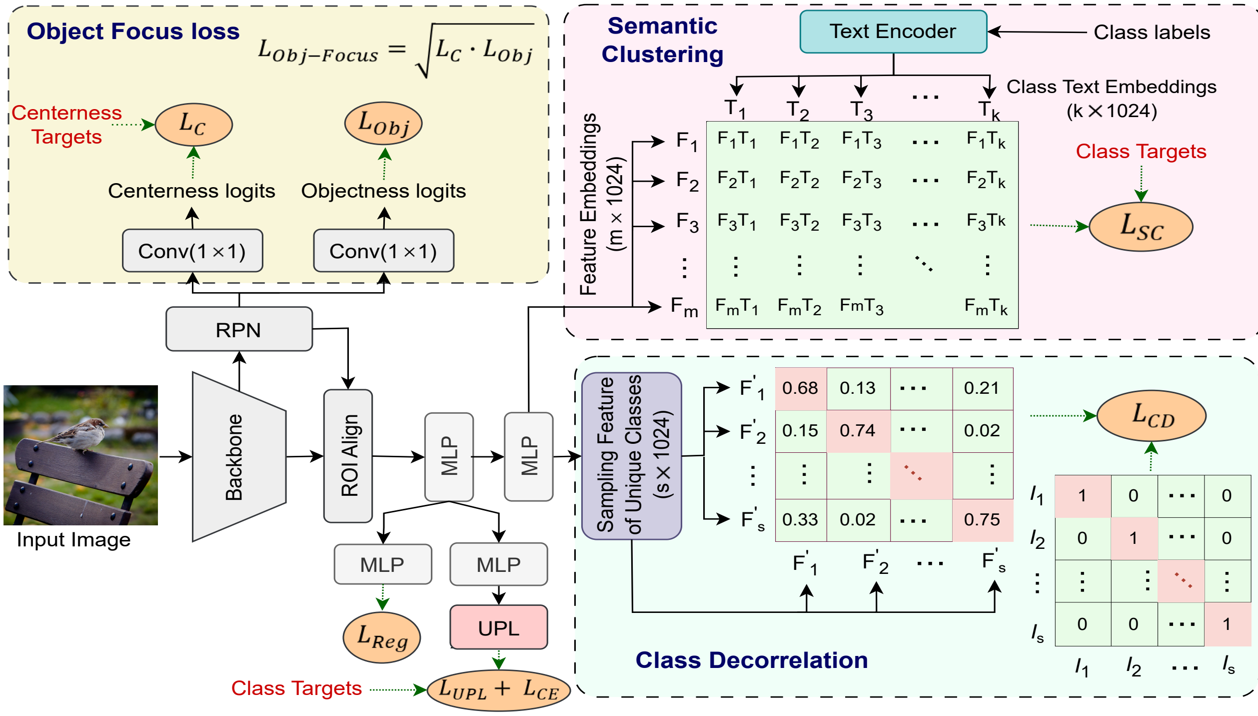


Figure 2. **Overview of our proposed method**. **Object Focus loss:** Object focus loss is a combination of L_C (classification free loss) and L_{obj} (classification based loss). **Semantic Clustering:** $\{T_1, T_2, T_3, \ldots, T_k\}$ represents the class embeddings of k classes and $\{F_1, F_2, F_3, \ldots, F_m\}$ represents m feature embeddings. Each F_i gets aligned with its corresponding class embedding. **Class Decorrelation:** We sample one feature per unique class to create $\{F'_1, F'_2, F'_3, \ldots, F'_s\}$, where s represents the total count of unique classes within an iteration. These sampled features are subsequently orthogonalized against the remaining features to ensure orthogonality.

object classes using these learned features. Recently, Kim *et al.* [10] incorporates centerness features to learn objectness, demonstrating that substituting classification-based loss with a classification-free variant enhances performance for open world proposals. In [29], Zhou *et al.* utilized a classification-free RPN approach for detecting unknown object proposals. Wu *et al.* [22] investigated the direct integration of CFOD into an open-set setting, resulted in a 70% decrease in unknown object recall. Inspired by these findings, we propose a new object focus loss that uses a combination of classification-free and classification-based object detection for learning objectness.

