ALWOD: Active Learning for Weakly-Supervised Object Detection

Yuting Wang[1], Velibor Ilic[2], Jiatong Li[1], Branislav Kisačanin[3,2], and Vladimir Pavlovic[1]

[1]Rutgers University, NJ, USA
[2]The Institute for Artificial Intelligence Research and Development of Serbia, Novi Sad, Serbia
[3]Nvidia Corporation, TX, USA
yw632@rutgers.edu, velibor.ilic@ivi.ac.rs, jiatong.li@rutgers.edu,
b.kisacanin@ieee.org, vladimir@cs.rutgers.edu

Abstract

Object detection (OD), a crucial vision task, remains challenged by the lack of large training datasets with precise object localization labels. In this work, we propose ALWOD, a new framework that addresses this problem by fusing active learning (AL) with weakly and semi-supervised object detection paradigms. Because the performance of AL critically depends on the model initialization, we propose a new auxiliary image generator strategy that utilizes an extremely small labeled set, coupled with a large weakly tagged set of images, as a warm-start for AL. We then propose a new AL acquisition function, another critical factor in AL success, that leverages the student-teacher OD pair disagreement and uncertainty to effectively propose the most informative images to annotate. Finally, to complete the AL loop, we introduce a new labeling task delegated to human annotators, based on selection and correction of model-proposed detections, which is both rapid and effective in labeling the informative images. We demonstrate, across several challenging benchmarks, that ALWOD significantly narrows the gap between the ODs trained on few partially labeled but strategically selected image instances and those that rely on the fully-labeled data. Our code is publicly available on https://github.com/seqam-lab/ALWOD.

1. Introduction

Object detection (OD) is a critical vision problem. To solve it, many fully-supervised object detection (FSOD) methods have been developed to build deep neural network architectures with high detection performance [35, 15, 3] and fast inference [33, 34]. Typically, these networks are trained on large fully-annotated (FA) data, which require humans to manually identify, with accurate bounding boxes and category labels, each object in an image. However, manual annotation of a large dataset is time-consuming [44], limiting the scalability of FSOD as the number of images, categories, and objects grows. To address this, many weakly-supervised object detection (WSOD) methods [2, 45, 19, 51] have been developed. WSOD aims to reduce the object annotation cost by leveraging cheaper, weakly-annotated (WA) data, where an image instance is tagged according to the objects present in it, without the need to specify bounding boxes for each object.

Existing WSOD methods often struggle to distinguish between object parts and objects, or between objects and groups of objects [36]. The performance of WSOD methods lags behind that of FSODs since WSODs rely on weaker annotation signals. Recently proposed semi-supervised [32, 43, 50] and few-shot [1, 29] learning approaches demonstrated that a good trade-off between annotation effort and detection performance can be achieved by first fully annotating a set of random images, followed by training the detector on a combination of large WA and small FA data. Active learning (AL) methods [53, 5, 54, 49] aim to further reduce the size of the FA sets using acquisition functions to select the most informative images for human labeling. AL methods can be either warm-start, which begin with a labeled set and iteratively select informative samples with feedback from the model, or cold-start, which select all informative samples at once without the need for an initial labeled set. Our work focuses on the warm-start setting.

To reduce the annotation cost and maximize the detection performance, we introduce an ***A**ctive **L**earning for **W**eakly-Supervised **O**bject **D**etection* (ALWOD) framework that combines semi-supervised learning with active learning by dynamically augmenting the semi-supervised set with a small set of actively selected and then fully annotated images, as illustrated in Fig. 1. However, traditional warm-start AL methods commence with an FSOD model trained on a random set of FA data [5, 8], typically hundreds or thousands of images, or a WSOD model trained on a large

set of WA data [49]. While the former results in effective learning of OD models, it still requires a significant annotator effort at initialization; the latter strategy is less effective and necessitates more rounds of AL. To circumvent this and further reduce the annotation cost, we design an image generator that leverages few FA images to synthesize a large auxiliary FA FSOD training set. Together with the WA data, the two sets are used for semi-supervised pre-training of an OD. The auxiliary FA data can serve as a warm-start for existing AL approaches, which require initial FA data.

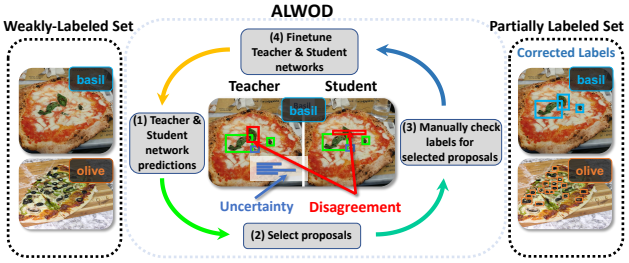


Figure 1: Overview of ALWOD. The teacher and student networks are first semi-supervised trained on WA data and auxiliary FA training data generated by an image generator, and then fine-tuned in successive stages using a few new well-selected FA and large remaining WA data. We propose a model disagreement and image uncertainty based acquisition function to select images, and build a new annotation tool to fully label selected images with decreased annotation workload compared to traditional AL labeling strategies.

The effectiveness of AL is predicated on the selection of the most informative samples for human labeling. To that end, we propose a novel active learning function that naturally leverages a combination of the model disagreement between student-teacher based ODs [25, 50] and the image uncertainty of the teacher network on the WA training data: images, where the network pairs disagree most and where the teacher exhibits high uncertainty, are passed on to human annotators for further labeling. This strategy allows us to select informative samples for both the low and the high-performing classes, particularly on class-unbalanced data.

To further reduce the AL cost, we replace the typical annotator task of drawing bounding boxes and assigning to each an object class label with the task of *correcting* the class labels of selected predicted bounding boxes, assigning bounding box quality scores, and removing falsely predicted bounding boxes. This type of annotation requires a significantly lower annotation effort while only slightly dampening the performance compared to the standard workload. The new FA and the remaining WA data will be used to fine-tune the student-teacher OD. This process is repeated until a desired annotation budget is met. As shown in Sec. 4, our combined AL and WSOD approach can achieve performance on par with FSOD, while reducing the need to precisely annotate large image datasets.

Our contributions are three-fold: (1) We present a new framework ALWOD to improve annotation efficiency and annotation quality by combining active learning with weakly and semi-supervised object detection paradigms. (2) We introduce a new acquisition function that considers the model disagreement between student-teacher networks, coupled with image uncertainty. (3) We propose an auxiliary domain to warm up the learning process utilizing small labeled sets. Our experimental results demonstrate that ALWOD achieves state-of-the-art performance across several benchmarks.

2. Related Work

Weakly-Supervised Object Detection (WSOD). WSOD methods generally aim to reduce detection annotation costs by exploiting only image-level annotations. Most existing methods mainly formulate WSOD as a multiple-instance learning (MIL) problem [2, 45, 11, 56, 36, 19, 18, 51]. Although many promising results have been achieved by WSOD, they are still not comparable to FSOD. Our work utilizes a small amount of FA data and a large amount of WA data based on active learning strategies and semi-supervised learning to achieve better performance.

