Generalized Open-World Semi-Supervised Object Detection

Garvita Allabadi **Ana Lucic** **Siddarth Aananth** **Tiffany Yang**
Yu-Xiong Wang **Vikram Adve**
University of Illinois at Urbana-Champaign
{garvita4,alucic2,aananth2,yang244,yxw,vadve}@illinois.edu

Abstract

Traditional semi-supervised object detection methods assume a fixed set of object classes (in-distribution or ID classes) during training and deployment, which limits performance in real-world scenarios where unseen classes (out-of-distribution or OOD classes) may appear. In such cases, OOD data is often misclassified as ID, thus harming the ID classes accuracy. Open-set methods address this limitation by filtering OOD data to improve ID performance, thereby limiting the learning process to ID classes. We extend this to a more natural open-world setting, where the OOD classes are not only detected but also incorporated into the learning process. Specifically, we explore two key questions: 1) how to accurately detect OOD samples, and, most importantly, 2) how to effectively learn from the OOD samples in a semi-supervised object detection pipeline without compromising ID accuracy. To address this, we introduce an ensemble-based OOD Explorer for detection and classification, and an adaptable semi-supervised object detection framework that integrates both ID and OOD data. Through extensive evaluation on different open-world scenarios, we demonstrate that our method performs competitively against state-of-the-art OOD detection algorithms and also significantly boosts the semi-supervised learning performance for both ID and OOD classes.

1 Introduction

Object detection based on deep learning techniques is highly dependent on two assumptions (1) availability of large-scale labeled datasets, and (2) a common, fixed set of object classes that appear in both training and unlabeled data, i.e., the "closed-world" assumption. While these assumptions may hold in ideal conditions, they often manifest as limitations in real-world situations. Semi-supervised object detection (SSOD) approaches [Sohn et al., 2020, Berthelot et al., 2019, Jeong et al., 2019] address the first limitation by leveraging unlabeled data to boost model performance. However, these methods assume the classes in both the training and unlabeled data are sampled from the same distribution, i.e., in-distribution (ID) objects. In practice, the unlabeled data might contain objects that were *unknown* and *unseen* during training, which we call out-of-distribution (OOD) objects.

Prior work [Dhamija et al., 2020, Liu et al., 2022, Miller et al., 2021] shows that the existing SSOD techniques perform worse in the open-set setting, i.e in the presence of OOD objects. This issue is termed as the 'semantic expansion' problem, where OOD objects end up being detected as ID objects and incorrectly used as pseudo-labels for unlabeled data. Liu et al. [2022] addresses the open-set semi-supervised object detection problem, by proposing an offline OOD detector to filter out (reject) any OOD data, to improve the performance of ID classes. These methods, however, disregard any novel OOD data, thereby limiting the learning process to a predefined set of classes.

In this work, we propose an extension of the open-set to a more generalizable open-world setting that not only improves the performance for ID classes but also discovers and includes OOD classes in the

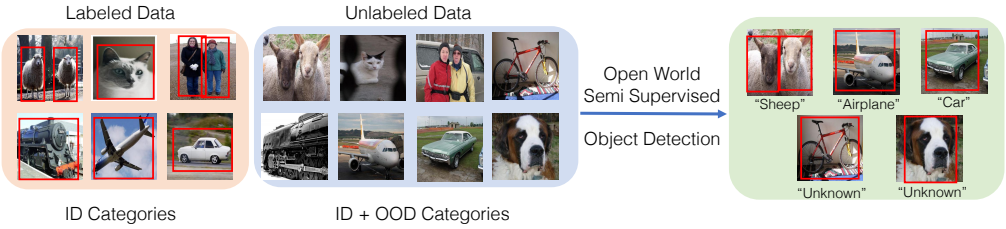

Figure 1: In our task of generalized open-world semi-supervised object detection (OWSSD), the model observes a labeled dataset with ID data and an unlabeled dataset which can contain both ID and OOD data. The objective is to train a detection model that is able to (1) localize and classify instances belonging to seen classes; and (2) localize instances not belonging to seen classes and group them into a new 'unknown' class.

learning process. Specifically, in the generalized open-world setting, we are given a small labeled dataset that includes ID categories of data and a large unlabeled dataset that contains both ID and OOD categories. The aim is to use both the labeled and unlabeled data to improve the performance of ID categories and simultaneously discover OOD data as 'unknown' and adapt the semi-supervised object detection pipeline (Figure 1).

To achieve this, we propose an *OOD Explorer* which is designed to address two primary tasks: the classification of objects as OOD or ID and the localization of novel OOD objects. For the OOD classification task, we propose a simple yet effective OOD detector that leverages an ensemble of lightweight auto-encoder networks trained *only* on ID data. For OOD localization, we systematically analyze multiple unsupervised, class-agnostic object detection techniques, as they align with the inherently open-world nature of the problem. The second component is an *OOD-aware semi-supervised learning framework*, structured into two learning stages. The first stage leverages labeled data, while the second stage incorporates both labeled and unlabeled data, with the help of the OOD Explorer that introduces OOD data.

