BAOD: Budget-Aware Object Detection

Alejandro Pardo *[2], Mengmeng Xu *[2], Ali Thabet[2], Pablo Arbeláez[1], and Bernard Ghanem[2]

[1]Universidad de los Andes, Colombia
[2]King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Abstract

We study the problem of object detection from a novel perspective in which annotation budget constraints are taken into consideration, appropriately coined Budget Aware Object Detection (BAOD). When provided with a fixed budget, we propose a strategy for building a diverse and informative dataset that can be used to optimally train a robust detector. We investigate both optimization and learning-based methods to sample which images to annotate and what type of annotation (strongly or weakly supervised) to annotate them with. We adopt a hybrid supervised learning framework to train the object detector from both these types of annotation. We conduct a comprehensive empirical study showing that a handcrafted optimization method outperforms other selection techniques including random sampling, uncertainty sampling and active learning. By combining an optimal image/annotation selection scheme with hybrid supervised learning to solve the BAOD problem, we show that one can achieve the performance of a strongly supervised detector on PASCAL-VOC 2007 while saving 12.8% *of its original annotation budget. Furthermore, when* 100% *of the budget is used, it surpasses this performance by* 2.0 *mAP percentage points.*

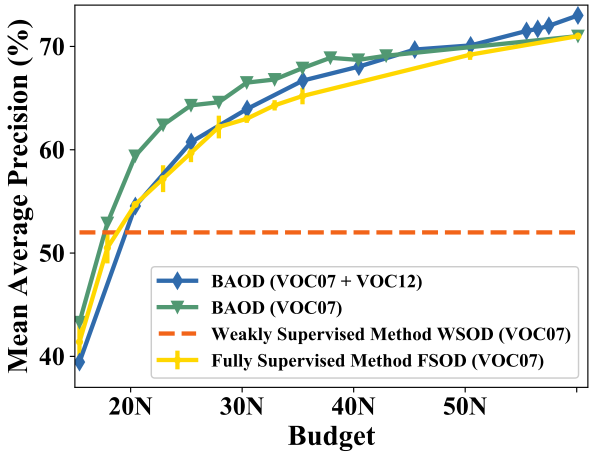

Figure 1: **Budget-Aware performance of detectors with different levels of supervision.** The models are trained on PASCAL VOC 2007 trainval (VOC07), PASCAL VOC 2012 trainval (VOC12) as specified in the legend. Our proposed Budget aware object detection (BAOD) (green -∇- curve) has a higher mAP than FSOD (yellow -|- curve) and WSOD (orange - - curve) methods at most budgets. Given a larger unlabeled image pool (Blue -◇- curve, VOC07+VOC12), our BAOD can reach a higher mAP using the same budget needed to annotate VOC07 with instance-level labels. Since the dataset is finite, WSOD cannot increase its performance with more budget.

1. Introduction

Object detection in images is a fundamental computer vision problem with applications in many tasks including face/pedestrian detection [43, 1, 20, 40, 11, 56], counting [5, 22], and visual search [9, 38]. Building a successful object detector encompasses three main dimensions: (1) **the image dataset** to be annotated for training the detector. A larger dataset allows for a more accurate detector, but the number of training images is limited by the annotation budget; (2) **the annotation scheme** used to label the training images. One could annotate either image-level labels (the categories of the objects are known but their locations are unknown, denoted weakly supervised annotation) or instance-level labels (both categories and locations are known, denoted strongly supervised annotation) [40, 34, 28, 39]; (3) **the detection model**. Most works on object detection fix the first two dimensions and only explore the third. In fact, they tend to focus on optimizing the detection model based on one or more training datasets that typically provide the same kind of annotations. In this paper, we fix the detection model and investigate solutions in the first two dimensions of the problem.

*denotes equal contribution

A large group of object detectors fall under the umbrella of Fully-Supervised Object Detection (FSOD) [44, 45, 14, 13, 10, 46, 17, 33]. It has been shown in recent years that these techniques can reach high detection performance, especially with the introduction of large datasets with strong annotations [51, 12, 34]. This requirement makes FSOD methods expensive and time consuming. In contrast, Weakly-Supervised Object Detection (WSOD) [3, 8, 23, 4, 62, 15, 31, 19, 48, 54, 21] aims at building object detectors from cheaper but less informative image-level or weak annotations.