Semi-Supervised Object Detection (SSOD). SSOD work focuses on training an object detector with a combination of FA, WA, or un-annotated (UA) data. A traditional paradigm of SSOD is to construct a multi-stage self-training pipeline [38, 43, 50]: (1) pre-train a model on FA data; (2) generate pseudo-labels on WA or UA data; (3) fine-tune the model on both FA and pseudo-labeled data; (4) repeat this process if needed. Some work [25, 50, 46, 26, 4, 23] relies on a student-teacher framework, where the teacher network generates pseudo-labels for the student network. Our work is also based on a student-teacher framework, but extends it with active learning strategies for selecting most informative WA or UA data, which are fully annotated by humans.

Active Learning for Object Detection (ALOD). The traditional active learning strategies [20, 10, 17, 13, 41] are designed for classification tasks. Few active learning methods specifically focus on object detection [8, 49, 5, 52, 53], which is more challenging with complex instance distributions. ALOD methods aim to improve detection performance by selecting the most informative images to be fully annotated by humans. They define an acquisition function used to assign a single score representing the informativeness of each weakly or unlabeled image. Most work [53, 5, 54] considers the instance-based uncertainty as the acquisition signal. The work of [21] introduces localization tightness and localization stability metrics to quantitatively evaluate the localization uncertainty of an object detector. The work of [5] proposes the aleatoric and epistemic uncertainty in both image and instance levels based on the Gaussian mixture model. Yoo *et al*. [53] propose the active learning method with the loss prediction module to predict

the loss of an input data point. Elezi *et al*. [8] introduce a class-agnostic active learning function based on the robustness of the network. In [52], the instance-level uncertainty and diversity are jointly considered in a bottom-up manner. Our work is related but different from the aforementioned methods. Similarly to these methods, we consider the image uncertainty as a part of the acquisition score. Unlike them, we also consider the model disagreement between student-teacher based ODs as a part of the acquisition function which is even more reliable for both the low and the high performing classes.

A key to making ALOD approaches effective is to appropriately initialize the OD model. Choi *et al*. [5] pre-train an FSOD model on a random set of fully-annotated data including 2,000 images, which requires a large annotation effort, and achieves 62.4% AP50 on VOC2007 dataset. Vo *et al*. [49] pre-train a WSOD model on weakly-annotated data with a small annotation effort, and achieves 47.7% AP50 on VOC2007 dataset. It is essential to balance initial detection performance with annotation cost. In contrast to traditional approaches, our OD is pre-trained on a large fully-labeled auxiliary domain constructed with minimal effort from only a few (as low as 50) fully-annotated images and a large weakly-labeled domain in a semi-supervised manner.

Annotation workflow. Traditional active learning approaches aim to query strong labels for data. However to reduce annotation costs, Desai *et al*. [6] first query weak localization information by requiring humans to click the centers of objects. This point information is stronger than the image-level tag. Pardo *et al*. [31] first decide the type of annotation for each selected image then optimizes the detection model on the hybrid supervised dataset. To reduce annotation cost while maintaining high annotation quality, our method allows one to select an imprecise bounding box for each object, a stronger label signal than the marked points.

3. Methodology

Our proposed object detection framework, `ALWOD`, aims to address the lack of accurate object localization information in the real-world training data by formulating WSOD as a combination of semi-supervision and active learning.

3.1. Preliminaries and Problem Statement

Consider an iterative, semi-supervised OD model learning setting where at each iteration $t = 1, 2, \ldots$ the model M^t is learned from a dynamic combination of weakly and fully labeled data. Let W^t, where $|W^t| = N - n$, and F^t, where $|F^t| = n$, denote the sets of indices of images in the training set $\mathcal{S}$ with the weak and full annotations, respectively, where N is the number of images in $\mathcal{S}$. An RGB image $\mathbf{X}_j \in \mathbb{R}^{h \times w \times 3}$, where h and w are its height and width, is said to be *fully annotated*, $j \in F^t$, if it is associated with the label $\mathbf{Y}_j^f = \{(\mathbf{b}_k, c_k, p_k)\}_{k=1}^{n^f}$ for each of the n^f objects present (labeled) in that image. The label consists of $\mathbf{b}_k \in \mathbb{R}^4$, the k-th object's localization bounding box defined by $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ that specifies its top-left corner $(x_{\min}, y_{\min})$ and its bottom-right corner $(x_{\max}, y_{\max})$. The label also contains the class label $c_k \in \{1, \ldots, C\}$, where C is the number of object categories, and the bounding box quality score $p_k \in \{1, 0\}$[1]. The same image is said to be *weakly annotated*, $j \in W^t$, if the image label contains only the classes of objects present in that image but not the objects' locations, *i.e*., $\mathbf{Y}_j^w = \{c_k\}_{k=1}^{n^w}$, where $1 \leq n^w \leq C$ is the number of object classes in that image. We denote this "version" of the dataset $\mathcal{S}^t := \mathcal{S}(W^t, F^t)$.

After model M^t is learned at cycle t from $\mathcal{S}(W^t, F^t)$, an active learning acquisition function $\alpha(\mathcal{S}(W^t, F^t), M^t)$ will select a set of B weakly annotated images with indices $A^{t+1} \subseteq W^t$, $|A^{t+1}| = B$, which will be passed on to a human annotator to label. The selection will be based on an assessment of model M^t's performance on $\mathcal{S}^t$ according to existing full and weak labels over $F^t \cup W^t$. In this fashion, we will arrive at an updated "version" $\mathcal{S}(W^{t+1}, F^{t+1})$, where $F^{t+1} = F^t \cup A^{t+1}$ and $W^{t+1} = W^t \setminus A^{t+1}$, which will have B more fully annotated images $|F^{t+1}| = |F^t| + B$ and B fewer weakly annotated images $|W^{t+1}| = |W^t| - B$ than the previous $\mathcal{S}^t$. This process is illustrated in Fig. 2.

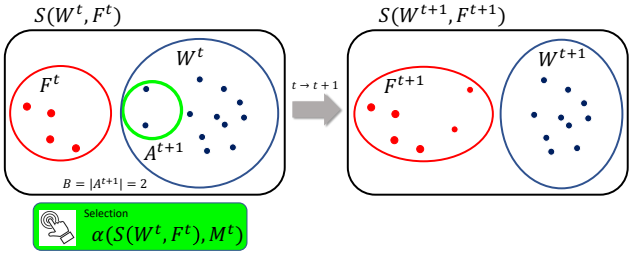


Figure 2: Training set notation for two steps of active learning for object detectors M^t.

Our goal here is to design the acquisition function $\alpha()$ and the semi-supervised learning approach for learning the OD models M^t to minimize the annotator workload proxied through the annotation budget $T \cdot B$, where T is the total number of cycles of active learning, while maximizing the final model detection accuracy. As a part of this process, we also aim to design the initialization procedure for model M^0. The following sections describe our proposed approach to achieving these goals.

3.2. Object Detection with Student-Teacher Networks

Fig. 3 gives an overview of our framework. Motivated by the recent success of student-teacher networks for

[1]The score corresponds to a subjective (annotator) notion of whether the (predicted) bounding box is precise, $p_k = 1 : \text{IoU} \geq 0.9$, or imprecise, $p_k = 0 : 0.5 < \text{IoU} < 0.9$. See Sec. 3.4.

