Semi-supervised Open-World Object Detection

Sahal Shaji Mullappilly[1], Abhishek Singh Gehlot[1], Rao Muhammad Anwer[1], Fahad Shahbaz Khan[1,2], Hisham Cholakkal[1]

[1]Mohamed bin Zayed University of Artificial Intelligence
[2]Linköping University
{sahal.mullappilly, abhishek.gehlot, rao.anwer, fahad.khan, hisham.cholakkal}@mbzuai.ac.ae

Abstract

Conventional open-world object detection (OWOD) problem setting first distinguishes known and unknown classes and then later incrementally learns the unknown objects when introduced with labels in the subsequent tasks. However, the current OWOD formulation heavily relies on the external human oracle for knowledge input during the incremental learning stages. Such reliance on run-time makes this formulation less realistic in a real-world deployment. To address this, we introduce a more realistic formulation, named semi-supervised open-world detection (SS-OWOD), that reduces the annotation cost by casting the incremental learning stages of OWOD in a semi-supervised manner. We demonstrate that the performance of the state-of-the-art OWOD detector dramatically deteriorates in the proposed SS-OWOD setting. Therefore, we introduce a novel SS-OWOD detector, named SS-OWFormer, that utilizes a feature-alignment scheme to better align the object query representations between the original and augmented images to leverage the large unlabeled and few labeled data. We further introduce a pseudo-labeling scheme for unknown detection that exploits the inherent capability of decoder object queries to capture object-specific information. On the COCO dataset, our SS-OWFormer using only 50% of the labeled data achieves detection performance that is on par with the state-of-the-art (SOTA) OWOD detector using all the 100% of labeled data. Further, our SS-OWFormer achieves an absolute gain of 4.8% in unknown recall over the SOTA OWOD detector. Lastly, we demonstrate the effectiveness of our SS-OWOD problem setting and approach for remote sensing object detection, proposing carefully curated splits and baseline performance evaluations. Our experiments on 4 datasets including MS COCO, PASCAL, Objects365 and DOTA demonstrate the effectiveness of our approach. Our source code, models and splits are available here https://github.com/sahalshajim/SS-OWFormer.

1 Introduction

Conventional object detectors are built upon the assumption that the model will only encounter 'known' object classes that it has come across while training (Girshick et al. 2014; Carion et al. 2020; Zong, Song, and Liu 2023). Recently, the problem of open-world object detection (OWOD) has received attention (Joseph et al. 2021; Gupta et al. 2022),

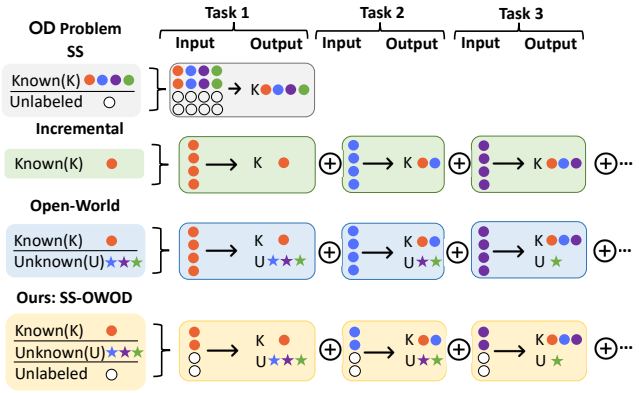

Figure 1: Comparison of our SS-OWOD with other closely related object detection problem settings.

where the objective is to detect known and 'unknown' objects and then incrementally learn these 'unknown' objects when introduced with labels in the subsequent tasks. In this problem setting, the newly identified unknowns are first forwarded to a human oracle, which can label new classes of interest from the set of unknowns. The model then continues to learn and update its understanding with the new classes without retraining on the previously known data from scratch. Thus, the model is desired to identify and subsequently learn new classes of objects in an incremental way when new data arrives.

As shown in Fig.1 *Semi-supervised* (SS) object detection learns a set of known classes (●●●●), while being fed labeled and unlabeled (○) data. In *incremental learning*, classes are learned in steps, as illustrated, the model learns ● in task 1, then fed (●) in the next task and learns to detect (●) without forgetting previously learned class (●), repeating the same process for the subsequent tasks. *Open-world* object detection aims at detecting unknowns (⋆⋆⋆) along with known classes (●). Unknown classes labeled by a human oracle are learned by the model in the next task as illustrated: the unknown (⋆) is learned as a known (●) in the next task while continuing to detect remaining unknowns (⋆,⋆). The same procedure is repeated in the subsequent tasks, where the unknown (⋆) is learned as a known (●). In contrast, we propose the **SS-OWOD** setting that aims to reduce the labeling cost

of the incoming data of detected unknowns (⋆⋆⋆),by leveraging the unlabeled data (○).

Open-world object detection (OWOD) provides a more realistic setting in two ways: (i) It assumes that not all the data in terms of semantic concepts are available during the model training and (ii) it assumes that the data points are non-stationary. Although standard OWOD provides flexibility to detect unknown object categories and then incrementally learn new object categories, the general problem of incremental learning of new classes comes with the need to be trained in a *fully supervised* setting (Fini et al. 2022). To this end, current OWOD approaches rely on strong oracle support to consistently label *all* the identified unknowns with their respective semantics classes and precise box locations.

The objective of this paper is to decrease the aforementioned reliance on the human oracle to provide annotations at run time for the unknown classes (see Fig.1). We argue that it is less realistic to assume that an interacting oracle is going to provide annotations for a large amount of data. The annotation problem becomes extremely laborious in domains like satellite object detection requiring a much higher number of dense oriented box annotations, in the presence of background clutter and small object size. Moreover, existing OWOD methods rely on naive heuristics such as simple averaging across backbone feature channels (Gupta et al. 2022) or clustering of latent feature vectors (Joseph et al. 2021) to pseudo-label unknown objects, thereby struggling to accurately detect the unknowns. To this end, we propose a novel transformer-based method, named SS-OWFormer, that collectively addresses both the issues of improving unknown detection and reducing the annotation cost for identified unknowns during the life span of model learning.