3. Proposed Framework

3.1. Problem Statement & Notations

Given an object detection training dataset $D_{tr} = \{(x, y), x \in X_{tr}, y \in Y_{tr}\}$ with known classes $C_k = \{c_1, c_2, ..., c_k\}$ and testing dataset $D_{te} = \{(x', y'), x' \in X_{te}, y' \in Y_{te}\}$ containing k known classes (C_k) as well as u unknown classes (C_u), the objective of OSOD is to accurately detect all known objects belonging to C_k, while also identifying novel objects as "unknown" class. In this context, X_{tr} and X_{te} represent the input images of training and testing dataset, respectively, while Y_{tr} and Y_{te} are set of training and testing labels containing corresponding class labels and bounding boxes. As it is unfeasible to enumerate infinite unknown classes, we denote unknown classes as $C_u = c_{k+1}$.

The proposed framework introduces three alignment modules to effectively segregate the known class clusters and detect unknown classes accurately. (i) The CLIP-based semantic clustering module that facilitates the formation of clusters in the semantic space. (ii) The class decorrelation module, which enforces an orthogonality constraint among features of different clusters and helps to separate the clusters. (iii) The object focus loss that enhances the unknown detection capabilities. Detailed explanations of these modules are discussed in the subsequent subsections.

3.2. Semantic Clustering

Contrastive learning has significantly enhanced the deep clustering of images by capturing distinct visual attributes customized to each instance [7, 9, 15, 19, 27]. Nonetheless, its capacity to explicitly infer class decision boundaries still needs to be explored. This can be attributed to the absence of class sensitivity within the instance discrimination strategy, leading to clusters in the feature space that are not effectively aligned with class decision boundaries. This issue is further compounded when applied in the OSOD setting, where an unknown object is misclassified to its semantically closer known class.

To address this issue, we have introduced a semantic clustering module inspired by CLIP [16]. This module facilitates the formation of clusters in the semantic space, thereby establishing clear semantic decision boundaries. Consequently, the module mitigates the confusion between semantically similar objects, thus minimizing the misclassification of unknown objects to known classes. In this regard, we utilize the CLIP-based text encoder to generate label embeddings $\{T_1, T_2, T_3, \dots, T_k\}$ that correspond to k ground truth VOC classes [3]. Subsequently, a MLP layer is appended after the ROI align module to generate 1024-dimensional image features $\{F_1, F_2, F_3, \dots, F_m\}$ for m sampled proposals. For each proposal feature F_i, we calculate the cosine similarity with each text embedding. The cosine similarity between the i^{th} image feature and the j^{th} label feature is then calculated as

$$\text{cos-sim}_{ij} = \frac{F_i \cdot T_j}{\|F_i\| \cdot \|T_j\|}, \tag{1}$$

The resulting output (i.e., cos-sim_{ij}) is subsequently utilized to evaluate the cross-entropy loss with respect to the ground-truth labels. The final semantic clustering loss is defined as

$$L_{SC} = \sum_{i=1}^{m} \sum_{j=1}^{k} \Upsilon_{ij} \cdot \log\left(\frac{e^{\text{cos-sim}_{ij}}}{\sum_{l=1}^{n} e^{\text{cos-sim}_{il}}}\right). \tag{2}$$

Here, Υ_{ij} denotes the one-hot encoding belonging to i^{th} feature proposal and j^{th} class. This approach assists in achieving compact semantic clusters of the known classes and it creates clear separation boundaries between clusters of similar semantics (see Figure 1(a)).

3.3. Class Decorrelation

To encourage separation among known clusters, we propose a class decorrelation module inspired by [24]. This module imposes an orthogonality constraint on the proposal features, effectively enhancing the inter-cluster distance, thus facilitating better separation of known classes. Initially, one feature from the set of proposal features $\{F_1, F_2, \dots, F_m\}$ is sampled for each distinct class within the batch, resulting in a subset of s features $\{F'_1, F'_2, \dots, F'_s\}$ for s unique classes. Subsequently, we proceed to orthogonalize this subset of features that reduces potential correlations among features, enhancing cluster separation and enabling the model to focus on class-specific differences.

To perform orthogonalization, we first compute the cosine similarity for all the sampled features, forming a similarity matrix.

$$\text{sim}_{i,j} = \frac{F'_i \cdot F'_j}{\|F'_i\| \cdot \|F'_j\|} \tag{3}$$

Then we calculate the corresponding correlation between the i^{th} and j^{th} feature (i.e., $\text{corr}_{i,j}$) as given in Equation 4.

$$\text{corr}_{i,j} = \frac{e^{\text{sim}_{i,j}}}{\sum_{l=1}^{k} e^{\text{sim}_{i,l}}} \tag{4}$$