In our experiments, we examine different open-world scenarios and evaluate the performance of (1) the proposed OOD Explorer on both OOD classification and localization metrics (2) OOD-aware semi supervised object detection. The results show that our method performs competitively against state-of-the-art OOD detection algorithms and that the proposed OOD-aware semi-supervised learning method significantly improves the robustness of ID objects classification and identification through its ability to detect OOD objects and integrate them into the model and learn from them.

Our contributions are summarized as follows:

- We address the problem of generalized open-world semi-supervised object detection problem, by improving the performance on *both* ID and OOD classes.
- We propose a novel *OOD Explorer* for handling novel OOD data, capable of both classification and localization) (Section 3.2). In combination with the OOD Explorer, we develop an OOD-aware semi-supervised object detection algorithm capable of detecting and labeling objects of both ID and OOD classes (Section 3.3).
- We demonstrate that our proposed method is able to detect OOD classes while substantially improving the robustness of ID class detection. (Section 4).

2 Related Work

Semi-Supervised Object Detection. Semi-supervised object detection (SSOD) approaches have become popular to reduce the need for labeling Sohn et al. [2020], Berthelot et al. [2019], Jeong et al. [2019]. Pseudo-labeling based methods such as FlexMatch Zhang et al. [2021], TSSDL Shi et al. [2018], and others Iscen et al. [2019], Luo et al. [2018], Yan et al. [2019], Liu et al. [2021], Xu et al. [2021], first train a teacher model using only labeled data and then use that model to create pseudo-labels for unlabeled images. The pseudo-labels are then used along with the original labeled data to train a student model. On the other hand, consistency regularization approaches such as

Table 1: Comparison of our method with related settings

Setting	Detect Novel OOD Data	Semi-Supervised	Learns from Novel OOD Data
SSOD	✗	✓	✗
Open-World OD	✓	✗	✓
Open-Set SSOD	✓	✓	✗
Our Method	✓	✓	✓

Sajjadi et al. [2016], Laine and Aila [2017], Tarvainen and Valpola [2017], Liu and Tan [2021], Luo et al. [2018], Jeong et al. [2019], Iscen et al. [2019], Liu et al. [2021], Xu et al. [2021], aim to minimize a consistency loss between differently augmented versions of an image. All of these semi-supervised learning approaches assume a "closed-world" setting with a fixed set of classes in both training and testing, which is not a valid assumption in real-world applications.

Open-World Object Detection. Open-world object detection enables the detection of novel objects by incrementally adding novel object classes to the set of known classes. Previous work Kim et al. [2022], Kuo et al. [2015], O Pinheiro et al. [2015], Wang et al. [2020], Maaz et al. [2022] has studied different methods of object proposals for novel objects by attempting to remove the notion of class (all objects are regarded the same). ORE Joseph et al. [2021] is the first to propose an open-world object detector that identifies novel classes as 'unknown' and proceeds to learn the unknown classes once the labels become available. Han et al. [2022] aims to identify unknown objects by separating high/low-density regions in the latent space. Both these approaches work in a fully-supervised setting. Our setup goes a step further and situates the open-world problem in the context of semi-supervised learning, with limited amounts of labeled ID data *only*, that more closely resembles the real-world settings.

Unsupervised Object Localization. Recently proposed methods such as CutLER Wang et al. [2023], FreeSolo Wang et al. [2022], LOST Siméoni et al. [2021], and MOST Rambhatla et al. [2023] propose to localize objects in an unsupervised manner, either by segmentation masks or bounding boxes. Some of these Wang et al. [2023], Siméoni et al. [2021], Rambhatla et al. [2023] use features from self-supervised trained transformers to localize objects in the scene. In our work, we evaluate the capabilities of such methods for localizing OOD objects, as they present open-world capabilities. Based on our evaluation (4.3), we use CutLER as part of the OOD Explorer to localize OOD classes. Section 4 provides the details of our evaluation.

Open-Set/Open-world Semi-Supervised Object Detection. The open-set semi-supervised object detection problem Liu et al. [2022] addressed some of the limitation of the above mentioned work. Furthermore, they address like the performance of ID classes in the presence of OOD data, but they do not learn from it or improve OOD performance. They propose an offline OOD detector to filter out OOD data, thus limiting the risk of ID performance in the presence of OOD data. In contrast, our approach *both* improves performance for ID classes *as well as* OOD classes, i.e., our proposed framework solves a strictly stronger problem. Specifically speaking, Liu et al. [2022] solves for identifying novel classes and filters it out, but does not re-introduce the classes back into the training pipeline in order to be able to learn its features. Mullappilly et al. [2024] addresses some of the limitations of the previous mentioned methods by extending the problem to a semi-supervised setting. However, their problem setting is similar to an incremental learning setting, access to unknown class labels is provided in subsequent tasks. Our generalized setting, on the other hand, does not require access to any unknown class labels.