Contributions: The primary contributions of this research encompass the following aspects:

(i) We introduce a novel Semi-supervised Open-World Object Detection (SS-OWOD) problem setting that reduces the strong dependence on external human oracles to provide annotations for *all* incoming data in incremental learning stages. We further propose a *Semi-supervised Open-World object detection Transformer* framework, named, SS-OWFormer, designed to detect a newly introduced set of classes in a semi-supervised open-world setting. SS-OWFormer utilizes a feature alignment scheme to effectively align the object query representations between the original and augmented copy of the image for exploiting the large unlabeled and fewer labeled data.

(ii) We introduce a pseudo-labeling scheme to better distinguish the unknown objects by exploiting the inherent capability of the detection detector object queries to capture object-specific information. The resulting modulated object queries provide multi-scale spatial maps to obtain the objectness confidence scores which in turn are used for the pseudo-labeling process.

(iii) Comprehensive experiments on OWOD COCO split (Joseph et al. 2021) are performed to demonstrate the effectiveness of our approach. Our SS-OWFormer achieves favorable detection performance for both the 'known' and 'unknown' classes in all the tasks, compared to the state-of-the-art OW-DETR (Gupta et al. 2022). SS-OWFormer achieves superior overall detection performance when using only 10% of the labeled data, over OW-DETR using 50% labeled data. In terms of 'unknown' detection, SS-OWFormer achieves an absolute gain of 4.8%, in terms of unknown recall, over OW-DETR.

(iv) Lastly, we explore for the first time the SS-OWOD problem for remote sensing domain. We show the effectiveness of our SS-OWFormer on satellite images, where the labeling task is even more laborious and time-consuming. Moreover, we have proposed open world splits for the Object365 dataset having large number of categories. Our experiments on 4 datasets including MS COCO, PASCAL, Objects365 and DOTA demonstrate the effectiveness of our approach.

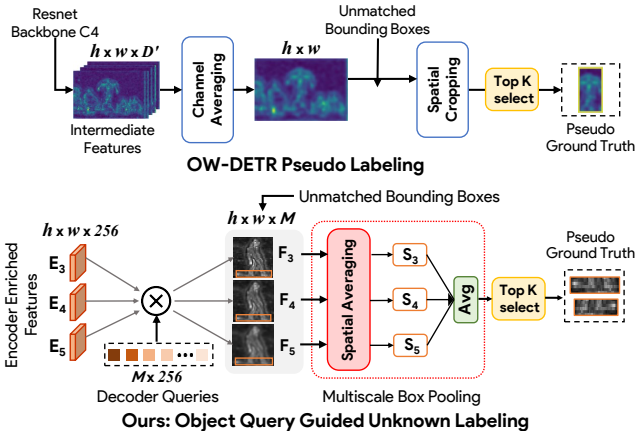

Figure 2: Comparison of our object query guided pseudo-labeling with feature averaging used in OW-DETR baseline. The baseline framework performs a channel averaging over single-scale features from the backbone, spatially crops them at the predicted bounding box positions, and selects the top-k to obtain pseudo-labels. In contrast, our approach strives to leverage object-specific information from multi-scale encoder features *and* decoder object queries. We modulate the decoder object queries with multi-scale encoder feature maps and perform multi-scale box pooling at predicted box locations to obtain objectness scores and select the top-k bounding box proposals as pseudo labels.

2 Preliminaries

Let $\mathcal{D}^t = \{\mathcal{I}^t, \mathcal{Y}^t\}$ be a dataset containing N images $\mathcal{I}^t = \{I_1, I_2, ..., I_N\}$ with corresponding labels $\mathcal{Y}^t = \{Y_1, Y_2, ..., Y_N\}$. Here, each image label $Y_i = \{y_1, y_2, ..., y_k\}$ is a set of box annotations for all k object instances in the image. The open-world object detection (OWOD) follows the incremental training stages on the progressive dataset $\mathcal{D}^t$ having only $\mathcal{K}^t = \{C_1, C_2, ..., C_n\}$ known object classes at time t. A model trained on these $\mathcal{K}^t$ known classes is expected to not only detect known classes but also detect (localize and classify) objects from unknown classes $\mathcal{U} = \{C_{n+1}, ...\}$ by predicting an unknown class label for all unknown class instances. An overview of closely related object detection settings is shown in Fig.1.

Proposed SS-OWOD Problem Setting: Here, each image

label $Y_i = \{y_1, y_2, ..., y_k\}$ is a set of box annotations for all k object instances in the image. The instance annotation $y_k = [l_k, o_k^x, o_k^y, h_k, w_k]$ consists of $l_k \in \mathcal{K}^t$ is the class label for a bounding box having a center at (o_k^x, o_k^y), width w_k, height h_k. In this work, we argue that it is laborious and time-consuming for the human oracle to obtain bounding box annotations for all training images used for learning. Hence, we propose a new *semi-supervised open-world object detection* problem setting, where only a partial set of images (N_s) are annotated by the human oracle and the remaining N_u images are unlabeled (see Fig.1). This aims to reduce the strong dependence on the human oracle for adding knowledge to the model's learning framework. Here, during learning stages in an open-world setting, the model is expected to utilize both labeled and unlabeled sets of training images (N_s+N_u) to learn about the new $\mathcal{K}^{t+1}$ classes, without forgetting previously known $\mathcal{K}^t$ classes, thereby enabling detection of unknown objects at the same time.

2.1 Baseline Framework

We base our approach on the recently introduced OW-DETR (Gupta et al. 2022). It comprises a backbone network, transformer encoder-decoder architecture employing deformable attention, box prediction heads, objectness, and novelty classification branches to distinguish unknown objects from known and background regions. Here, the transformer decoder takes a set of learnable object queries as input and employs interleaved cross- and self-attention modules to obtain a set of object query embeddings. These object query embeddings are used by the prediction head for box predictions as in (Zhu et al. 2020). It selects bounding boxes of potential unknown objects through a pseudo-labeling scheme and learns a classifier to categorize these potential unknown object query embeddings into a single unknown class as in (Gupta et al. 2022). Here, potential unknown objects are identified based on average activations at a selected layer (C4 of ResNet50) of the backbone feature map at regions corresponding to predicted box locations (see Fig.2). Among all potential unknown object boxes, only boxes that are non-overlapping with the known ground-truth boxes are considered pseudo-labels for potential unknowns. It learns a binary class-agnostic objectness branch to distinguish object query embeddings of known and potential unknown objects from background regions. In addition, it learns a novelty classification branch having unknown as an additional class along with $\mathcal{K}^t$ known classes as in (Gupta et al. 2022). We refer to this as our baseline framework.