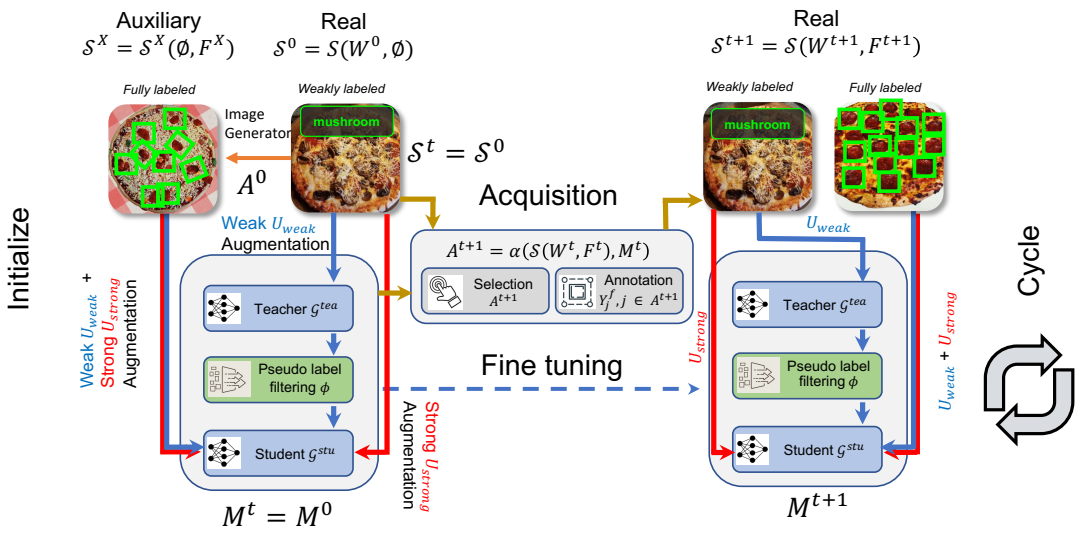


Figure 3: Architecture of ALWOD. The framework couples student-teacher WSOD and SSOD with active learning. The initial OD M^0 is pre-trained on $\mathcal{S}^0 \cup \mathcal{S}^X$ in a semi-supervised manner. The auxiliary $\mathcal{S}^X$ is created using an image generator and a small, fully-annotated A^0. At each cycle $t > 0$, we select B images using a novel acquisition function $\alpha(\cdot)$, which are passed on to annotators for full labeling using our annotation tool. The M^t is fine-tuned on the updated $\mathcal{S}^{t+1}$.

semi-supervised learning [25, 50] and transformer-based object detector [3, 37], our semi-supervised object detector is composed of a student detection network $\mathcal{G}^{stu}(\mathbf{X}|\theta_{stu})$ and a teacher detection network $\mathcal{G}^{tea}(\mathbf{X}|\theta_{tea})$. Both $\mathcal{G}^{stu}$ and $\mathcal{G}^{tea}$ are transformer-based object detectors or Faster-RCNN [35][2]. Thus, $M^t := \{\mathcal{G}^{stu}(\cdot|\theta^t_{stu}), \mathcal{G}^{tea}(\cdot|\theta^t_{tea})\}$, where θ^t_{stu} and θ^t_{tea} are the models' parameters at stage t of active learning.

3.2.1 Initial Learning ($t = 0$)

Assumptions: Our learning of the OD model begins with an assumption that the entire initial set of real-world images $\mathcal{S}$ is only tagged with weak labels, *i.e.*, $W^0 = \{1, \ldots, N\}, F^0 = \emptyset$, which we will denote $\mathcal{S}^0 := \mathcal{S}(W^0, F^0)$. This, while realistic, is an extremely challenging assumption and will lead to poor initial OD models and high annotation workload, *i.e.*, large B. To address this challenge, we make the second assumption that one can "cheaply" construct a large *auxiliary* "synthetic" fully-labeled dataset, which we denote $\mathcal{S}^X := \mathcal{S}^X(W^X = \emptyset, F^X = \{1, \ldots, N_X\})$. We will use $\mathcal{S}^0 \cup \mathcal{S}^X$ to initialize our student-teacher OD model[3].

Auxiliary set: Performance of OD learning critically depends on model initialization. It is thus essential to pre-train an FSOD model that achieves good detection performance with the least annotation cost. To that end, $\mathcal{S}^X$ is created by combining real-world background images with synthetically "pasted" but otherwise real-in-appearance foreground objects. These realistic foreground objects are cropped from the FA images in A^0, which are randomly selected from S^0. The annotation cost of this large fully-annotated auxiliary set is identical to that of the B images in A^0. However, as demonstrated in Sec. 4.2, the role of $\mathcal{S}^X$ is essential, when combined with $\mathcal{S}^0$, to learn an effective ALOD.

Burn-in: $\mathcal{G}^{stu}$ is first trained only on fully-labeled data $\mathcal{S}^X$. The teacher $\mathcal{G}^{tea}$ is initialized by duplicating the burn-in student model, $\theta_{tea} = \theta_{stu}$.

Student-teacher learning: We build upon the student-teacher learning paradigm of [50]. Therein, two types of augmentation (strong and weak) are used to regularize learning of the student-teacher network pair. Here, we train this pair using $\mathcal{S}^0 \cup \mathcal{S}^X$.

Specifically, we apply both weak $\mathrm{U}_{weak}(\cdot)$ and strong $\mathrm{U}_{strong}(\cdot)$ augmentation to the data, expanding the fully-labeled initial set to $\mathrm{U}_{weak}(\mathcal{S}^X) \cup \mathrm{U}_{strong}(\mathcal{S}^X)$ and the weakly labeled real-world set to $\mathrm{U}_{weak}(\mathcal{S}^0) \cup \mathrm{U}_{strong}(\mathcal{S}^0)$. We then train the student network using $\mathrm{U}_{weak}(\mathcal{S}^X) \cup \mathrm{U}_{strong}(\mathcal{S}^X)$ as well as the pseudo-labeled $\mathrm{U}_{strong}(S^0)$, with labels proposed by $\mathcal{G}^{tea}(\mathrm{U}_{weak}(S^0))$[4] and filtered by

[2]Since transformer-based object detectors are stronger than Faster-RCNN, we focus on transformer-based object detectors.

[3]Note that $\mathcal{S}^X$ is only used in the initial cycle $t = 0$.