3 Generalized Open-World Semi-Supervised Object Detection

We propose an integrated framework for *Generalized Open-World Semi-Supervised Object Detection* consisting of an Ensemble-Based OOD Explorer (Section 3.2) and an OOD aware semi-supervised pipeline (Section 3.2). The OOD Explorer plays a crucial role in identifying OOD objects, encompassing both localization and classification into ID or ''unknown' class. The OOD-aware semi-supervised learning pipeline (Section 3.3) ensures that the model assimilates knowledge from both ID and OOD data without risking the forgetting of previously learned ID classes. This integrated approach aims to

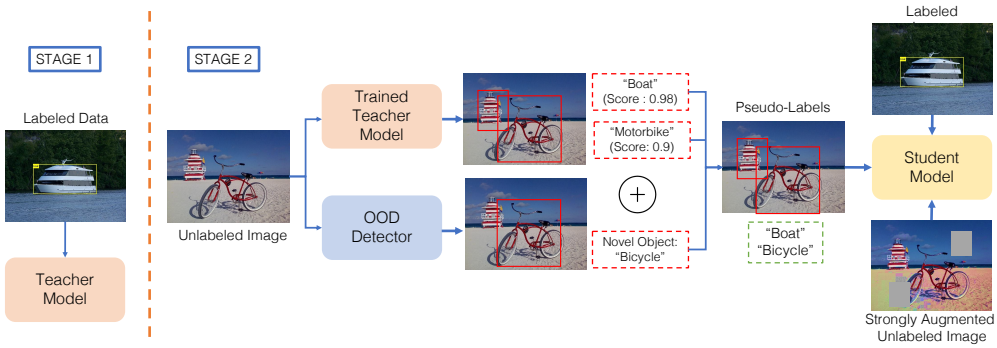

Figure 2: A conceptual overview of our **OOD-aware semi-supervised learning** framework. In the first stage, a **Teacher Model** is trained from the available labeled images with ID categories. In the second stage, unlabeled images that could contain ID and OOD categories are passed through (1) the trained **Teacher Model** to generate pseudo-labels for the ID categories that the Teacher model was trained on; (2) an **OOD Detector** to discover any novel OOD categories present in the unlabeled data. The pseudo-labels generated by the two modules are then fused such that in cases of conflict (high overlap), the novel OOD labels get preference. This helps eliminate false negatives that could be introduced when an OOD category is wrongly classified as an ID category. Finally, a **Student model** is trained jointly on the labeled and unlabeled data to minimize the total loss.

enhance the adaptability and robustness of object detection models in the open-world context. Figure 2 provides a visual summary of our proposed framework.

3.1 Problem Formulation

For our task of generalized open-world semi-supervised object detection, we are given a small labeled dataset $D_l = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ and a large unlabeled dataset $D_u = \{u_1, u_2, \ldots, u_m\}$, where $\{x_i\}$ and $\{u_i\}$ are input images, and $\{y_i\}$ are annotations. D_l consists of a set of ID categories denoted by C_{id} (*i.e.*, $C_l = C_{\text{id}}$), and D_u consists of both C_{id} and a set of OOD categories denoted by C_{ood} (*i.e.,* $C_u = C_{\text{id}} \cup C_{\text{ood}}$). Each annotation y_i in D_l includes a set of object bounding box coordinates and the corresponding class labels from C_{id}.

Our objective is to train a detection model on D_l and D_u jointly that is able to (1) localize and classify instances belonging to one of the ID categories C_{id}; and (2) identify instances belonging to C_{ood} and localizing them. Note that the model does not have access to any C_{ood} bounding box coordinates or labels. The evaluation is performed on a held-out dataset D_v that consists of both objects from C_{id} and C_{ood} categories (*i.e.*, $C_v = C_{\text{id}} \cup C_{\text{ood}}$).

3.2 Ensemble-Based OOD Explorer

In the context of object detection, the integration of OOD data involves two primary tasks: OOD Classification, which distinguishes OOD data from ID data, and OOD Localization, which concurrently provides the precise location of OOD objects within the image.

OOD Classification. We propose an ensemble model that is trained using labeled ID data *only* and without OOD samples. Our ensemble model comprises of multiple auto-encoder networks, each trained on samples from a specific ID class. Each model within the ensemble is dedicated to learning a lower-dimensional representation (encoding) for its respective category. An autoencoder network uses an encoder ($z = E_\phi(x), E_\phi : \mathcal{X} \rightarrow \mathcal{Z}$) to compress its input data and a decoder ($\hat{x} = D_\theta(z), D_\theta : \mathcal{Z} \rightarrow \mathcal{X}$)to reconstruct the original input.