Here, $\text{corr}_{i,j}$ gauges both the self-correlation of a feature vector and its dissimilarity from other vectors. The primary objective of this step is to establish inter-dependencies among the chosen features, thereby reflecting the innate properties of various classes present within the feature set. Finally, the class decorrelation loss (i.e., L_{CD}) is calculated via cross-entropy between resulting correlation matrix and s unique class-based identity matrix (i.e., I).

$$L_{\text{CD}} = \sum_{i=1}^{s} \sum_{j=1}^{s} I_{i,j} \cdot \log\left(\text{Corr}_{i,j}\right) \tag{5}$$

Our objective revolves around diagonalizing this correlation matrix, which enforces decorrelation among the features associated with distinct classes. This not only enhances the clarity of class boundaries but also contributes to improving cluster formation (see Figure 1(a)).

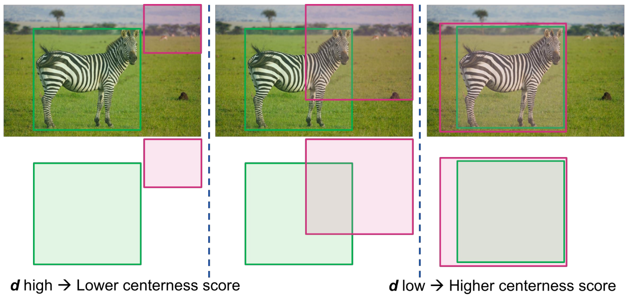
3.4. Object Focus Loss

In Faster R-CNN [17], the RPN places significant emphasis on ground truth objects to acquire knowledge of the objectness score. Nevertheless, this methodology frequently results in the issue of overfitting on the known training classes. This inclination towards overfitting impedes the model's ability to identify new and previously unseen object classes during inference efficiently.

To address this limitation, we introduce object focus loss, which leverages the concept of centerness [10, 20]. This centerness enables the model to consider critical attributes such as object location, shape, and geometric properties to learn objectness. As a result, we develop a class-agnostic object proposal detection mechanism encompassing a broader range of object characteristics. The centerness loss is computed using centerness logits (i.e., C_{logits}) and centerness targets (i.e., $C_{targets}$). In our proposed framework, we have added a single convolution layer after RPN network, which gives us the C_{logits}. Furthermore, the corresponding $C_{targets}$ are generated inspired by [20, 29]. It measures how far the center of an object proposal is from the center of a ground-truth bounding box as depicted in Figure 3(a). The details of generating $C_{targets}$ are presented in Supplementary material. The centerness loss (i.e., L_C) can be accomplished by evaluating the disparity between the C_{logits} and $C_{targets}$ as given below.

$$L_C = |C_{logits} - C_{targets}|_1 \tag{6}$$

Finally, the object focus loss is calculated as the geometric mean of the centerness loss (L_C) and the classification-



(a) Illustration of centerness score in the calculation of object focus loss. Green colored bounding box indicate ground-truth annotation while magenta colored bounding box is predicted proposal annotation. ***d*** denotes the distance between ground-truth annotation and predicted proposal annotation.

(b) Importance of object focus loss in detecting unknown objects

Figure 3. Demonstrates the process of object focus loss along with a visual comparison with and without object focus module.

based objectness loss (L_{Obj}) [17] as

$$L_{Obj-Focus} = \sqrt{L_C \cdot L_{Obj}}. \tag{7}$$

The object focus loss (i.e., $L_{Obj-Focus}$) bridges the gap between classification-free and classification-based detection. The introduction of the centerness loss is a powerful remedy, enhancing the model's capability to generalize and detect more objects belonging to unknown classes during evaluation. By shifting the focus towards the geometric aspects of objects, we enable the model to develop a more comprehensive understanding of objects' inherent features. This results in an enhanced object detection that can effectively handle known and previously unseen classes, improving overall performance and generalization capabilities. We demonstrate its efficacy in Figure 3(b), where we compare results on methods with and without object focus loss. We observe that in selective cases, without object focus loss leads to the failure of detecting certain unknown objects, a problem rectified by incorporating object focus loss.

3.5. Architectural Details

We have utilized Faster R-CNN [17] as the baseline network. Further, we add a convolutional layer with a kernel size of 1×1 into the RPN head to enable the regression of centerness logits and the objectness logits generated by the convolutional layer. Moreover, to cluster the m proposal features in the semantic space, two MLP layers are added after the ROI align sampling process. The resultant 1024-dimensional m proposal features are aligned with their corresponding 1024-dimensional text embeddings. The CLIP text encoder [16] is employed to generate the 1024 text embeddings. Unlike the CLIP approach [16], we adopt the class name as a prompt instead of one or more prompts. This setting is validated via ablation analysis presented in Supplementary material. Furthermore, m proposals are also forwarded to the class decorrelation module, where they are preprocessed to sample one feature from each distinct class in the batch.