[4]We use the following notation for brevity: $\mathrm{U}_a(\mathcal{S})$ means that the image component $\mathbf{X}$ of $(\mathbf{X}, \mathbf{Y}) \in \mathcal{S}$ is transformed by the augmentation operator $\mathrm{U}_a(\cdot)$, i.e., $\mathrm{U}_a(\mathcal{S}) := \{(\mathbf{X}', \mathbf{Y}') : \mathbf{X}' = \mathrm{U}_a(\mathbf{X}), \mathbf{Y}' = \mathrm{U}_a(\mathbf{Y}), \forall (\mathbf{X}, \mathbf{Y}) \in \mathcal{S}\}$. Augmentation of labels is applied as necessary to enforce label coherence, e.g., when $\mathrm{U}_a(\cdot)$ is L-R flip, the object class labels are maintained while the object locations are "flipped". Similarly, $\mathcal{G}^m(\mathcal{S})$ stands for the set constructed by replacing the label component $\mathbf{Y}$ of the $(\mathbf{X}, \mathbf{Y}) \in \mathcal{S}$ pair using the predictive model $\mathcal{G}^m(\cdot)$. I.e., $\mathcal{G}^m(\mathcal{S}) := \{(\mathbf{X}', \mathbf{Y}') : \mathbf{X}' = \mathbf{X}, \mathbf{Y}' = \mathcal{G}^m(\mathbf{X}), \forall (\mathbf{X}, \mathbf{Y}) \in \mathcal{S}\}$.

$\phi(\cdot)$:

$$\theta^0_{stu} \leftarrow \min \sum_{(\mathbf{X},\mathbf{Y})\in\mathcal{T}^0} \mathcal{L}(\mathcal{G}^{stu}(\mathbf{X}|\theta_{stu}), \mathbf{Y}), \quad (1)$$

where

$$\mathcal{T}^0 = \mathrm{U}_{weak}(\mathcal{S}^X) \cup \mathrm{U}_{strong}(\mathcal{S}^X) \cup \phi(\mathcal{G}^{tea}(\mathrm{U}_{weak}(S^0))), \quad (2)$$

and $\mathcal{L}$ is the classification and bounding box regression loss used in transformer-based detectors [3, 37]. Our pseudo-labeling filtering approach $\phi(\cdot)$ uses the bounding box quality score and the bounding box tag while [50] does not. For details of $\phi(\cdot)$, please refer to the Supplement **Sec. 1**.

The teacher network is updated by the exponential moving average (EMA) from the student network [48],

$$\theta_{tea} \leftarrow q\theta_{tea} + (1-q)\theta_{stu}, \quad (3)$$

where $q \in (0,1)$, set empirically to $q = 0.9996$. This EMA updated teacher can be seen as a temporal ensemble of student models along the training trajectories [50]. After training M^0, we set $F^0 \leftarrow F^0 \cup A^0$ and $W^0 \leftarrow W^0 \setminus A^0$.

In Sec. 3.3, we introduce a novel way of exploiting the model disagreement and the image uncertainty between the student and the teacher networks to identify informative samples, which are subsequently used to minimize the human annotation effort.

3.2.2 Active Learning Cycle ($t > 0$)

At each cycle $t > 0$, we select B images from the current weakly-annotated data W^t using our acquisition function $A^{t+1} = \alpha(\mathcal{S}(W^t, F^t), M^t)$. The human annotator will annotate images in this acquisition set A^{t+1} with bounding boxes, object class labels, and bounding box quality scores, helped by predictions generated from M^t. This results in a new set of full labels for images in this set, $\mathbf{Y}^f_j, j \in A^{t+1}$, replacing their existing weak labels. In this manner, we have created a new training set $\mathcal{S}^{t+1} = \mathcal{S}(W^{t+1}, F^{t+1})$, increasing by B images the fully labeled image set F^t to F^{t+1}. Since the new fully-annotated images may include imprecise bounding boxes b_k, where $p_k = 0$, we propose a new pseudo-labeling filtering method embodied in $\phi(\cdot)$. For an imprecise bounding box b_k, using the strategy of [50], we find the best matched predicted bounding boxes $\hat{b}_k$. We then generate a final pseudo-labeled bounding box by interpolating the coordinates of the imprecise and matched boxes. The precise bounding boxes coincide with the final pseudo-labeled boxes. $\mathcal{S}^{t+1}$ will be used to fine-tune the student-teacher pair $(\theta^t_{stu}, \theta^t_{tea})$, using the new pseudo-labeling filtering, resulting in an updated OD M^{t+1}. This cycle will repeat T times until the annotation budget $T \cdot B$ is met or the model converges in validation loss. The detection performance is evaluated with the teacher network.

The key to making this active learning process effective is to define an optimal acquisition function $\alpha(\cdot)$ and the annotation workflow. In the next section, we propose a novel approach to designing such a function together with a new annotation procedure.

3.3. Active Learning Strategies

We aim to select the most informative training samples A^{t+1} from $\mathcal{S}^t$, with a small annotation budget B, that will lead to the largest reduction in loss (improvement in detection performance), based on the trained student and teacher networks. For this, we consider the following two signals: (a) the disagreement between the teacher-student network pair, and (b) the uncertainty of predictions on each image. We first define the scores for each signal, then fuse them to arrive at the final acquisition function.

Model Disagreement. The EMA updated teacher behaves as a stochastic average of consecutive student models [47]. In an ideal case, the student network's predictions would be consistent with the teacher's predictions. This naturally leads to using the disagreement of predictions between student and teacher networks as one of the acquisition signals to create A^{t+1}. Specifically, we define the model disagreement acquisition score $\beta_{MD}(\cdot)$ on image $\mathbf{X}$ as

$$\beta_{MD}(\mathbf{X}|\mathcal{S}^t, M^t) := 1 - \frac{\sum_c AP_c(\mathcal{G}^{stu}(\mathbf{X})|\mathcal{G}^{tea}(\mathbf{X}))}{n^w}, \quad (4)$$

where $c \in \{c_k\}_{k=1}^{n^w}$ are the known classes of objects present in $\mathbf{X}$. This scores an image according to the value of the average precision per-class score of the student model predictions when treating the teacher model predictions as the "ground truth". The higher the score, the less agreement the models have on image $\mathbf{X}$, indicating that the image may be a plausible candidate for manual annotation.

Image Uncertainty. Another traditional signal that points toward the need to manually annotate an image is the class prediction uncertainty. Given the uncertainty for each prediction in an image, we define the uncertainty of the image by aggregating the score over all predicted objects. Here, we define the image uncertainty score as the maximum entropy $\beta_{IU}(\cdot)$ on image $\mathbf{X}$ for the n^f objects predicted by the teacher network:

$$\beta_{IU}(\mathbf{X}|\mathcal{S}^t, M^t) := \max_{k=1}^{n^f} H(c_k|\theta_{tea}), \quad (5)$$

where $H(c_k|\theta_{tea})$ represents the entropy over the distribution c_k of predictions generated from the teacher network. The higher the entropy, the more uncertain the model is about its prediction, indicating that the image may need to be manually annotated.

Acquisition Function. We propose the final acquisition function for each image, which fuses the model disagree-

ment and the image uncertainty signals:

$$\alpha_{\Sigma}(\mathcal{S}(W^t, F^t), M^t) \coloneqq \underset{j \in W^t}{\text{argmax}_{\text{B}}}\, \beta_{MD}(\mathbf{X}_j|\mathcal{S}^t, M^t) + \beta_{IU}(\mathbf{X}_j|\mathcal{S}^t, M^t), \quad (6)$$

or

$$\alpha_{\Pi}(\mathcal{S}(W^t, F^t), M^t) \coloneqq \underset{j \in W^t}{\text{argmax}_{\text{B}}}\, \beta_{MD}(\mathbf{X}_j|\mathcal{S}^t, M^t) \cdot \beta_{IU}(\mathbf{X}_j|\mathcal{S}^t, M^t), \quad (7)$$

which selects B images from the weakly labeled set W^t with the highest values of the total score. We empirically find that taking the product of the model disagreement and the image uncertainty scores performs better than taking the sum[5]. Intuitively, for each cycle t of active learning, we will be selecting those images where the student-teacher models disagree the most and the teacher model predictions are the most uncertain.