The autoencoder learns by minimizing the reconstruction error R, which gives a measure of how well the output was reconstructed compared to the input. R is estimated with a dissimilarity function d, such that $d(x, \hat{x})$ measures how much x differs from $\hat{x}$ ($R = d(x, D_\theta(E_\phi(x)))$)

When an auto-encoder model is trained only with a specific category of ID data, the reconstruction error, R, for samples of that category is low. At test time, when confronted with an OOD sample,

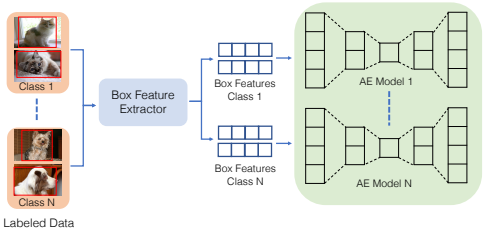

Figure 3: Training Stage

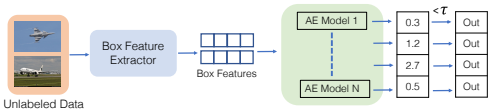

Figure 4: Testing Stage

Figure 5: Proposed Ensemble Based OOD Detector. In the training stage, ground truth proposals from the labeled data are passed through the feature extractor, the output of which is used to train our OOD detector to learn the representation of ID classes. In the OOD prediction stage, unlabeled images are passed through class-agnostic and open-world friendly proposal generators to generate proposals that are then predicted as ID/OOD by the trained OOD detector.

the model is unable to accurately reconstruct the input, resulting in a higher reconstruction error. If this error value is less than a threshold value (μ), then the sample belongs to the corresponding ID class which the model was trained on. This process is repeated for all auto-encoder models in the ensemble. Each model votes if the sample belongs to its class based on the observed R. A sample is deemed OOD if none of the models "claim" it. This simple heuristic illustrates how the decision boundary between ID and OOD samples is established (Figure 5 shows this process).

This threshold is calibrated during training by observing the reconstruction error of ID categories. Since each model is responsible only for its own category, all other categories function as pseudo-OOD data for determining μ. Finally, given that a single image may contain multiple objects of diverse categories, we leverage the box-level features of each object present in the image (Refer to the appendix for details on the training and test procedure).

As seen above, training an autoencoder network is inherently unsupervised as it does not require any information about the class label. By introducing an ensemble of such networks trained on individual categories, we introduce a specialization for each model. *Our key insight is that this introduced specialization is a useful and necessary requirement for OOD detection with very limited labeled data.* When an autoencoder model is trained using the objects of a single category, the model learns the most salient and informative characteristics for that category. Training such an ensemble of models ensures that a model makes confident predictions when encountered with objects from the same ID category and results in a high reconstruction error when encountered with an object different from the one it was trained on. Section 4 describes the variety of experiments we conduct which demonstrate that an ensemble model performs better compared to not only a common autoencoder model trained on all ID classes, but also a variety of state-of-the-art OOD detection algorithms.

OOD Localization. In standard two-stage detectors like Faster R-CNN, object localization is achieved by using a Region Proposal Network (RPN) to generate object proposals, a selection of locations likely to contain objects. This network, however, is trained on a fixed set of ID classes and thus fails to generalize to novel OOD classes.

To address this challenge, we employ CutLER [Wang et al., 2023], an unsupervised method that generates class-agnostic proposals. This characteristic proves advantageous in our open-world setting, as the proposals are not confined to in-distribution (ID) classes. We utilize the models pre-trained on unsupervised ImageNet, including both Cascade Mask RCNN and Mask RCNN. Subsequently, we filter the proposals based on a confidence score.

Additionally, we also analyzed two other localization methods for OOD localization, namely OLN [Kim et al., 2022] and MOST [Rambhatla et al., 2023]. CutLER and MOST use the features extracted from a transformer [Vaswani et al., 2017] network trained with with a self-supervised manner proposed in DINO [Caron et al., 2021] to localize multiple objects in an image. OLN [Kim et al., 2022], on the other hand, estimates the objectness of a candidate region by relying on geometric cues such as location and shape of an object, regardless of its category. We show the performance of these methods when incorporated in our OOD Explorer and SSL pipelines in Section 4.3.

		$AP^{50}\uparrow$			$AR\uparrow$		
Method	Localization	All	ID	OOD	All	ID	OOD
Labeled Only	RPN	57.5	61.3	-	42.8	45.6	-
SSL	RPN	60.3	64.3	-	46.6	49.7	-
SSL + OOD $_{\text{Filtering}}$	RPN	60.4	64.4	-	46.8	49.9	-
SSL + OOD $_{\text{Expansion}}$	RPN + CutLER	61.0	64.4	10.2	48.2	50.0	21.2
SSL + OOD $_{\text{Expansion}}$	RPN + CutLER*	61.2	64.7	8.8	48.4	50.1	21.8

Table 2: Generalized Open-world semi-supervised object detection results on the VOC-15 data split. We examine different open-world settings. The first three settings - Labeled only (Supervised), Semi-supervised (SSL) and OOD Filtering are focused on improving the performance for the original ID classes in the presence of OOD objects. The OOD Expansion setting measures model performance when (unlabeled) OOD objects are also included.