Overall Optimization: The proposed framework is trained using final loss L_{Final} i.e., the combination of RPN loss (i.e., L_{RPN}) and detector loss (i.e., L_{Det}).

$$\begin{aligned} L_{Final} &= L_{RPN} + L_{Det} \ \ where, \\ L_{Det} &= \alpha_1 L_{SC} + \alpha_2 L_{CD} + L_{UPL} + L_{Reg} + L_{CE} \\ L_{RPN} &= \alpha_3 L_{Obj-Focus} + L_{RPN-Reg} \end{aligned} \tag{8}$$

Here, L_{RPN} is a weighted combination of object focus loss (i.e., $L_{Obj-Focus}$) and RPN-based regression loss (i.e., $L_{RPN-Reg}$). L_{Det} is a weighted combination of semantic cluster loss (i.e., L_{SC}), class decorrelation loss (i.e., L_{CD}), unknown probability loss L_{UPL} and standard Faster R-CNN based regression loss (i.e., L_{Reg}) and cross-entropy loss (i.e., L_{CE}). The L_{UPL} is employed from [6] to learn the unknown probability of a proposal for each instance based on the uncertainty of predictions.

4. Experiments & Result Analysis

4.1. Implementation details

Dataset details: Following [6], we use the widely-used PASCAL VOC [3] and MS COCO [11] benchmark datasets for training and testing purpose. The trainval set of VOC is employed for closed-set training, while 20 VOC classes and 60 non-VOC classes from COCO are adopted to assess the effectiveness of our approach under diverse open-set conditions. Specifically, two settings have been designed, namely VOC-COCO-{T1, T2}.

- **VOC-COCO-T1:** The first setting involves gradually increasing open-set classes to create three joint datasets. Each dataset comprises $n = 5000$ VOC testing images, as well as $\{n, 2n, 3n\}$ COCO images with $\{20, 40, 60\}$ non-VOC classes, respectively.

- **VOC-COCO-T2:** This setting involves gradually increasing the Wilderness Ratio (WR) to create four joint datasets. Each dataset comprises $n = 5000$ VOC testing images and $\{0.5n, n, 2n, 4n\}$ COCO images that are disjoint with VOC classes.

Table 1. Comparison with SOTA methods on VOC-COCO-T1 setting. The best-performing measures are highlighted with **bold font**, while the second-best is highlighted with *underlined italic font*. * indicates the re-trained methods.

Method	VOC	VOC-COCO-20					VOC-COCO-40					VOC-COCO-60				
	$mAP_k \uparrow$	$WI \downarrow$	$AOSE \downarrow$	$mAP_k \uparrow$	$AP_u \uparrow$	*HMP* $\uparrow$	$WI \downarrow$	$AOSE \downarrow$	$mAP_k \uparrow$	$AP_u \uparrow$	*HMP* $\uparrow$	$WI \downarrow$	$AOSE \downarrow$	$mAP_k \uparrow$	$AP_u \uparrow$	*HMP* $\uparrow$
ResNet50 as Backbone																
Faster R-CNN [17]	*80.10*	18.39	15118	58.45	0.00	0.00	22.74	23391	55.26	0.00	0.00	18.49	25472	55.83	0.00	0.00
ORE [8]	79.80	18.18	12811	58.25	2.60	4.98	22.40	19752	55.30	1.70	3.30	18.35	21415	55.47	0.53	1.05
DS [12]	80.04	16.98	12868	58.35	5.13	9.43	20.86	19775	55.31	3.39	6.39	17.22	21921	55.77	1.25	2.45
PROSER [28]	79.68	19.16	13035	57.66	10.92	18.36	24.15	19831	54.66	7.62	13.38	19.64	21322	55.20	3.25	6.14
OpenDet [6]	80.02	14.95	11286	*58.75*	*14.93*	*23.81*	18.23	16800	55.83	*10.58*	*17.79*	14.24	**18250**	56.37	*4.36*	*8.09*
Openset RCNN [29]	**82.94**	*11.58*	*10839*	**59.19**	—	—	*14.48*	*16652*	**56.12**	—	—	*12.41*	19631	**57.01**	—	—
Our (proposed)	78.06	**9.55**	**9267**	58.52	**18.45**	**28.05**	**11.89**	**14057**	*56.10*	**12.56**	**20.52**	**10.96**	*19153*	*56.47*	**5.10**	**9.36**
ConvNet-small as Backbone																
Faster R-CNN* [17]	**84.88**	13.31	16019	*63.85*	0.00	0.00	17.15	24444	*61.25*	0.00	0.00	13.68	26077	*62.18*	0.00	0.00
DS* [12]	83.24	11.22	15257	62.80	6.75	12.19	14.62	23406	60.24	4.81	8.91	12.51	27951	60.58	1.79	3.48
PROSER* [28]	*84.78*	16.32	14222	62.72	*18.06*	*28.04*	20.39	21296	60.20	11.97	19.97	15.40	21889	61.03	4.57	8.50
OpenDet* [6]	83.22	*9.43*	*11700*	63.22	17.49	27.40	*11.79*	*16823*	60.83	*12.28*	*20.43*	*9.61*	*19963*	61.98	*4.72*	*8.77*
Our (proposed)	83.28	**9.06**	**9441**	**64.15**	**20.60**	**31.19**	**11.20**	**13695**	**61.56**	**14.00**	**22.81**	**9.46**	**17561**	**62.26**	**5.34**	**9.84**