3.4. Annotation Procedure and Tool

In traditional OD annotation, for each image humans are asked to draw tight bounding boxes around the objects to be detected and then select categories for each bounding box, an expensive and tedious task [7, 9, 14]. For instance, [44] reports an average drawing box annotation time of 34.5 seconds (25.5 seconds for drawing one box and 9.0 seconds for verifying quality). Extreme clicking [30] relaxes the task of clicking on four extreme points of the object and requires about 7 seconds. Considering the trade-off between annotation efficiency and annotation quality, we develop an annotation tool that leverages our acquisition scores from Sec. 3.3 to improve the efficacy of labeling. An example of the tool's frontend is shown in Fig. 4.

Specifically, all images from A^{t+1} are presented to the user in an ordered list according to (7). Each image contains a large number of predicted bounding boxes with predicted class labels generated from both student and teacher networks. After applying non-maximum suppression and confidence threshold, some proposals with bounding boxes and class labels are given to the annotators. The annotators are asked to: (1) select the bounding boxes from proposals that overlap with true objects ($> 50\%$ IoU) and include at least one of the four extreme points (top, bottom, left-most, right-most), (2) correct the bounding box categories, and (3) assess the bounding box quality. The remaining unselected bounding boxes are removed. If there is no bounding box over an actual object, the annotators directly draw a tight bounding box and select the object label.

[5]We use $\texttt{ALWOD}_{\Sigma}$ to denote the model based on $\alpha_{\Sigma}(\mathcal{S}(W^t, F^t), M^t)$ and $\texttt{ALWOD}_{\Pi}$ to denote the model that uses $\alpha_{\Pi}(\mathcal{S}(W^t, F^t), M^t)$ active learning strategy. When not specified, ALWOD refers to $\texttt{ALWOD}_{\Pi}$.

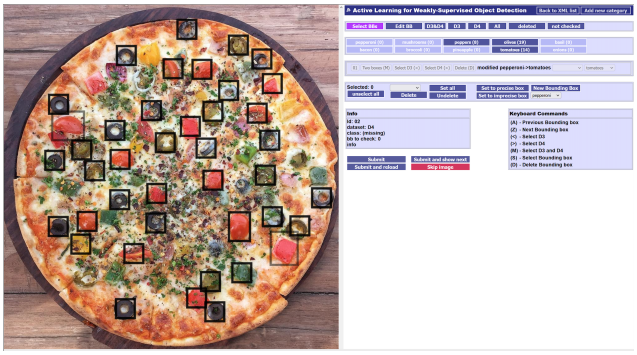


Figure 4: Web tool for manual checking, editing bounding boxes, and correcting classes.

Compared to [44, 30], our annotation process is significantly cheaper: on average, it can be completed in only 2 seconds, a reduction of over 70% over [30]. This results in a significantly decreased overall workload, as demonstrated in Sec. 4. For details of the annotation procedure, please refer to the Supplement **Sec. 3**.

4. Experimental Results

Datasets and Evaluation. We evaluate our method on three object detection benchmarks, VOC2007 [9], COCO2014 [22], and RealPizza10 [51]. Following previous works [49, 51], we use the trainval split of VOC2007 for training and the test split for evaluation, respectively containing 5,011 and 4,952 images. On COCO2014, we train detectors with the train split (82,783 images) and evaluate on the validation split (40,504 images). We use trainval split of RealPizza10 with 5,029 images for training and 552 test split images for testing. On RealPizza10, COCO2014, and VOC2007 datasets, each image contains 19.1, 7.7, and 2.5 instances on average, respectively. To evaluate the detection performance, we use the average precision metrics AP50 and AP, computed with the IoU threshold of $\tau = 0.5$ and the threshold set $\tau \in \{0.5, 0.55, \ldots, 0.95\}$, respectively. A predicted box is treated as a positive example when the IoU between the ground truth bounding box and the predicted object box exceeds τ.

Auxiliary Image Generator. The image generator creates auxiliary images in $\mathcal{S}^X$ by composing background images and object templates, as illustrated in Fig. 5. The object templates are created by cropping the object instances of fully-annotated images in A^0. Image augmentations such as rotation and scaling are performed on these templates, after which they are placed at random locations over the background images by employing a copy-paste augmentation technique sourced in [55, 12]. For details of the image generator, please refer to the Supplement **Sec. 2**.

Implementation Details. The VGG16 [42], ResNet50 [16], and Swin-T [27] models pre-trained

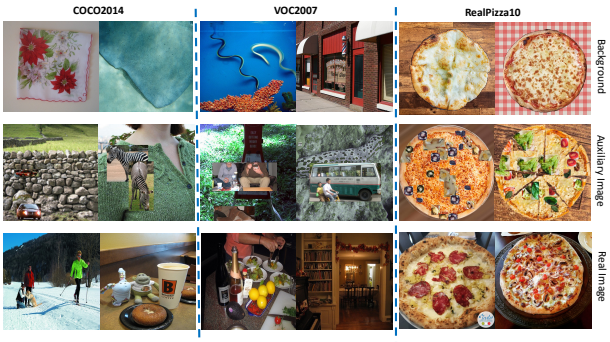


Figure 5: Examples of background images, auxiliary images, and real images on COCO2014, VOC2007, and RealPizza10 datasets.

on ImageNet [39] were used as backbones. We adopt Sparse DETR [37] or Faster-RCNN [35] for both $\mathcal{G}^{stu}$ and $\mathcal{G}^{tea}$. The confidence threshold for pseudo-labeling is 0.7. Following [25], for strong augmentations, we apply random color jittering, grayscale, Gaussian blur, and cutout patches. For weak augmentations, only random horizontal flipping is used. We report the results of "$L\%$" experiments, where about $L\%$ of the training weakly-labeled set are selected to be fully annotated. The minimal image height and width are set to 800 pixels. On VOC2007 and RealPizza10 datasets, $N_X = 2N$, and on COCO2014, $N_X = 0.5N$.

Baselines. We mainly focus on comparing against the state-of-the-art FSOD baselines: Faster-RCNN [35], SSD [24], and Sparse DETR [37]; WSOD baselines: WSDDN [2], OICR [45], C-MIDN [11], WSOD2 [56], MIST(+Reg) [36], CASD [19], W2N [18], and D2DF2WOD [51]; SSOD baselines: BCNet [29], OAM [1], Unbiased teacher v2 [26], and Label Matching [4]; and ALOD baselines: BAOD [31], SSDGMM [5], NALAE [8], Active Teacher [28], and BiB [49].