3.3 OOD-Aware Semi-Supervised Learning

We now describe how the OOD samples identified by the OOD Explorer are introduced for training along with the ID samples, in our all *OOD-aware* semi-supervised learning framework.

We adopt the Teacher-Student paradigm and use a two-stage training process (Fig. 2). In the first stage, we use the labeled data to train a **Teacher model**. This model then operates on weakly-augmented unlabeled images to generate bounding boxes and class predictions for the *ID* classes. A subset of confident predictions that are higher than a pre-determined threshold are treated as pseudo-labels.

In the open-world setting, such pseudo-labels could be highly noisy – a novel OOD sample might be wrongly classified as an ID category. To overcome this difficulty, the Teacher model predictions that have an Intersection over Union (IoU) score greater than a threshold τ with the OOD pseudo labels are filtered out from the final list.

In the second stage, a **Student model** is trained using both the labeled data and strongly augmented unlabeled data and the consistency regularization paradigm is used which enforces a model to output the same prediction for an unlabeled sample even after augmentation. The final loss is computed for the labeled samples and unlabeled samples (which includes both ID and OOD categories): $\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_u$

$$\mathcal{L}_s = \sum_i \mathcal{L}_{cls}^{rpn}(x_i, \hat{y}_i) + \mathcal{L}_{reg}^{rpn}(x_i, \hat{y}_i) + \mathcal{L}_{cls}^{roi}(x_i, \hat{y}_i) + \mathcal{L}_{reg}^{roi}(x_i, \hat{y}_i) \tag{1}$$

The supervised loss $\mathcal{L}_s$ is the standard loss for Faster RCNN network comprising of the the RPN classification loss (L_{cls}^{rpn}), the RPN regression loss (L_{reg}^{rpn}), the ROI classification loss (L_{cls}^{roi}), and the ROI regression loss (L_{reg}^{roi}). This loss is computed on the model prediction compared to the ground truth labels. In contrast, the unsupervised loss $\mathcal{L}_u$ is computed on the model prediction on a strongly augmented image compared to the pseudo-label, also consists of the four loss terms. λ is a hyper-parameter that controls contribution of the unsupervised loss to the total loss. The details of data augmentation and hyper-parameter settings are described in Sec. 4.

Note that since we are introducing a new OOD class to the model, instead of using the weights provided by the Teacher model, the Student model is retrained from scratch to alleviate the *catastrophic forgetting* issue observed in incremental learning scenarios [Li and Hoiem, 2017].

4 Experiments

4.1 Experimental Setting and Datasets

Datasets. We evaluate OWSSD on the PASCAL VOC [Everingham et al., 2010] object detection datasets. The trainval sets of VOC07 and VOC12 consist of 5,011 and 11,540 images respectively, both from 20 object categories. The VOC07 test set consists of 4,952 images from 20 object categories.

Table 3: Evaluation of Unsupervised Localization Methods for the OOD Explorer.

Method	$AP^{50}\uparrow$	AR$\uparrow$
OLN	3.5	14.4
MOST	5.2	16.9
Oracle	36.8	45.1
CutLER	10.2	21.2
CutLER*	8.8	21.8

Table 4: Evaluation of OOD detection methods for object detection.

Method	$F_1 \uparrow$	FPR $\downarrow$	AUROC $\uparrow$
LOF	0.33	0.44	0.68
IF	0.37	0.32	0.70
OneSVM	0.33	0.49	0.68
KNN	0.35	0.44	0.70
Ours	**0.46**	**0.14**	**0.72**

We evaluate the open world semi supervised object detection by creating a VOC-15 set that samples the first 15 categories from the VOC07 trainval set as the ID classes. The VOC12 trainval set, with the original 20 categories (15 ID classes and 5 OOD classes), is used as the unlabeled dataset. The model performance is evaluated on the VOC07 test set.

Evaluation Metrics. We report two kinds of metrics: (1) Object-detection metrics for end-to-end evaluation: We report the the average precision at IoU=0.50 (AP^{50}) and average recall (AR) for all, ID, and OOD object categories. (2) Open-world metrics: To demonstrate the effectiveness of OOD detection, we report the F_1 score, False Positive Rate (FPR) and Area Under the ROC curve (AUROC) metrics.

4.2 Main Results

Experiments with OOD Expansion: The aim for these experiments is to evaluate the efficacy of our proposed technique for open world semi supervised object detection. We examine different open-world settings. The Labeled only, Semi-supervised (SSL) and OOD Filtering settings are focused on improving the performance for the original ID classes in the presence of OOD objects. The OOD Expansion setting measures model performance when (unlabeled) OOD objects are also included in the training process. Our results are presented in Table 2.