Entropy Thresholding: We employ a technique called entropy thresholding to penalize the low-confidence outputs during evaluation. This involves calculating the entropy of the classification head output logits and comparing them with a fixed threshold. In cases where the entropy $(-\mathbf{p} \cdot \log(\mathbf{p}))$ of a logit exceeds the threshold, we classify it as the "unknown" label. This approach ensures that if the open-set detector is not confident in its prediction regarding an object proposal, it is most likely an unknown object that is not encountered during training. After empirical analysis, we have selected a threshold 0.85 for all settings. This thresholding mechanism helps to prevent misclassification of unknown objects as known classes.

Evaluation metrics: We have performed the evaluation on open-set metrics such as Wilderness Impact (WI) [2], Absolute Open-Set Error ($AOSE$) [12], and AP_u (Average Precision of unknown classes), along with closed-set metric i.e., mAP_k (mean Average Precision of known classes). The purpose of WI [2] is to determine the degree to which unknown objects have been misclassified into known classes.

$$WI = \left(\frac{P_k}{P_{k \cup u}} - 1\right) \times 100, \tag{9}$$

where P_k and $P_{k \cup u}$ denote precision of closed-set and open-set classes, respectively. We report the WI under 0.8 recall level as suggested by [8]. Furthermore, $AOSE$ [12] is utilized to quantify the number of unknown objects that have been misclassified. Lowering the WI and $AOSE$ scores indicates better detection performance.

Harmonic Mean Precision: We introduce a new metric called Harmonic Mean Precision (*HMP*) that encapsulates the performance of a detector on both known and unknown classes in one metric. This is achieved by calculating the harmonic mean of the mAP_k and AP_u.

$$HMP = \frac{2 \times mAP_k \times AP_u}{mAP_k + AP_u} \tag{10}$$

Training details: For a fair comparison with existing OSOD methods, we have adopted the Faster R-CNN [17] architecture as our baseline and incorporated our modules to transform it into an open-set detector. During the training phase, for experiments with the ConvNet backbone, we utilized the AdamW optimizer with a learning rate of 1e-4 and trained for 50,000 iterations. Our training process has involved using a single GPU with a batch size of 6. In the case of experiments with the ResNet50 backbone, we opted for the SGD optimizer with a learning rate of 0.002 and trained for 32,000 iterations as well. Like the ConvNet experiments, the training has been conducted using a single GPU but with a larger batch size of 16. We empirically set the weight coefficients α_1 and α_2 to 0.05, and α_3 to 1. For comparison, we choose recent OSOD methods such as Faster R-CNN [17], ORE [8], DS [12], PROSER [28], OpenDet [6], and Openset RCNN [29]. These methods are re-trained on the exact configuration to compare the results on the CovnNet backbone. Furthermore, we have also compared our method with open world object detection methods such as ORE [8], OW-DETR [5], PROB [31] on OSOD-based evaluation protocol proposed in [8] and the corresponding results are presented in Supplementary material.