4.1. Main Results

In total, we annotate 5% of training images on VOC2007 and RealPizza10, 1% of training images on COCO2014. We perform five annotation cycles with a budget of $B = 50$ images per cycle on VOC2007 and RealPizza10, and $B = 160$ images per cycle on COCO2014.

We compare our method `ALWOD` to state-of-the-art FSOD ($n/N = 100\%$ FA data in $\mathcal{S}$), WSOD ($n/N = 0\%$ FA data in $\mathcal{S}$), and SSOD methods ($n/N = 10\%$ or $n/N = 5\%/1\%/5\%$ FA data in $\mathcal{S}$), where n is the total number of fully-labeled samples. Tab. 1 summarizes the detection results on different benchmarks.

Our method outperforms the ALOD baselines. As shown in Tab. 1, on VOC2007 our method reaches 68.0% AP50, outperforming SSDGMM [5] and BiB [49] by 39.2% and 7.4% absolute points, using the same number of fully-annotated images. On the more challenging COCO2014, our method reaches 38.4% AP50, outperforming BiB [49] by 5.2%. Using the same number of annotated training images, our detection performance benefits from the combination of semi-supervised learning and active learning compared with SSDGMM and WSOD-based BiB. On RealPizza10, our method outperforms SSDGMM [5] using 80% fully-annotated data by 14.5% using only 5% fully-annotated data. The auxiliary set is comparable with the large fully-labeled real data. **Our method also outperforms the SSOD baselines.** On VOC2007 in Tab. 1, we observe that our method improves by over $n/N = 10\%$ SSOD baselines, while using only $n/N = 5\%$ full-annotated data. **Our method also outperforms the WSOD baselines.** Compared to WSOD baselines, our method obtains significantly better results, across the three benchmarks, with only a small amount of full-annotated data. **Our method generalizes across different datasets,** performing particularly well on datasets that contain multiple and varied object instances per image, *i.e.*, RealPizza10 and COCO2014. **Our method generalizes to different backbones.** Our method approaches to FSOD state-of-the-art Sparse DETR [37] by 3.4% AP50 difference with only 5% fully-annotated training images on RealPizza10 using ResNet50 backbone. **Our method generalizes to different detectors.** We compare Faster-RCNN with Sparse DETR for both $\mathcal{G}^{stu}$ and $\mathcal{G}^{tea}$. They both perform similarly. Hence, the performance gain is largely not affected by the detector architecture. `ALWOD` shows a better trade-off between detection performance and annotation effort than FSOD, WSOD, and SSOD. Qualitative results can be found in Supplement **Sec. 5**.

4.2. Ablation Study

Impact of $\mathcal{S}^X$. We analyze the impact of our warm-start active learning approach using the auxiliary set $\mathcal{S}^X$, described in Sec. 3.2.1, on RealPizza10 and VOC2007 datasets. `ALWOD` first annotates randomly selected images $\mathcal{A}^0$, then creates $\mathcal{S}^X$ constructed from task-specific augmentations of these fully-labeled images $\mathcal{A}^0$. The initial student-teacher model is trained on weakly-annotated $\mathcal{S}^0$ and fully-annotated $\mathcal{A}^0$ or $\mathcal{S}^X$. In Tab. 2 $\mathcal{S}^X \cup \mathcal{A}^0 \cup \mathcal{S}^0$ denotes that a model is trained on the fully-annotated auxiliary images $\mathcal{S}^X$ coupled with weakly-annotated real images $\mathcal{S}^0$ in a semi-supervised manner, where $\mathcal{A}^0$ is not directly used for training, while used for creating $\mathcal{S}^X$. To our knowledge, this is the first use of this concept in active learning. We compare our framework with SSDGMM [5], and BiB [49] using different training sets. In our framework, we only use fully-labeled images $\mathcal{A}^0$ or auxiliary domain $\mathcal{S}^X$ to initialize the student and teacher networks. In SSDGMM [5], the initial FSOD model is only trained on fully-annotated images $\mathcal{A}^0$ or auxiliary domain $\mathcal{S}^X$. In BiB, we first pretrain the base weakly-supervised detector MIST [49] on $\mathcal{S}^0$. Then we fine-tune MIST either on $\mathcal{A}^0$ selected by BiB AL

Table 1: Results (AP50 and AP in %) for different methods in different settings on VOC2007, COCO2014, and RealPizza10. In FSOD and WSOD settings, each baseline considers the same fraction of the FA images across datasets. In SSOD and ALOD settings, different approaches consider different fractions of FA images in different benchmarks, *e.g.*, ALWOD and BiB consider 5% of FA data on VOC2007, 1% on COCO2014, and 5% on RealPizza10, respectively, denoted as $5\%/1\%/5\%$. Red figures denote the best-performing non-FSOD method, followed by the second-best in blue. * denotes reproduced results.

Setting	n/N	Backbone (Detector)	Method	VOC2007 AP50	COCO2014 AP50	COCO2014 AP	RealPizza10 AP50
FSOD	100%	VGG16	Faster-RCNN [35]	69.9	42.1	20.5	39.1
		VGG16	SSD [24]	68.0	42.1	24.1	32.8
		VGG16	Sparse DETR [37]	72.1	58.1	33.4	41.2
		ResNet50	Faster-RCNN [35]	74.1	59.1	38.4	40.2
		ResNet50	Sparse DETR [37]	88.4	65.8	45.5	42.7
		Swin-T	Sparse DETR [37]	90.2	69.2	48.2	43.8
WSOD	0%	VGG16	WSDDN [2]	34.8	11.5	-	-
			OICR [45]	41.2	-	-	4.7
			C-MIDN [11]	52.6	21.4	9.6	-
			WSOD2 [56]	53.6	22.7	10.8	-
			MIST(+Reg) [36]	54.9	24.3	11.4	-
			CASD [19]	56.8	26.4	12.8	12.9
			W2N [18]	65.4	-	-	-
			D2DF2WOD [51]	66.9	-	-	25.1
SSOD	10%	VGG16	BCNet [29]	61.8	38.3	22.9	35.4
	10%	VGG16	OAM [1]	63.3	-	-	-
	5%/1%/5%	ResNet50 (Faster-RCNN)	Unbiased teacher v2 [26]	61.2	37.5	22.3	32.8
	5%/1%/5%	ResNet50 (Faster-RCNN)	Label Matching [4]	61.7	37.7	22.4	34.1
ALOD	10%	VGG16 (Faster-RCNN)	BAOD [31]	50.9	-	-	-
	80%/8%/80%	VGG16(SSD)	SSDGMM [5]*	62.1	28.8	15.3	23.4
	80%/12.5%/80%	VGG16 (SSD)	NALAE [8]*	67.7	25.5	11.9	32.3
	5%/1%/5%	VGG16 (SSD)	SSDGMM [5]	28.8	8.7	4.3	16.4
		VGG16 (SSD)	NALAE [8]	36.3	8.8	3.4	22.2
		Resnet50 (Faster-RCNN)	Active Teacher [28]	49.7	33.5	18.0	30.8
		VGG16 (MIST [36])	BiB [49]*	60.6	33.2	16.5	15.7
		ResNet50 (Faster-RCNN)	ALWOD	69.1	39.8	24.5	38.0
		VGG16 (Sparse DETR)	ALWOD	68.0	38.4	23.7	37.9
		ResNet50 (Sparse DETR)	ALWOD	70.5	41.8	26.0	39.3
		Swin-T (Sparse DETR)	ALWOD	71.7	42.5	27.2	40.2

Table 2: Effectiveness of the auxiliary domain $\mathcal{S}^X$ on RealPizza10 and VOC2007. For RealPizza10, the set cardinalities are: $|\mathcal{A}^0| = 50$, $|\mathcal{S}^X| = 10,058$, and $|\mathcal{S}^0| = 5,029$. For VOC2007, the set cardinalities are: $|\mathcal{A}^0| = 50$, $|\mathcal{S}^X| = 10,022$, and $|\mathcal{S}^0| = 5,011$. To focus on the auxiliary domain, we reproduce all numbers by applying each training set to each method.