We find that introducing OOD filtering through our ensemble OOD detector, combined with SSL, enhances the performance of SSL across ID and All classes, compared to the Label Only and SSL baselines. Additionally, when adapting SSL for OOD Expansion using our proposed *OOD Explorer*, we observe substantial improvements, including an increase of +5.3 mean Average Precision (mAP) for all classes and +5.4 mAP for ID classes, compared with using labeled images only, and +0.7 and +0.5 compared with the baseline SSL. Furthermore, our method, in conjunction with CutLER for OOD Localization, demonstrates increased capabilities in detecting OOD objects, achieving an AP^{50} of 10.2 and an Average Recall (AR) of 21.2. Cutler [Wang et al., 2023] uses the MaskRCNN, and using a variant of CutLER with Cascade Mask RCNN (which we label CutLER*) further improves performance slightly for ID objects but hurts slightly for OOD objects.

Comparison of OOD Detection Methods: We compare the OOD detection capabilities of our proposed ensemble method with other OOD detection methods, namely Local Outlier Factor (LOF) [Breunig et al., 2000], Isolation Forest (IF) [Liu et al., 2008], One Class Support Vector Machine (OneSVM) [Schölkopf et al., 2001] and K-Nearest Neighbors (KNN). Table 4 summarizes our findings. OWSSD's OOD detector features a significantly higher F_1 score compared to existing OOD detection methods with the same amount of limited labeled data. Also, the OWSSD OOD detector results in the highest AUROC score that summarizes the ability of the model to distinguish between the seen and unseen classes. The False Positive Rate reflects the rate of false positive OOD instances: the lower the rate the better performance of the model. The FPR rate for OWSSD OOD is the lowest, which explains why our technique is able to outperform the other methods.

4.3 Ablation Studies

Evaluation of OOD Localization methods We evaluated 3 different localization methods for our *OOD Explorer*, namely CutLER [Wang et al., 2023], MOST [Rambhatla et al., 2023] and OLN [Kim et al., 2022]. We also compared the performance against an Oracle that generates perfect proposals for OOD classes, to establish an upper bound. Based on the proposals provided by each of these

Table 5: Choice of reconstruction error threshold for ID vs OOD categorization

Threshold	$F_1 \uparrow$	FPR $\downarrow$	AUROC $\uparrow$
0.05	0.40	0.35	0.74
0.2	0.29	0.003	0.59
Ours 0.1	0.46	0.14	0.72

Table 6: Impact of using an Ensemble compared to a common AE model.

Method	$F_1 \uparrow$	FPR $\downarrow$	AUROC $\uparrow$
Common AE	0.42	0.09	0.67
Ours	0.46	0.14	0.72

methods, we use our ensemble based OOD detector to identify the OOD proposals. The performance for OOD classes is reported after training the student model.

As seen in Table 4.2, CutLER outperforms both OLN and MOST for OOD object detection. Nevertheless, considering the inherent challenges in localizing OOD data without any labels in the context of open-world learning, there remains potential for enhancement, as evidenced by the much better performance of the oracle proposal generator. These results show that accurate localization is a key bottleneck we must address for further performance improvements on open-world semi-supervised object detection.

Impact of different error thresholds for OOD detector: OWSSD's OOD detectors consisting of an ensemble of autoencoder networks categorizes data into ID or OOD classes based on a threshold value μ. We study how the threshold value impacts the OOD detector's performance. The lower reconstruction error indicates that one of the autoencoder models has been able to reconstruct the object and that object class has been claimed. If, however, none of the classes can effectively reconstruct the object featured in the bounding box, the voting mechanism of separate encoders would result in a higher reconstruction error. We tested a series of threshold (μ) values and show results for the following: 0.05, 0.1, 0.2. As Table 4 indicates, the best balance between precision and recall, as reflected through the F_1 score, is achieved with the 0.1 threshold.

Evaluating ensemble approach in the OOD detector. OWSSD uses an ensemble of autoencoder networks for OOD detection. We compared the performance of OWSSD's OOD detector to a common autoencoder model trained for all ID categories with limited labeled data. Table 4.3 shows the results. The ensemble method results in a better recall and F_1 score for OOD detection compared to a common model trained on all labeled ID data, showing that it is able to distinguish OOD classes effectively. As noted previously, the ensemble approach enables each autoencoder to have high accuracy in distinguishing the specific ID class it was trained on, which leads to better discrimination of ID vs OOD objects. (The small increase in False Positive Rate did not affect the F1 and AUROC score advantage that the Ensemble AE method has over the Single AE.)

5 Conclusion

We introduce an generalized OOD-aware semi-supervised object detection approach, based on a Teacher-Student paradigm, which includes both ID and OOD data. We also propose, an OOD Explorer to achieve two goals: 1) effective separation of OOD objects from ID objects and 2) localization of OOD objects. Our results show that this method significantly improves the robustness of semi-supervised object detection methods for ID objects through its capability to not only detect OOD objects but also integrate them into the training process and learn from them.