Limitations: As discussed above, the baseline framework employs a heuristic method for pseudo-labeling with a simple averaging across channels of a single-scale feature map to compute objectness confidence where only single-scale features from the backbone are utilized. However, such a feature averaging to identify the presence of an object at that spatial position is sub-optimal for the accurate detection of unknown objects. To improve unknown object detection, it is desired to leverage the object-specific information available in both deformable encoder and decoder features (see Fig.2). Existing state-of-the-art OWOD frameworks, including our baseline, typically require bounding box supervision for *all* images used during the incremental learning of novel classes in the OWOD tasks. However, this makes the OWOD model strongly dependent on an external human oracle to provide dense annotations for all the data in the subsequent tasks. Next, we introduce our SS-OWFormer approach that collectively addresses the above issues in a single framework.

3 Method

3.1 Overall Architecture

Fig.3 shows the overall architecture of our *Semi-supervised Open-World object detection Transformer (SS-OWFormer)* framework. It comprises a backbone network, deformable encoder, deformable decoder, object query-guided pseudo-labeling, and prediction heads.

The backbone takes an input image of spatial resolution $H \times W$ and extracts multi-scale features for the deformable encoder-decoder network having M learnable object queries at the decoder. The decoder employs interleaved cross- and self-attention and outputs M object query embeddings (z). These query embeddings are used in the box prediction head, objectness and novelty classification branches. In addition, these query embeddings (z) are used in our semi-supervised learning framework to align the current network ($\mathcal{M}_{cur}/z$) with a detached network from the previous task ($\bar{\mathcal{M}}_{prev}/\bar{z}$). We take augmented images as input to the current network and corresponding query embeddings (z^a) are transformed to the latent space of the detached network using a mapping network ($\mathcal{G}$). These transformed embeddings are aligned with the embeddings ($\bar{z}$) obtained for the same images from the detached network using a feature-aligning strategy detailed in Sec.3.3.

We employ fully supervised learning for the first task (task-1) where the object detector is trained with initial known object categories. During task-1 inference, the model is expected to detect all known and unknown object categories. Then, in the subsequent task, the model is trained with new object categories in our novel semi-supervised incremental learning setting where we have annotations only for a partial set of training data. Here, the objective is to learn new object categories using labeled and unlabeled data without forgetting the task-1 categories. To this end, we use a detached network whose weights are fixed during our incremental learning and an identical current network where the network weights are updated. We learn the current network (by taking the detached network as a reference) using labeled and unlabeled data, followed by fine-tuning the current network using available labeled data. Next, we introduce our object query-guided unknown-labeling scheme.

3.2 Object Query Guided Pseudo-Labeling

As discussed, we need to accurately detect unknown objects out of the known set of classes in open-world object detection. Here, the model is expected to transfer its known object detection knowledge to detect unknown objects. Our baseline utilizes a single-scale pseudo-labeling scheme which is a simple heuristic approach with a naive way of averaging Resnet features for pseudo-labeling unknowns. We aim to utilize learnable properties intrinsic to the deformable