Backbone	Setting	Training set	AP50 (%)					
			RealPizza			VOC2007		
			SSDGMM	BiB	ALWOD	SSDGMM	BiB	ALWOD
VGG16	ALOD	$\mathcal{A}^0$	10.7	-	4.3	14.4	-	3.3
		$\mathcal{S}^0$	-	11.8	-	-	47.7	-
		$\mathcal{S}^X$	10.8	-	7.0	14.6	-	39.2
		$\mathcal{A}^0 \cup \mathcal{S}^0$	-	15.2	11.9	-	54.5	55.4
		$\mathcal{S}^X \cup \mathcal{A}^0 \cup \mathcal{S}^0$	-	14.0	18.8	-	50.6	60.1

strategy or $\mathcal{S}^X$ constructed from $\mathcal{A}^0$.

Tab. 2 shows the key role of $\mathcal{S}^X$ in lifting the initial detection performance of ALWOD from 11.9% to 18.8% in terms of AP50 on RealPizza10. As shown in Tab. 2 our approach significantly outperforms the initialization strategy of SSDGMM [5] by utilizing the auxiliary domain.

We also analyze the impact of our auxiliary domain $\mathcal{S}^X$ in the BiB framework. As shown in Tab. 2, in ALWOD $\mathcal{S}^X$ helps to initialize the student-teacher model. $\mathcal{S}^X$ improves the performance of MIST, while due to the domain gap between $\mathcal{S}^X$ and $\mathcal{S}^0$, the improvement is less than the improvement using $\mathcal{A}^0$.

Comparison of active learning strategies. We investigate the performance of our semi-supervised framework on three benchmarks under different acquisition functions using our proposed annotation strategy. We complete $T = 5$ annotation cycles with a budget of $B = 50$ images per cycle on VOC2007 and RealPizza10, $B = 160$ images

per cycle on COCO2014 datasets. We report the performance using the average of AP50 over three repetitions. As shown in Fig. 6, across the three benchmarks, our acquisition method $\texttt{ALWOD}_{\Pi}$ consistently achieves improvement over other strategies. The performance of BiB trails our $\texttt{ALWOD}_{\Pi}$, since BiB strategy aims to select the "best" training samples to "fix" the mistakes of the base weakly-supervised detector [49], while ALWOD disagreement-based AL strategy leverages the key property of the student-teacher networks. Entropy-sum obtains significantly worse results than other strategies on RealPizz10 and COCO2014 datasets as shown in Fig. 6a and Sec. 4.2. Core-set [40] and loss [53] underperform uniform sampling. The performance of uniform sampling is always worse than our proposed functions in each AL cycle. At $t = 3$, we see a significant improvement compared with the previous steps. As shown in Fig. 6, our product strategy $\texttt{ALWOD}_{\Pi}$ for the final fused acquisition function exceeds the sum strategy $\texttt{ALWOD}_{\Sigma}$. Per class results can be found in Supplement **Sec. 4**

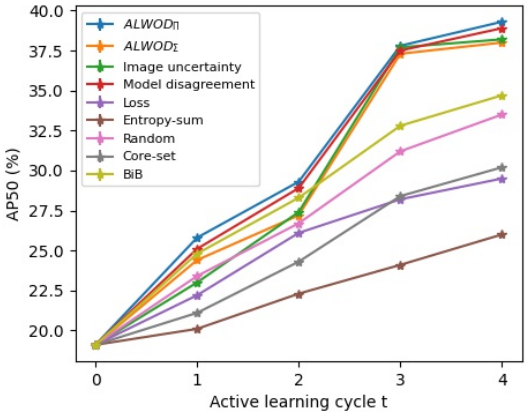
Comparison of annotation strategies. As shown in Tab. 3, our final fused acquisition function α_{Π} combined with our "selecting the box" annotation strategy obtains comparable results to the traditional "drawing the box" annotation strategy by only a 0.6% difference in terms of AP50.

Table 3: Ablation study on affect annotation tools based on our semi-supervised framework on RealPizza10 dataset using ResNet50 backbone.

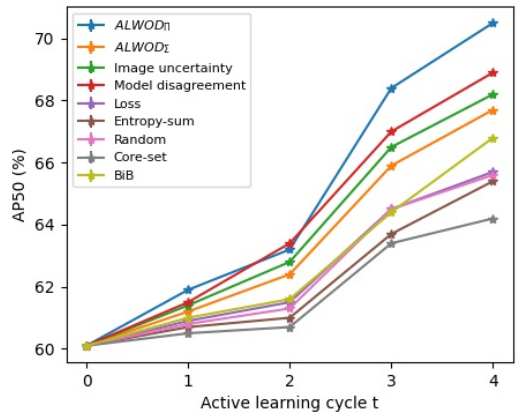
Labeling Cycle (t)	AP50 (%) 0	1	2	3	4
Drawing	19.1	27.1	37.3	38.8	39.9
Selecting (ALWOD)	19.1	25.8	29.3	37.7	39.3

5. Discussion and Conclusion

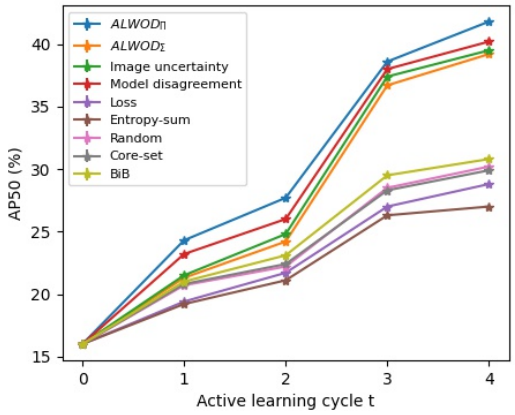
We propose a new approach to boost the detection performance of weakly-supervised object detectors by combining semi-supervised learning and active learning. We introduce a simple yet effective image generator to create large auxiliary fully-annotated data by leveraging few fully-annotated real images to warm-start active learning. Our framework introduces a novel acquisition function based on the fusion of the student-teacher OD model disagreement and the traditional image uncertainty, combined with an effective, low-effort annotation strategy. Empirical evaluations show that our method significantly outperforms the state-of-the-art on several key benchmarks and is particularly adept at tackling challenging multi-object, multi-class scenarios such as those in COCO2014 and RealPizza10 datasets. **Limitation.** Our annotation strategy requires the framework to automatically select certain object proposals on each image for manual checking. The default selection



(a) RealPizza10



(b) VOC2007



(c) COCO2014

Figure 6: Detection performance across different active learning strategies in our framework on the three benchmarks using ResNet50 backbone.

criteria may either introduce noisy or false positive bounding boxes or ignore bounding boxes with true objects in them; this may negatively affect the annotation quality or the annotation speed and, subsequently, the OD accuracy.