Limitations and future work. While our method advances the state of the art, there is much room for improvement with respect to the localization and classification of OOD objects. Most importantly, new techniques are needed to localize OOD objects more precisely. Our future work will include a repeated and continual testing of the proposed approach with a varying number of new classes. New approaches such as continuous, in-field learning and the incorporation of domain understanding into current data-driven models may hold promise in achieving these goals.

References

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, page 93–104, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581132174. doi: 10.1145/342009.335388. URL `https://doi.org/10.1145/342009.335388`.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

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.

Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, jun 2010. ISSN 0920-5691. doi: 10.1007/s11263-009-0275-4. URL `https://doi.org/10.1007/s11263-009-0275-4`.

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.

Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5070–5079, 2019.

Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Advances in neural information processing systems*, 32, 2019.

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.

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.

Weicheng Kuo, Bharath Hariharan, and Jitendra Malik. Deepbox: Learning objectness with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2479–2487, 2015.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=BJ6oOfqge`.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008. doi: 10.1109/ICDM.2008.17.

Lu Liu and Robby T Tan. Certainty driven consistency loss on multi-teacher networks for semi-supervised learning. *Pattern Recognition*, 120:108140, 2021.

Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Yen-Cheng Liu, Chih-Yao Ma, Xiaoliang Dai, Junjiao Tian, Peter Vajda, Zijian He, and Zsolt Kira. Open-set semi-supervised object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 143–159. Springer, 2022.

Yucen Luo, Jun Zhu, Mengxi Li, Yong Ren, and Bo Zhang. Smooth neighbors on teacher graphs for semi-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8896–8905, 2018.

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 *17th European Conference on Computer Vision*. Springer, 2022.

Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1485–1488, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589336. doi: 10.1145/1873951.1874254. URL `https://doi.org/10.1145/1873951.1874254`.

Dimity Miller, Niko Sünderhauf, Michael Milford, and Feras Dayoub. Uncertainty for identifying open-set errors in visual object detection. *IEEE Robotics and Automation Letters*, 7(1):215–222, 2021.

Sahal Shaji Mullappilly, Abhishek Singh Gehlot, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Hisham Cholakkal. Semi-supervised open-world object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4305–4314, 2024.

Pedro O O Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. *Advances in neural information processing systems*, 28, 2015.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

Sai Saketh Rambhatla, Ishan Misra, Rama Chellappa, and Abhinav Shrivastava. Most: Multiple object localization with self-supervised transformers for object discovery. *arXiv preprint arXiv:2304.05387*, 2023.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016.

Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7):1443–1471, jul 2001. ISSN 0899-7667. doi: 10.1162/089976601750264965. URL `https://doi.org/10.1162/089976601750264965`.

Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. Transductive semi-supervised deep learning using min-max features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 299–315, 2018.

Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. November 2021.

Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. In *arXiv:2005.04757*, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Rui Wang, Dhruv Mahajan, and Vignesh Ramanathan. What leads to generalization of object proposals? In *European Conference on Computer Vision*, pages 464–478. Springer, 2020.

Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, and Jose M Alvarez. Freesolo: Learning to segment objects without annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14176–14186, 2022.

Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3124–3134, 2023.

Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video salient object detection using pseudo-labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7284–7293, 2019.

Bang Xiang Yong and Alexandra Brintrup. Do autoencoders need a bottleneck for anomaly detection? *IEEE Access*, 10:78455–78471, 2022. URL `https://api.semanticscholar.org/CorpusID:247155132`.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021.

A Implementation Details

We use Faster R-CNN equipped with a Feature Pyramid Network and a ResNet-50 backbone as our object detector. The backbone is initialized with ImageNet pre-trained weights in both stages of training. The implementation is based on the torchvision library [Marcel and Rodriguez, 2010] from PyTorch [Paszke et al., 2019]. For the open-world settings, we use $\mu = 0.1$ as the threshold for VOC-15 data split. The autoencoder model architecture in this paper uses one hidden layer in the encoder and decoder parts of the network. A Feature Pyramid Network vector, 1,024 in length, is used as input to the model that is then compressed to 256 and eventually to 64 in the bottleneck layer. For the semi-supervised settings, we use $\tau = 0.9$ as the pseudo-labeling threshold, $\lambda = 2$ as the unlabeled loss weight. Our data augmentation strategy consists of RandomHorizontalFlip for the first stage of training, and a combination of ScaleJitter, FixedSizeCrop, and RandomHorizontalFlip for the second stage of training. For Cutler, we use the ImageNet pre-trained models using Cascade Mask R-CNN and Mask R-CNN as detectors.