4.2. Result Analysis

Comparison on T1 setting: Table 1 presents a comparison of state-of-the-art (SOTA) methods on VOC-COCO-T1 setting. This comparison is conducted by utilizing the ResNet50 and ConvNet backbone in terms of open-set and closed-set metrics. The results from Table 1 indicate that, in the context of ResNet50 backbone comparison, the proposed method improves AP_u by 17-24% from OpenDet [6] all dataset settings. Our method also reduces $AOSE$ by 450-2600 and gains 11-18% in WI compared to previous best-performing Openset RCNN [29] method. For the open-set mAP_k measure, the proposed model performs better than the OpenDet [6] method on all the settings except VOC-COCO-20. When comparing the results based on the ConvNet backbone, the proposed framework

Table 2. Comparison with SOTA methods on VOC-COCO-T2 setting. The best-performing measures are highlighted with **bold font**, while the second-best is highlighted with *underlined italic font*. * indicates the re-trained methods.

Methods	VOC-COCO-n					VOC-COCO-2n					VOC-COCO-4n				
	$WI \downarrow$	$AOSE \downarrow$	$mAP_k \uparrow$	$AP_u \uparrow$	$HMP \uparrow$	$WI \downarrow$	$AOSE \downarrow$	$mAP_k \uparrow$	$AP_u \uparrow$	$HMP \uparrow$	$WI \downarrow$	$AOSE \downarrow$	$mAP_k \uparrow$	$AP_u \uparrow$	$HMP \uparrow$
ResNet50 as Backbone															
Faster R-CNN [17]	16.14	12409	74.52	0.00	0.00	24.18	24636	*70.07*	0.00	0.00	32.89	48618	63.92	0.00	0.00
ORE [8]	15.36	10568	74.34	1.80	3.51	23.67	20839	70.01	2.13	4.13	32.40	40865	64.59	2.14	4.14
DS [12]	15.43	10136	73.67	4.11	7.79	23.21	20018	69.33	4.84	9.05	31.79	39388	63.12	5.64	10.35
PROSER [28]	16.65	10601	73.55	8.88	15.85	25.74	21107	69.32	10.31	17.95	34.60	41569	63.09	11.15	18.95
OpenDet [6]	11.70	8282	**75.56**	*12.30*	*21.16*	*18.69*	*16329*	**71.44**	*14.96*	*24.74*	26.69	*32419*	**65.55**	*16.76*	*26.69*
Openset RCNN [29]	*11.59*	*7705*	*74.90*	—	—	—	—	—	—	—	*25.14*	34382	*64.88*	—	—
Our (proposed)	**9.51**	**6875**	73.47	**14.13**	**23.70**	**15.39**	**13615**	69.44	**16.36**	**26.48**	**22.31**	**27362**	64.28	**17.77**	**27.84**
ConvNet-small as Backbone															
Faster R-CNN* [17]	14.23	11611	79.04	0.00	0.00	23.47	23190	73.63	0.00	0.00	33.52	46418	67.11	0.00	0.00
DS* [12]	14.18	11166	76.29	5.15	9.65	22.85	22200	71.39	6.46	11.85	32.84	44648	64.98	7.69	13.75
PROSER* [28]	13.35	10053	78.04	12.85	22.07	22.01	20255	72.63	15.07	24.96	32.39	40389	65.98	17.02	27.06
OpenDet* [6]	**9.05**	*7615*	*79.27*	*13.79*	*23.49*	*15.80*	*15150*	*74.79*	*16.64*	*27.22*	*24.53*	*30159*	*69.20*	*18.74*	*29.49*
Our (proposed)	*9.43*	**6969**	**79.69**	**15.33**	**25.71**	**15.60**	**11977**	**75.74**	**17.87**	**28.92**	**24.29**	**27654**	**70.27**	**19.47**	**30.49**

Figure 4. Visual comparison between our proposed model and baseline methods such as Faster R-CNN [17] and OpenDet [6]. More results can be visualized from the Supplementary materials. (Zoom-in for a better view)

outperforms other methods across all metrics, achieving an increase by 13-18% in AP_u and reducing $AOSE$ by approximately 2200-3200 compared to the previous best-performing method [6]. We also show a gain in HMP by 11-18% across all settings on both backbones against [6].

Comparison on T2 setting: In Table 2, a comparison between the proposed and existing SOTA methods on the VOC-COCO-T2 setting is presented. In the context of ResNet50-based comparison, the proposed model outperforms other methods in terms of WI, $AOSE$, and AP_u across all dataset settings. For example in VOC-COCO-n, we show an increase of 14.9% in AP_u from [6]. Similarly, in comparison on the ConvNet backbone, the proposed method outperforms other methods on all dataset settings, except on the WI measure in the VOC-COCO-n setting, where the proposed method achieves a comparable measure with OpenDet [6]. With ConvNet backbone, our method achieves a 4-12% increase in AP_u, and consistently improves upon $AOSE$ by a large margin of 600-3200 compared to [6]. In HMP, we improve upon [6] by 3-12% across all settings on both backbones.