Acknowledgement: This material is based upon work supported by NSF IIS Grant #1955404.

References

[1] Carlo Biffi, Steven McDonagh, Philip Torr, Aleš Leonardis, and Sarah Parisot. Many-shot from low-shot: Learning to annotate using mixed supervision for object detection. In *ECCV*, 2020. 1, 7, 8

[2] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016. 1, 2, 7, 8

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1, 4, 5

[4] Binbin Chen, Weijie Chen, Shicai Yang, Yunyi Xuan, Jie Song, Di Xie, Shiliang Pu, Mingli Song, and Yueting Zhuang. Label matching semi-supervised object detection. In *CVPR*, 2022. 2, 7, 8

[5] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, and Jose M Alvarez. Active learning for deep object detection via probabilistic modeling. In *ICCV*, 2021. 1, 2, 3, 7, 8

[6] Sai Vikas Desai, Akshay L Chandra, Wei Guo, Seishi Ninomiya, and Vineeth N Balasubramanian. An adaptive supervision framework for active learning in object detection. *arXiv preprint arXiv:1908.02454*, 2019. 3

[7] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *CVPR*, 2009. 6

[8] Ismail Elezi, Zhiding Yu, Anima Anandkumar, Laura Leal-Taixe, and Jose M Alvarez. Not all labels are equal: Rationalizing the labeling costs for training object detection. In *CVPR*, 2022. 1, 2, 3, 7, 8

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 2010. 6

[10] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. 2

[11] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *ICCV*, 2019. 2, 7, 8

[12] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*, 2021. 6

[13] Jiannan Guo, Haochen Shi, Yangyang Kang, Kun Kuang, Siliang Tang, Zhuoren Jiang, Changlong Sun, Fei Wu, and Yueting Zhuang. Semi-supervised active learning for semi-supervised models: Exploit adversarial examples with graph-based virtual labels. In *ICCV*, 2021. 2

[14] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 6

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[17] Siyu Huang, Tianyang Wang, Haoyi Xiong, Jun Huan, and Dejing Dou. Semi-supervised active learning with temporal output discrepancy. In *ICCV*, 2021. 2

[18] Zitong Huang, Yiping Bao, Bowen Dong, Erjin Zhou, and Wangmeng Zuo. W2N: Switching from weak supervision to noisy supervision for object detection. In *ECCV*, 2022. 2, 7, 8

[19] Zeyi Huang, Yang Zou, BVK Kumar, and Dong Huang. Comprehensive attention self-distillation for weakly-supervised object detection. In *NeurIPS*, 2020. 1, 2, 7, 8

[20] Velibor Ilić and Jovan Tadić. Active learning using a self-correcting neural network (alscn). *Applied Intelligence*, 2022. 2

[21] Chieh-Chi Kao, Teng-Yok Lee, Pradeep Sen, and Ming-Yu Liu. Localization-aware active learning for object detection. In *ACCV*, 2018. 2

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 6

[23] Liang Liu, Boshen Zhang, Jiangning Zhang, Wuhao Zhang, Zhenye Gan, Guanzhong Tian, Wenbing Zhu, Yabiao Wang, and Chengjie Wang. Mixteacher: Mining promising labels with mixed scale teacher for semi-supervised object detection. In *CVPR*, 2023. 2

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 7, 8

[25] 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. *arXiv preprint arXiv:2102.09480*, 2021. 2, 4, 7

[26] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *CVPR*, 2022. 2, 7, 8

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 6

[28] Peng Mi, Jianghang Lin, Yiyi Zhou, Yunhang Shen, Gen Luo, Xiaoshuai Sun, Liujuan Cao, Rongrong Fu, Qiang Xu, and Rongrong Ji. Active teacher for semi-supervised object detection. In *CVPR*, 2022. 7, 8

[29] Tianxiang Pan, Bin Wang, Guiguang Ding, Jungong Han, and Jun-Hai Yong. Low shot box correction for weakly supervised object detection. In *IJCAI*, 2019. 1, 7, 8

[30] Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. Extreme clicking for efficient object annotation. In *ICCV*, 2017. 6

[31] Alejandro Pardo, Mengmeng Xu, Ali Thabet, Pablo Arbeláez, and Bernard Ghanem. Baod: budget-aware object detection. In *CVPR*, 2021. 3, 7, 8

[32] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *CVPR*, 2018. 1

[33] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 1

[34] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017. 1

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1, 4, 7, 8

[36] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *CVPR*, 2020. 1, 2, 7, 8

[37] Byungseok Roh, JaeWoong Shin, Wuhyun Shin, and Saehoon Kim. Sparse detr: Efficient end-to-end object detection with learnable sparsity. In *ICLR*, 2022. 4, 5, 7, 8

[38] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. *Carnegie Mellon University. Journal contribution*, 2005. 2

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 7

[40] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018. 9

[41] Oriane Siméoni, Mateusz Budnik, Yannis Avrithis, and Guillaume Gravier. Rethinking deep active learning: Using unlabeled data at model training. In *ICPR*, 2021. 2

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLP*, 2015. 6

[43] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 1, 2

[44] Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing annotations for visual object detection. In *AAAI*, 2012. 1, 6

[45] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *CVPR*, 2017. 1, 2, 7, 8

[46] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. In *CVPR*, 2021. 2

[47] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 2017. 5

[48] A Tarvainen and H Valpola. Weight-averaged consistency targets improve semi-supervised deep learning results. corr abs/1703.01780. *arXiv preprint arXiv:1703.01780*, 2017. 5

[49] Huy V Vo, Oriane Siméoni, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, and Jean Ponce. Active learning strategies for weakly-supervised object detection. In *ECCV*, 2022. 1, 2, 3, 6, 7, 8, 9

[50] Pei Wang, Zhaowei Cai, Hao Yang, Gurumurthy Swaminathan, Nuno Vasconcelos, Bernt Schiele, and Stefano Soatto. Omni-detr: Omni-supervised object detection with transformers. In *CVPR*, 2022. 1, 2, 4, 5

[51] Yuting Wang, Ricardo Guerrero, and Vladimir Pavlovic. D2DF2WOD: Learning object proposals for weakly-supervised object detection via progressive domain adaptation. In *WACV*, 2023. 1, 2, 6, 7, 8

[52] Jiaxi Wu, Jiaxin Chen, and Di Huang. Entropy-based active learning for object detection with progressive diversity constraint. In *CVPR*, 2022. 2, 3

[53] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *CVPR*, 2019. 1, 2, 9

[54] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *CVPR*, 2021. 1, 2

[55] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 6

[56] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. WSOD2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In *ICCV*, 2019. 2, 7, 8