B OOD detection algorithms

Algorithm 1: OOD Detector Training Algorithm

Data: D_l: Images and annotations consisting of class labels and bounding boxes for objects of C_{id} classes
Result: Ensemble of auto-encoder trained models

foreach *c in* C_{id} **do**
  foreach *bb in* D_l *where class* = *c* **do**
    $x = extract_features(bb)$
    $z = E_\phi(x)$
    $\hat{x} = D_\theta(z)$
    $L(x, \hat{x}) = \frac{1}{N}\sum_{i=1}^{N} ||x_i - \hat{x}_i||^2$

Algorithm 2: OOD Detector Testing Algorithm

Data: D_u: Images and class-agnostic proposals for objects of C_{id} and C_{ood} classes;
Ens: Ensemble of autoencoder trained models
Result: List of OOD proposals

foreach *bb in* D_u *where class* = *c* **do**
  foreach *model in Ens* **do**
    $X = extract_features(bb)$
    $z = E_\phi(x)$
    $\hat{x} = D_\theta(z)$
    $L_(x, \hat{x}) = \frac{1}{N}\sum_{i=1}^{N} ||x_i - \hat{x}_i||^2$
    if $L_(x, \hat{x}) < \mu$ **then**
      id = **True**
  if *not id* **then**
    ood_list.add(bb)
return ood_list

C Additional ablation studies

C.1 Impact of choices in our OOD-aware semi-supervised pipeline.

In our OOD-aware semi-supervised pipeline, if generated pseudo-labels for ID classes (i.e. high quality predictions made by the Teacher model) have an Intersection over Union (IoU) score greater than a threshold with the OOD pseudo labels, they are filtered out from the final list of pseudo labels. In this study, we examine the impact of not filtering out the overlapping pseudo-labels. As seen in table C.1, when OOD pseudo-labels and ID pseudo-labels are merged with filtering based on their IoU scores, the overall performance for both ID and OOD classes increases, as no false positives are introduced.

C.2 Impact of number of layers in the autoencoders in the OOD detector

We conducted an ablation study on the architecture of the OOD detector. We increased the number of layers from 3 to 5, including the input and output layers. In our Ensemble autoencoder network, we start with 1,024 features that are compressed to 256 and then to 64 in the bottleneck layer. In the architecture with 5 layers, we start with the 1,024 features that are compressed to 512, 256, 128, and

Table 7: Impact of the OOD-awareness in the semi-supervised object detection pipeline. Our approach results in increased performance by reducing the false positive pseudo labels.

mAP	All	ID	OOD
Without adding OOD data	60.3	64.3	-
With OOD data, without filtering	63.4	65.8	27.1
With OOD data + filtering	64.3	66.1	36.8

64 in the bottleneck layer. The decoder part is symmetrical to the encoder part. In both architectures, each layer is followed by the Rectified Linear Activation (ReLU) activation function, the number of epochs is held constant at 30, the mean squared error is used as the loss function, and the learning rate is set at 0.001. As Table C.2 indicates, we see a slight improvement in the AUROC score with the network with more layers (0.74 versus 0.72). The F_1 score, however, has seen a slight decrease (0.43 compared to 0.46) and the FPR has seen an increase (0.23 compared to 0.14). We have not observed a clear benefit to increasing the number of layers in the Ensemble AE architecture. We leave more extensive testing of the autoencoder network hyperparameters, such as non-bottlenecked architecture[Yong and Brintrup, 2022] that does not include a bottleneck layer for future work.

Table 8: Study on the number of layers in the Ensemble AE architecture

Method	$F_1 \uparrow$	FPR $\downarrow$	AUROC $\uparrow$
Ensemble with 5 layers	0.43	0.23	0.74
Ensemble with 3 layers	0.46	0.14	0.72

D Qualitative Results

Figures 6, 7, 8, 9 shows the qualitative results of our approach. We show how different localization methods for OOD classes impact the quality of results. In all cases, the model was trained on 15 ID classes and tested on 20 classes consisting of 15 ID and 5 OOD classes. The OOD classes includes instances of 'potted plant', 'sofa', 'train' and 'tv monitor'. The aim is to detect instances of both ID categories with their class labels and OOD categories with an 'unknown' class label. As seen in the figures, CutLER [Wang et al., 2023] is able to localize OOD objects more effectively compared to OLN [Kim et al., 2022] and MOST [Rambhatla et al., 2023]. Both OLN and MOST result in a high number of proposals for OOD objects, some of which are inaccurate or miss relevant objects completely. On the other hand, images with 'sheep' are sometimes misclassified as 'cow', we attribute this to the high degree of similarity between these two classes and similar background that the objects appear in.

Figure 6: Qualitative results using our OOD Detector and CutLER (Cascade) [Wang et al., 2023] as the open world localization method.

Figure 7: Qualitative results using our OOD Detector and CutLER (MaskRCNN) [Wang et al., 2023] as the open world localization method.

Figure 8: Qualitative results using our OOD Detector and MOST [Rambhatla et al., 2023] as the open world localization method.

Figure 9: Qualitative results using our OOD Detector and OLN [Kim et al., 2022] as the open world localization method.