Visual Comparison: In addition to quantitative analysis, a qualitative comparison is provided in Figure 4 to demonstrate the improvement of our method over baseline methods, i.e., Faster R-CNN [17]) and previous best-performing OpenDet [6]. It can be visualized that the proposed method accurately classifies unknown objects that are semantically closer to known classes which other methods fails to do. For example, OpenDet [6] misclassifies 'bed' as 'dining table' due to their semantic similarity. However, our model, having learned semantic-based clusters, correctly labels 'bed' as the 'unknown' class. A similar analysis is conducted on 'elephant' and 'toilet', which OpenDet [6] misclassifies as 'cow' and 'chair', respectively. Our method, however, accurately identifies these as 'unknown' classes.

Table 3. Ablation analysis to validate the proposed components: semantic clustering (*SC*), class decorrelation (*CD*) and Object Focus loss (*OF*) on VOC-COCO-40 setting.

	SC	*CD*	*OF*	$WI\downarrow$	$AOSE\downarrow$	$mAP_k\uparrow$	$AP_u\uparrow$	$HMP\uparrow$
Case 1	✓			11.47	15561	60.22	12.29	20.41
Case 2		✓		12.26	16041	59.58	11.84	19.75
Case 3			✓	11.91	15268	60.34	12.22	20.32
Case 4	✓	✓		11.28	16301	61.12	12.31	20.49
Case 5	✓		✓	11.30	15225	61.52	13.59	22.26
Case 6		✓	✓	12.09	15642	60.42	13.95	22.67
Proposed	✓	✓	✓	**11.20**	**13695**	**61.56**	**14.00**	**22.81**

4.3. Ablation Studies & Analysis

To ensure a fair comparison, all ablation experiments were trained using ConvNet backbone and evaluated on the VOC-COCO-40 setting.

Effects of proposed method's components: In this investigation, we examine the impact of each component of the proposed framework, i.e., semantic clustering (*SC*), class decorrelation (*CD*) and object focus (*OF*)[1]. The proposed framework is trained utilizing either individual or combined introduced components to assess their efficacy on the comprehensive performance. The corresponding outcomes are presented in Table 3, which reveals that all three introduced components significantly improve the performance of the proposed framework in known and unknown evaluation metrics. Furthermore, one can notice that adding each proposed component improves the performance of detector. For instance, adding *OF* module to *SC* and *CD* module (i.e., Case 5 and Case 6) improves the known as well as unknown detection performance, proving the importance of *OF* module, similarly, adding *CD* module to *SC* i.e., Case 4, enhance the detection performance as compared to Case 1 and Case 2. This substantiates the consequence of our introduced modules in the proposed framework.

Effects of geometric mean operation in Object Focus loss: In proposed object focus loss, we have employed the geometric mean between L_C and L_{Obj} (see Equation 7). To validate of this operation, we have conducted a series of experiments utilizing various settings, including only RPN-based object loss (i.e., L_{Obj}), only centerness loss (i.e., L_C), as well as the addition and multiplication of L_{Obj} and L_C. The results of these experiments are presented in Table 4, which demonstrate that the combination of L_C and L_{Obj} through the geometric mean operation performs better than the other settings.

Effect of loss weights: The proposed framework is trained using a weighted combination of multi-task losses, wherein the weight coefficients for L_{SC}, L_{CD} and $L_{Obj-Focus}$ are represented by α_1, α_2 and α_3, respectively. To determine these weight coefficients, a few experiments were conducted and the corresponding results in terms of *HMP* measure are illustrated in Figure 5. The graph portrays the analysis accomplished by varying one weight coefficient while keeping the value of the remaining coefficients fixed. After conducting empirical analysis, it was discovered that the $\alpha_1 = 0.05$, $\alpha_2 = 0.05$ and $\alpha_3 = 1.0$ combinations give us better *HMP* measure than other settings.

[1] without object focus (*OF*) refers to only standard RPN based objectness loss L_{Obj}

Table 4. Ablation analysis to validate the geometric mean operation in object focus loss on VOC-COCO-40 setting.