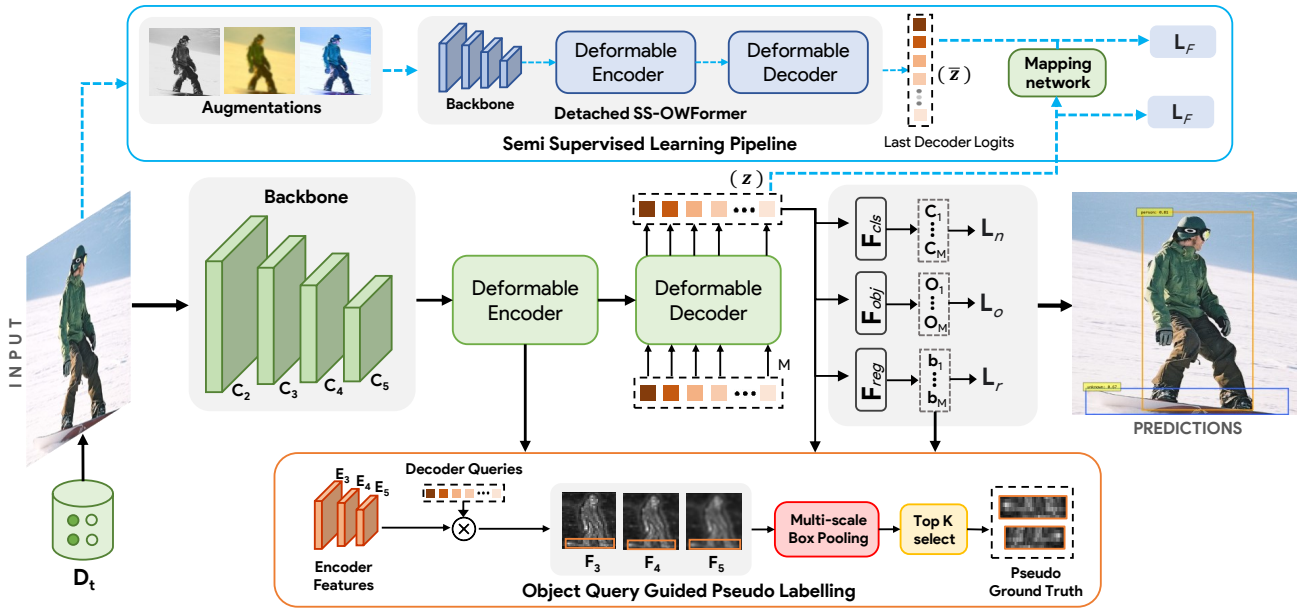

Figure 3: Overall architecture of our Semi-Supervised Open-World object detection Transformer (SS-OWFormer) framework. It comprises a backbone network, transformer-based deformable encoder-decoder, object query-guided pseudo-labeling, box prediction head, novelty classification, and objectness branches. The focus of our design is: (i) the introduction of a *object query-guided pseudo-labeling* (*orange* box at bottom row) that captures information from both transformer encoder and decoder for pseudo-labeling unknown objects. Object queries from the decoder are modulated with the multi-scale encoder features to obtain multi-scale spatial maps which are pooled at predicted box locations to obtain confidence scores for the unknown pseudo-labeling. (ii) The introduction of a novel semi-supervised learning pipeline ($\rightarrow$) for leveraging unlabelled data during incremental learning of a new set of object classes. In our *semi-supervised incremental learning* setting, the SS-OWFormer (current model) is trained along with its detached (frozen) copy (*blue* box on top row) together with a mapping network ($\mathcal{G}$). The mapping network ($\mathcal{G}$) projects the object queries from the current network to the detached network. Moreover, we use original and augmented images for the alignment of object query embeddings (z).

transformer architecture from encoder features and decoder queries. This is found to be more suitable for the objectness confidence levels to be used for pseudo-labeling. Let $F = \{\text{E3}, \text{E4}, \text{E5}\}$ be multi-scale encoder features and $y_k = [o_k^x, o_k^y, h_k, w_k]$ be a box proposal predicted for a given object query embedding. Let $\text{E}_i \in R^{H_i \times W_i \times D}$ be the encoder feature map at scale i and M queries $Q_j \in R^{M \times D}$ be the unmatched object queries at the decoder. Then, we modulate the encoder features with a transposed matrix multiplication to obtain query-modulated feature maps $\mathcal{F}_i \in R^{H_i \times W_i \times M}$. This query-modulated feature map results in better scoring for objectness since it leverages object-specific information from decoder queries along with encoder features. Then, we perform multi-scale box pooling over these maps $\mathcal{F}_i$ at predicted box locations of respective object queries. Our multi-scale box pooling performs spatial averaging over these spatial maps $\mathcal{F}_i$ to obtain objectness scores s_k corresponding to bounding boxes. For instance, the objectness score for a bounding box (b) can be calculated as,

$$\begin{aligned} \sum_{i=0}^{n} S_i(b) &= \frac{1}{h_b \cdot w_b} \sum_{i=0}^{n} \mathcal{F}_i \\ &= \frac{1}{h_b \cdot w_b} \sum_{i=0}^{n} E_i \cdot \sum_{j=0}^{M-K} Q_j^T \end{aligned} \tag{1}$$

These objectness scores are used to select the top k boxes which are then used as pseudo-labels to train the novelty classifier and objectness branches. The regression branch in the prediction head takes M object query embeddings from the decoder and predicts M box proposals. The bipartite matching loss in the decoder selects K queries (from M total queries) as positive matches for the known classes in the supervised setting.

3.3 Semi-supervised Open-world Learning

Previous open-world object detection works assume that all incoming data for novel classes are labeled while in a realistic scenario, it might prove to be costly. However, in our semi-supervised open-world object detection formulation, we employ semi-supervised learning for incremental learning. So, in our challenging setting, the model has to learn to detect novel object categories by using a limited amount of partially annotated data along with unlabeled data for the novel classes and detecting unknown objects, without forgetting previously learned categories.

As discussed in Sec.3.1, we introduce a subset of object categories to the model through subsequent tasks. For the first task, the model is trained like a standard OW object detector, and a set of classes $\mathcal{K}^1 = \{C_1, C_2, ...C_n\}$ are in-

troduced. Then for the subsequent tasks, semi-supervised learning is leveraged for the limited availability of annotations. Using a detached copy of the model from the previous task, $\bar{\mathcal{M}}_{pre}$, the current model $\mathcal{M}_{cur}$ is trained on labeled and unlabeled data with a *feature-aligning strategy*.

For semi-supervised learning using the next progressive dataset D^{t+1}, we employ strong augmentations such as color-jitter, random greyscaling, and blurring to obtain augmented data $\mathcal{D}_a^{t+1} = \{\mathcal{I}^a\}$. The augmentations here are selected such that they do not change the box positions in input images, hence better suitable for semi-supervised object detection. Furthermore, we do not use augmentations such as rotation, flipping, translation, cropping, etc that are likely to alter the feature representation in augmented images. We use a *detached model* $\bar{\mathcal{M}}_{pre}$ whose weights are fixed, a current model $\mathcal{M}_{cur}$ with learnable weights, and a mapping network $\mathcal{G}$ that maps the current model object queries to the detached model object queries. Here, a copy of the current model with fixed weights is used as the *detached model* $\bar{\mathcal{M}}_{pre}$. This detached model does not receive any gradient and remains detached during training.

For an image I_i from D^{t+1}, and its augmented version I_i^a from $\mathcal{D}_a^{t+1}$, we extract object query features using the current and detached models. i.e, We use the current model and obtain original image object query embedding feature $z = \mathcal{M}_{cur}(I_i)$ and augmented image query embedding, $z^a = \mathcal{M}_{cur}(I_i^a)$. Similarly, the detached model is used to obtain the embedding $\bar{z}^a = \bar{\mathcal{M}}_{pre}(I_i^a)$. Then, our mapping network $\mathcal{G}$ maps z^a to $\bar{z}^a$ instead of enforcing z^a to be similar to $\bar{z}^a$ as that may adversely affect the learning in the distillation loss $\mathcal{L}_D$. We perform feature alignment to bring the object queries $\mathcal{G}(z^a)$ and $\bar{z}^a$ together by employing a feature alignment loss $\mathcal{L}_F$. Here, we measure the cross-correlation matrix (Zbontar et al. 2021) between input embeddings and try to bring the object queries closer. The loss also helps to reduce the redundancy between embeddings and makes the representations robust to noise. In addition, the same loss is used to make the model invariant to augmentation, which in turn may help the object query representations z invariant to the state of the model. Then, the current model $\mathcal{M}_{cur}$ is trained using the following loss:

$$\begin{aligned}\mathcal{L}_{cur} &= \mathcal{L}_F\left(\boldsymbol{z}^a, \boldsymbol{z}\right) + \mathcal{L}_D\left(\boldsymbol{z}^a, \bar{\boldsymbol{z}}^{\boldsymbol{a}}\right) \\ &= \mathcal{L}_F\left(\boldsymbol{z}^a, \boldsymbol{z}\right) + \mathcal{L}_F\left(\mathcal{G}\left(\boldsymbol{z}^a\right), \bar{\boldsymbol{z}}^{\boldsymbol{a}}\right)\end{aligned} \tag{2}$$

3.4 OWOD in Satellite Images

Different from the OW-DETR that predicts axis-parallel bounding boxes in natural images, for satellite images we adapt our baseline framework to predict oriented bounding boxes along object directions for a more generalizable approach. For oriented object detection, we introduce an additional angle prediction head in OW-DETR and its standard bounding box prediction heads. Our Object Query Guided Pseudo-Labeling scheme is also found to be suitable for the challenges presented in satellite imagery such as large-scale variations, high object density, heavy background clutter, and a large number of object instances in satellite images. Moreover, the dependence on human oracles for open-world object detection in satellite imagery is highly problematic because of the requirement of a high number of dense-oriented bounding box annotations per image. Thereby, a semi-supervised open-world learning setting can prove beneficial.

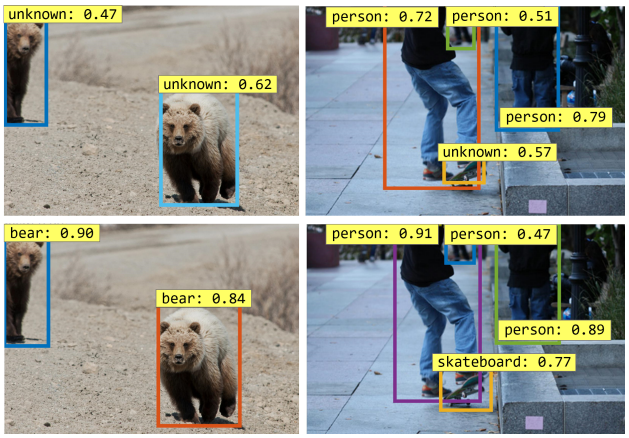

Figure 4: Qualitative results showing the detection performance on MS COCO examples. From the top row, the unknown classes are learned to be marked as a known category in the subsequent tasks as shown in the bottom row.

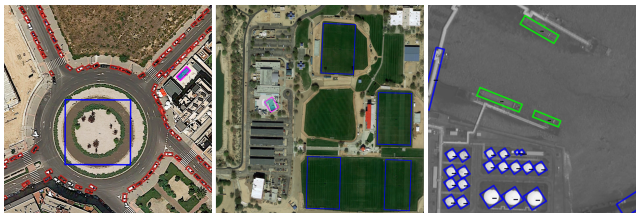

Figure 5: Qualitative results on satellite images with oriented bounding boxes. Oriented bounding boxes in *blue* depict unknown detections on the categories of roundabout, soccer field, and storage tanks in the images respectively. While other colors mark known categories of small-vehicle, swimming pool, and ship.

3.5 Training and Inference

Training: The overall loss formulation for the network can be written as:

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_r + \alpha\mathcal{L}_o + \mathcal{L}_{cur} \tag{3}$$

where $\mathcal{L}_c$, $\mathcal{L}_r$ and $\alpha\mathcal{L}_o$ respectively denote classification, bounding box regression, foreground objectness (class-agnostic) loss terms while $\mathcal{L}_{cur}$ stands for the loss from semi-supervised incremental learning from eq. 2.

The proposed framework follows multi-stage training. The first task is trained in a fully supervised manner using $\mathcal{L}_c$, $\mathcal{L}_r$, $\mathcal{L}_o$. Then, the subsequent tasks follow the *feature alignment* strategy using an additional $\mathcal{L}_{cur}$ loss. A detached model and a current model are trained on augmented unannotated data together with a mapping network $\mathcal{G}$ on top to bring the embeddings closer in latent space using feature-alignment.

Method	Task2				Task3				Task4		
	U-Recall	mAP			U-Recall	mAP			mAP		
		Prev	Cur	Both		Prev	Cur	Both	Prev	Cur	Both
ORE-EBUI	2.9	52.7	26	39.4	3.9	38.2	12.7	29.7	29.6	12.4	25.3
OW-DETR	6.2	53.6	33.5	42.9	5.7	38.3	15.8	30.8	31.4	17.1	27.8
OW-DETR (50%)	6.94	50.53	19.28	34.91	7.64	32.7	9.13	24.85	24.08	5.74	19.49
SS-OWFormer (50%)	**10.56**	**52.04**	**26.35**	**39.2**	**13.16**	**39.46**	**13.63**	**30.85**	**29.97**	**11.48**	**25.35**
OW-DETR (25%)	5.03	49.19	15.64	32.42	6.94	31.02	9.13	23.72	22.9	6.39	18.77
SS-OWFormer (25%)	10.47	52.21	21.16	36.68	12.22	36.4	10.83	27.87	26.91	8.72	22.36
OW-DETR (10%)	4.83	47.8	12.36	30.08	8.24	30.65	6.14	22.48	21.23	4.78	17.11
SS-OWFormer (10%)	**10.19**	**53.61**	**16.44**	**35.02**	**12.13**	**35.21**	**8.11**	**26.18**	**26.17**	**5.33**	**20.96**

Table 1: State-of-the-art comparison for the open-world object detection (OWOD) problem on natural images using MS COCO split of (Joseph et al. 2021). The comparison is presented in terms of unknown recall (U-Recall) and the previously known (Prev), current known (Cur) and Overall (both) AP for all tasks. U-Recall is not reported for task-4 since all classes are known. Our SS-OWFormer with just 10% labeled data outperforms the SOTA OW-DETR with 50% labeled data on all tasks.