	$WI\downarrow$	$AOSE\downarrow$	$mAP_k\uparrow$	$AP_u\uparrow$	$HMP\uparrow$
Only L_{Obj}	11.28	16301	61.12	12.31	20.49
Only L_C	14.53	6758	15.86	2.78	4.73
$L_{Obj} + L_C$	11.97	15260	60.36	11.87	19.84
$L_{Obj} \times L_C$	**11.14**	19645	43.70	10.67	17.15
$\sqrt{L_{Obj} \cdot L_C}$ (proposed)	11.20	**13695**	**61.56**	**14.00**	**22.81**

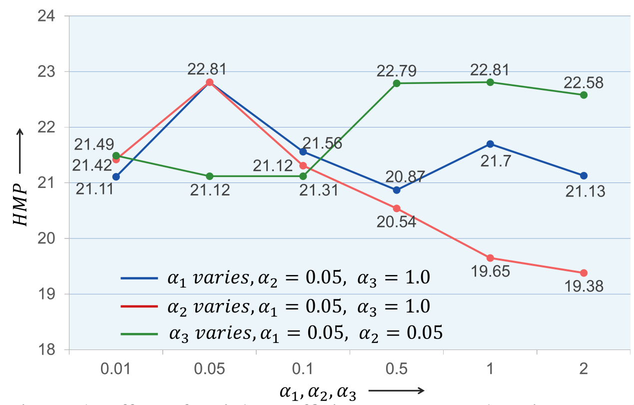


Figure 5. Effect of weight coefficients α_1, α_2 and α_3 in terms of *HMP* measure on VOC-COCO-40 setting.

5. Conclusion & Future Work

This paper proposes a new framework that effectively aligns known class representations to detect unknown objects accurately. The proposed model offers a solution to the issue of high-confidence unknown misclassification in OSOD. We demonstrate that clustering in the semantic space facilitates the formation of well-defined boundaries between clusters, particularly for semantically similar classes that are highly susceptible to misclassification. Additionally, we introduce a class decorrelation module that promotes inter-cluster separation and an object focus loss, wherein the objectness learning exhibits robustness in detecting novel and unseen objects. We also employ an entropy-thresholding-based evaluation technique that penalizes low-confidence outputs, thereby reducing the risk of misclassifying unknown objects. Finally, we carried out extensive experiments & ablation studies and found that the proposed method outperforms existing SOTA methods with significant margin. As the proposed approach have great potential to improve class alignment, it can be further extended to other open-set tasks like incremental object detection and open-set domain adaptation.

References

[1] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016. 2

[2] Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. The overlooked elephant of object detection: Open set. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1021–1030, 2020. 2, 6

[3] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 1, 4, 5

[4] ZongYuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*, 2017. 2

[5] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9235–9244, 2022. 2, 6

[6] Jiaming Han, Yuqiang Ren, Jian Ding, Xingjia Pan, Ke Yan, and Gui-Song Xia. Expanding low-density latent regions for open-set object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9591–9600, 2022. 1, 2, 5, 6, 7

[7] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[8] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5830–5840, 2021. 1, 2, 6, 7

[9] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 3

[10] 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*, 7(2):5453–5460, 2022. 2, 3, 4

[11] 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: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[12] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3243–3249. IEEE, 2018. 2, 6, 7

[13] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 613–628, 2018. 2

[14] Poojan Oza and Vishal M Patel. C2ae: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2307–2316, 2019. 2

[15] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021. 3

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 5

[17] 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, 2015. 4, 5, 6, 7

[18] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7352–7362, 2021. 2

[19] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7352–7362, 2021. 3

[20] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 4

[21] 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*, pages 35–40, 2022. 2

[22] 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*, pages 35–40, 2022. 3

[23] Zhiheng Wu, Yue Lu, Xingyu Chen, Zhengxing Wu, Liwen Kang, and Junzhi Yu. Uc-owod: Unknown-classified open world object detection. In *European Conference on Computer Vision*, pages 193–210. Springer, 2022. 2

[24] Kouta Nakata Yaling Tao, Kentaro Takagi. Clustering-friendly representation learning via instance discrimination and feature decorrelation. *Proceedings of ICLR 2021*, 2021. 2, 4

[25] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4016–4025, 2019. 2

[26] 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. 2

[27] Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label efficient semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10623–10633, 2021. 3

[28] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2021. 6, 7

[29] Zhongxiang Zhou, Yifei Yang, Yue Wang, and Rong Xiong. Open-set object detection using classification-free object proposal and instance-level contrastive learning. *IEEE Robotics and Automation Letters*, 8(3):1691–1698, 2023. 1, 2, 3, 4, 6, 7

[30] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2

[31] 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*, pages 11444–11453, 2023. 2, 6