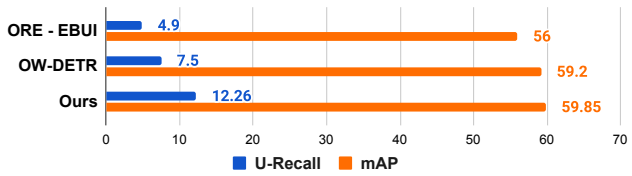

Figure 6: State-of-the-art comparison for OWOD Task-1. U-Recall reports the performance of the model in detecting unknown classes and mAP evaluates the performance of known classes. Owing to object query-guided pseudo-labeling our framework SS-OWFormer outperforms SOTA OW-DETR in terms of U-Recall and mAP.

Method	U-Recall	mAP
baseline	6.17	58.53
+ Enc_feature	8.06	58.56
SS-OWFormer	**12.26**	59.85
SS-OWFormer(Max)	11.25	59.51
SS-OWFormer(Mean)	**12.26**	59.85

Table 2: Ablation analysis of the impact on performance with various design choices for pseudo-labeling. The bottom section shows design choices for the selection of objections scores.

Model	Evaluation	mAP	U-Recall
Baseline	Task-1	64.9	2.5
	Task-2	68.1	-
SS-OWFormer	Task-1	66.7	**7.6**
	Task-2	**70.9**	-

Table 3: Comparison between the baseline and our SS-OWFormer on the proposed satellite OWOD splits. The comparison is shown in terms of mAP and unknown recall. The unknown recall metric assesses the model's ability to capture unknown object instances. While using full task-2 data, SS-OWFormer improves 5.1% in unknown recall over the baseline without compromising overall mAP.

Inference: The object queries for a test image I are obtained and the model predicts their labels from $\mathcal{K}^t+1$ classes along with a bounding box. A *top-k* selection with the highest scores is used for OWOD detection.

4 Experiments

Datasets: We evaluate our SS-OWOD framework on MS-COCO (Lin et al. 2014), Pascal VOC (Everingham et al. 2010), DOTA (Xia et al. 2018) and Objects365 (Shao et al. 2019) for OWOD problem. Non-overlapping tasks $\{T_1, T_2, ..., T_t, ...\}$ are formed from classes such that a class in T_λ is not introduced till $t = \lambda$ is reached as introduced in (Joseph et al. 2021). Our novel satellite OWOD split is prepared from DOTA based on the number of instances per image to ensure the corresponding representation of all categories.

Evaluation Metrics: The standard mean average precision (mAP) is the metric for known classes. However, mAP cannot be utilized as a fair metric to evaluate unknown detection since all possible unknowns are not labeled and can be more than given classes to be encountered in the next set of tasks. Therefore, we use average recall as a metric to test unknown object detection, as in (Bansal et al. 2018; Lu et al. 2016) under a similar context.

Implementation Details: The transformer architecture is a version of Deformable DETR (Zhu et al. 2020). Multi-scale feature maps are taken from ImageNet pre-trained ResNet50 (He et al. 2016; Zhang et al. 2022). Number of queries is set to $M = 250$ to account for the high number of instances in satellite images, while the threshold for the selection of pseudo-labels is set to top-10. Training is carried out for 50 epochs using ADAM optimizer (Kingma and Ba 2014) with weight decay (AdamW) and learning rate set to 10^{-4}.

4.1 State-of-the-art Comparison

We show a comparison with previous works (Joseph et al. 2021; Gupta et al. 2022) in OWOD for splits on MS COCO in Tab.1. The qualitative results can be seen in Fig.4. The

Partial Annotation	SSL	mAP		
		Overall	Previously Known	Current Known
100%	✓	70.9	77.1	61.6
75%	✗	65.2	74.8	50.9
	✓	69.04	76.2	58.3
50%	✗	63	74.5	45.8
	✓	68.06	75.1	57.5

Table 4: Comparison of results with and without semi-supervised learning on proposed satellite OWOD splits. This demonstrates our semi-supervised incremental learning strategy at different proportions of partially annotated data with a steady improvement over baseline under similar settings. The semi-supervised learning pipeline enables us to take advantage of the unannotated data while maintaining the performance on previously known classes.

comparison is made in terms of known class mAP and unknown class recall (U-Recall). U-Recall quantifies the model's capacity to retrieve unknown object instances in the OWOD setting. It should be noted that, since all classes are known in task-4, U-Recall cannot be reported. For a fair comparison, we omit ORE's energy-based unknown identifier (EBUI) since that relies on a held-out validation set. Our contributions prove useful in a fully supervised setting for task-1 as depicted above in Fig.6. Compared to the state-of-the-art method OW-DETR, our SS-OWFormer with merely 10% achieves an absolute gain of up to 5% in unknown recall for task-2 and task-3 owing to the object query-guided pseudo-labeling. Furthermore, our SS-OWFormer with just 10% labeled data outperforms state-of-the-art OW-DETR trained with 50% labeled data in terms of mAP for all the tasks, while SS-OWFormer with 50% labeled data stands comparable to fully supervised state-of-the-art OW-DETR. This poses SS-OWFormer as closer to a realistic solution by overcoming the requirement of fully supervised incremental learning for the OWOD problem.

4.2 Ablation Studies

Tab. 2 shows improvements made in performance with different components. Just using encoder features for pseudo-labeling instead of backbone features gives a 2% improvement in Unknown Recall over the baseline. Our proposed object query-guided pseudo-labeling helps to gain another 4.2% in Unknown Recall and reach 12.26 by utilizing decoder queries finally providing a relative gain of nearly doubling over the baseline in terms of Unknown Recall. Other design choice trials for the pseudo-labeling scheme include taking the mean of and maximum among the objectness scores. Mean is currently used in the proposed SS-OWFormer framework, while taking maximum causes a slight drop to unknown recall as it falls to 11.25.

4.3 Experiments on Satellite Images

Tab. 3 reports unknown recall (U-recall) for task 1 supervised training which assesses the model's capability to capture unknown objects for the OWOD-S split. The qualitative results for satellite images can be seen in Fig.5. Our baseline achieves an unknown recall of 2.5 on Task-1, with an mAP of 64.9 on our OWOD-S task-1 benchmark. On the same task, SS-OWFormer achieves an unknown recall of 7.6 and 66.7 mAP and an mAP of 70.9 for task-2. The object query-guided pseudo-labeling scheme feeds into the novelty classification and objectness branches which helps build a better separation of unknown objects from knowns and background in the satellite images. Tab.4 shows a comprehensive comparison of our semi-supervised incremental learning strategy at different proportions of labeled and unlabeled data with a steady improvement over baseline under similar settings. This consistent improvement shown under the limited annotation availability setting emphasizes the importance of proposed contributions in a close to realistic satellite OWOD scenario without drastic forgetting of previously known classes.

CJ	GB	GR	PS	SL	U-Recall	mAP
✗	✗	✗	✓	✓	9.57	37.68
✓	✓	✓	✓	✓	10.03	38.13
✓	✗	✗	✗	✗	10.06	38.71
✗	✓	✓	✗	✗	10.11	38.89
✓	✓	✓	✗	✗	10.56	39.20

Table 5: Performance comparison when using different combinations of augmentation techniques applied together. The augmentations are abbreviated as CJ - Color Jitter, GB - Gaussian Blur, GR - Greyscale, PS - Posterize, and SL - Solarize. All the experiments are run using a fixed seed of 42 for Task 2 50% labeled data.

4.4 Augmentation Techniques

Tab. 5 shows performance comparison when using different combinations of augmentation techniques. We use color jitter, gaussian blur, random greyscale, posterizing, and solarizing as augmentations for the semi-supervised open-world learning pipeline. From our experiments, we observe that posterizing and solarizing degrade the overall performance of the model as they are not well suited for our problem setting.

4.5 Incremental Object Detection

As shown in Tab. 7 our SS-OWFormer performs favourably compared to previous works on incremental object detection (iOD) task. iOD experiments are performed on Pascal VOC (Everingham et al. 2010) benchmark on the 10 + 10 class setting as proposed in (Joseph et al. 2021). Our SS-OWFormer achieves 65.2 mAP while using only 50% labeled data in the incremental learning setting compared with ORE and OW-DETR.

4.6 Objects365 Benchmark

Tab. 6 compares our SS-OWFormer with OW-DETR on the Objects365 (Shao et al. 2019) benchmark. Experiments on Objects365 with 365 object categories learned incrementally

Method	Task 1		Task 2		Task 3		Task 4		Task 5
	UR	mAP	UR	mAP	UR	mAP	UR	mAP	mAP
OW-DETR	13.6	21.2	17.8	18.6	12.7	16.7	17.8	16.3	15.8
Ours (50%)	16.7	23.3	21.6	18.1	14.9	15.9	18.7	15.4	15.1

Table 6: Comparison over the proposed OWOD splits on Objects365 dataset. Our splits derived from a subset of Objects365, comprise 100k images and five different tasks. Task-1 has 85 categories, while Tasks-2:5 have 70 categories each. To our knowledge, we are the first to report OWOD results on Objects365, and our method performs favorably compared to OW-DETR (Gupta et al. 2022)

Method	Avg of 10 Base classes	Avg of 10 Novel classes	mAP
ORE	60.37	68.79	64.5
OW-DETR	63.48	67.88	65.7
Ours (50%)	63.85	66.53	65.2

Table 7: Incremental detection results on PASCAL VOC (10+10 setting) as in (Joseph et al. 2021) averaged over base, novel classes & overall mAP.

in a semi-supervised manner shows the effectiveness of our approach in a close to realistic setting. Our method consistently improves over the baseline OW-DETR while using 50% labeled data and to the best our knowledge we are the first to report results on Objects365 in the Open-world object detection paradigm.

5 Relation to Prior Art

Semi-supervised and incremental object detection have been active research areas in computer vision, and recent works have achieved promising results. For semi-supervised object detection, methods such as S^4L (Zhai et al. 2019) and FixMatch (Sohn et al. 2020) have been proposed to leverage unlabeled data by exploring consistency regularization techniques. S^4L incorporates self-supervised learning with semi-supervised learning and achieved state-of-the-art performance on various datasets. FixMatch utilizes a mix of labeled and unlabeled data, achieving competitive results with fully supervised approaches. Other approaches like (Li et al. 2019; Qiao et al. 2023) introduce meta-learning to further enhance performance. For incremental object detection, methods like COCO-FUNIT (Saito, Saenko, and Liu 2020), iCaRL (Rebuffi et al. 2017), and NCM (Ristin et al. 2014) have been proposed to incrementally update the object detector model with new classes. COCO-FUNIT utilizes domain adaptation techniques for incremental learning, while iCaRL and NCM utilize exemplar-based methods for incremental feature learning. Other approaches like (Lee, Kim, and Yoon 2021; Kirsch, van Amersfoort, and Gal 2019; Sinha, Ebrahimi, and Darrell 2019; Yoo and Kweon 2019) utilize active learning and discriminative features to further enhance performance.

Open-world object detection in natural images recently gained popularity due to its applicability in real-world scenarios. ORE (Joseph et al. 2021) introduces an open-world object detector based on the two-stage Faster R-CNN (Ren et al. 2015). Since unknown objects are not annotated for training in the open-world paradigm, ORE utilizes an auto-labeling step to obtain a set of pseudo-unknowns for training. The OW-DETR (Gupta et al. 2022) introduces an end-to-end transformer-based framework for open-world object detection with attention-driven pseudo-labeling, novelty classification, and an objectness branch to triumph over the OWOD challenges faced by ORE. Methods like (Saito, Saenko, and Liu 2020; Rebuffi et al. 2017; Ristin et al. 2014; Perez-Rua et al. 2020) have been proposed to incrementally update the object detector model with new classes. OW-DETR achieved state-of-the-art performance on open-world object detection on the MS COCO benchmark. Localizing objects in satellite imagery(Xia et al. 2018; Waqas Zamir et al. 2019; Cheng et al. 2022) is a challenging task(Aleissaee et al. 2022; Van Etten 2018; Gong et al. 2022). The state-of-the-art results on DOTA (Xia et al. 2018) dataset is achieved by (Wang et al. 2022) by adapting the standard vision transformer to remote sensing domain using rotated window attention. To the best of our knowledge, open-world object detection has been focused on natural images and we are the first to propose an open-world object detection problem for satellite images.

6 Conclusion

We present SS-OWFormer, a framework aiming to reduce reliance on external oracles in the OWOD problem. SS-OWFormer comprises object query-guided pseudo-labeling to overcome limitations faced by heuristic approaches followed in previous works. We further explore a semi-supervised open-world object detection framework and introduce an OWOD-S split on DOTA. Experiments reveal the benefits of our contributions, leading to improvements for both known and unknown classes. Lastly, we validate our contributions in natural and remote sensing domains, achieving state-of-the-art OWOD performance.

Ethics Statement

In alignment with the AAAI Ethics Policy, we address the ethical dimensions of our work on Semi-Supervised Open-World Object Detection. We have conscientiously credited the data sources and other open source works on which SS-OWFormer is built upon.The open-world object detection problem is an intriguing real-world scenario that gradually learns additional objects. However, there may be circumstances in which a certain object or fine-grained category must not be identified because of privacy or legal issues, whether in satellite images or otherwise. Moreover, although

the proposed SS-OWFormer can incrementally learn new object categories, it does not have an explicit mechanism to forget some of the previously seen categories. Developing open-world object detectors with explicit forgetting mechanisms will be an interesting future research direction. Our commitment to transparency is evident through the availability of open-source resources, and we value collaboration and accountability within the research community. In recognizing the broader societal impact of our research, we pledge to uphold ethical standards in the development and deployment of our model and its applications.

Acknowledgements

The computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725, and by the Berzelius resource, provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

References

Aleissaee, A. A.; Kumar, A.; Anwer, R. M.; Khan, S.; Cholakkal, H.; Xia, G.-S.; et al. 2022. Transformers in Remote Sensing: A Survey. *arXiv preprint arXiv:2209.01206*.

Bansal, A.; Sikka, K.; Sharma, G.; Chellappa, R.; and Divakaran, A. 2018. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 384–400.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. arXiv:2005.12872.

Cheng, G.; Wang, J.; Li, K.; Xie, X.; Lang, C.; Yao, Y.; and Han, J. 2022. Anchor-free oriented proposal generator for object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–11.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338.

Fini, E.; da Costa, V. G. T.; Alameda-Pineda, X.; Ricci, E.; Alahari, K.; and Mairal, J. 2022. Self-Supervised Models are Continual Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv:1311.2524.

Gong, H.; Mu, T.; Li, Q.; Dai, H.; Li, C.; He, Z.; Wang, W.; Han, F.; Tuniyazi, A.; Li, H.; et al. 2022. Swin-Transformer-Enabled YOLOv5 with Attention Mechanism for Small Object Detection on Satellite Images. *Remote Sensing*, 14(12): 2861.

Gupta, A.; Narayan, S.; Joseph, K.; Khan, S.; Khan, F. S.; and Shah, M. 2022. OW-DETR: Open-world Detection Transformer. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Joseph, K. J.; Khan, S.; Khan, F. S.; and Balasubramanian, V. N. 2021. Towards Open World Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirsch, A.; van Amersfoort, J.; and Gal, Y. 2019. BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Lee, J.; Kim, E.; and Yoon, S. 2021. Anti-Adversarially Manipulated Attributions for Weakly and Semi-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4071–4080.

Li, X.; Sun, Q.; Liu, Y.; Zhou, Q.; Zheng, S.; Chua, T.-S.; and Schiele, B. 2019. Learning to Self-Train for Semi-Supervised Few-Shot Classification. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Lu, C.; Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2016. Visual relationship detection with language priors. In *European conference on computer vision*, 852–869. Springer.

Perez-Rua, J.-M.; Zhu, X.; Hospedales, T. M.; and Xiang, T. 2020. Incremental few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13846–13855.

Qiao, Z.; Wang, P.; Wang, P.; Ning, Z.; Fu, Y.; Du, Y.; Zhou, Y.; Huang, J.; Hua, X.-S.; and Xiong, H. 2023. A Dual-Channel Semi-Supervised Learning Framework on Graphs via Knowledge Transfer and Meta-Learning. *ACM Trans. Web*. Just Accepted.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. iCaRL: Incremental Classifier and Representation Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Ristin, M.; Guillaumin, M.; Gall, J.; and Van Gool, L. 2014. Incremental Learning of NCM Forests for Large-Scale Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Saito, K.; Saenko, K.; and Liu, M.-Y. 2020. COCO-FUNIT: Few-Shot Unsupervised Image Translation with a Content Conditioned Style Encoder. In Vedaldi, A.; Bischof, H.;

Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 382–398. Cham: Springer International Publishing. ISBN 978-3-030-58580-8.

Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8430–8439.

Sinha, S.; Ebrahimi, S.; and Darrell, T. 2019. Variational Adversarial Active Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 596–608. Curran Associates, Inc.

Van Etten, A. 2018. You only look twice: Rapid multi-scale object detection in satellite imagery. *arXiv preprint arXiv:1805.09512*.

Wang, D.; Zhang, Q.; Xu, Y.; Zhang, J.; Du, B.; Tao, D.; and Zhang, L. 2022. Advancing Plain Vision Transformer Towards Remote Sensing Foundation Model. *arXiv preprint arXiv:2208.03987*.

Waqas Zamir, S.; Arora, A.; Gupta, A.; Khan, S.; Sun, G.; Shahbaz Khan, F.; Zhu, F.; Shao, L.; Xia, G.-S.; and Bai, X. 2019. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 28–37.

Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2018. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3974–3983.

Yoo, D.; and Kweon, I. S. 2019. Learning Loss for Active Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, 12310–12320. PMLR.

Zhai, X.; Oliver, A.; Kolesnikov, A.; and Beyer, L. 2019. S4L: Self-Supervised Semi-Supervised Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. arXiv:2203.03605.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv preprint arXiv:2010.04159*.

Zong, Z.; Song, G.; and Liu, Y. 2023. DETRs with Collaborative Hybrid Assignments Training. arXiv:2211.12860.