In this paper, we propose a trade-off between dataset annotation cost and model precision in order to combine both weak and strong annotations and train an object detector with hybrid supervision. We put all the detectors on the same footing when different annotation schemes are available. Ideally, detectors should only be compared when they are trained using image datasets that offer the same amount of information (not necessarily the same number of images). Since this notion is difficult to define quantitatively, we take the **training budget** of a detector as a unifying surrogate measure. Here, we define budget as the effort, or cost, to annotate a dataset, thus combining the first two dimensions of the detection problem: dataset scale and annotation scheme. In fact, with a fixed budget and a set of unlabeled images, the number of images we can label depends highly on the annotation cost for each image. This cost varies significantly between image- and instance-level annotations. Typically, annotating a bounding box around an object in an image is significantly more expensive than simply annotating its category [40, 51]. Therefore, an FSOD method with the same budget as a WSOD method contains fewer images in its training set.

We explore strategies to build better object detector models when constrained with a training budget, a novel problem we coin *budget-aware object detection* (BAOD). As such, we focus on choosing the best images for training and how to annotate them. For this, we survey several selection methods to sequentially choose both the image and type of annotation following an active learning paradigm. Additionally, we propose a novel hybrid training procedure for object detectors that can use both strong (instance-level) *and* weak (image-level) annotations. Figure 1 shows that actively selecting images and their annotations to sequentially train a hybrid supervised detector outperforms its FSOD and WSOD counterparts, when the budget is larger than 20% of the budget needed to annotate VOC2007 trainval at the instance-level. Moreover, the yellow curve shows that we can reach an even higher mAP if we use the same budget of VOC07 to annotate images from both datasets VOC07 and VOC12.

Contributions. **(1)** We propose the **BAOD** problem and present a new evaluation criterion **Budget-Average mAP** for object detection algorithms. This criterion takes into account both detection performance and budget. **(2)** We study several strategies (*e.g.* optimization and learning-based) to select both training images and their annotation scheme. **(3)** We propose a hybrid supervised learning strategy that combines category and localization information to train a robust detection model that handles both weak and strong annotations. **(4)** Following a BAOD approach (*i.e.* combining intelligent image and annotation scheme selection with hybrid supervised learning), we show that the mAP test performance on VOC07 can be improved by 2 percentage points for the same budget used to annotate the training set of VOC07 (by combining it with VOC12). We also show the opposite, *i.e.* that state-of-the-art test performance on VOC07 can be achieved, while saving 12.8% of the budget used in strongly annotating its training set.

2. Related Work

2.1. Fully Supervised Object Detection (FSOD)

With the development of deep learning, many CNN based methods have been proposed to solve the FSOD problem, such as YOLO [45], SSD [36], RCNN [14], and its variants [46, 13, 10, 17, 33]. Although these methods achieve impressive detection results, providing them with large-scale instance-level (bounding-box) annotations for training is costly.

In this paper, we constrain the annotation budget and focus on how to reach the best detection performance within this budget by intelligently selecting between both weak and strong annotations. Su *et al.* [53] reported that it takes 26 seconds to draw one bounding-box without quality control, and 42 seconds with it. There has been recent progress in developing tools to further reduce annotation time (most recently to 7 seconds) [39, 40, 28]. Our work investigates a complementary aspect of annotation to emphasize that some images do not need to be strongly annotated to achieve excellent performance. While our discussion of budget aware object detection only considers the labor cost for image annotation, we note that sequentially training detection models also consume computational resources. However, given the continual development of hardware acceleration and re-organization, which leads to steady decrease in their cost, manual annotation remains the more expensive component in the detection problem.

2.2. Weakly Supervised Object Detection (WSOD)

If there are no instance-level labels available at training time, a WSOD method can be trained. Most classical approaches cast WSOD as a Multiple Instance Learning (MIL) problem [52, 31, 19, 48, 3, 21]. Bilen *et al.*[4]

were the first to utilize a deep CNN (denoted as the weakly supervised deep detection network or WSDDN) to solve this problem. WSDDN selects positive samples by multiplying the score of recognition and detection. Zhang *et al.*[59] proposed a simple yet effective post-processing step to mine pseudo ground truth bounding boxes used for iterative FSOD training.

We use weak supervision to improve pseudo object labels. These labels are generated from a previously trained detector, and are post-processed through a pseudo label mining process. The image-level labels help remove false predictions that correspond to a wrong object category.

2.3. Hybrid Supervised Learning

In our work, we study hybrid supervised learning for object detection. This type of learning exploits multiple types of supervision during training. Semi-supervised learning is a special case of this family, since it learns a model from a set of labeled and unlabeled data. Several works [32, 49, 6, 29, 42, 57, 60] try to solve the generic semi-supervised learning problem through *teacher-student learning*. They first train a *teacher* model from the strongly annotated subset, then the predictions obtained from the teacher model on the unlabeled data are used to regress a second model that is called the student model. Chéron *et al.* [7] also used several kinds of annotations to train an activity recognition model, and they revealed that strongly annotating every training sample is not necessary to achieve noteworthy localization results in the video domain.

In our case, since images can be labeled either weakly or strongly, a hybrid supervised dataset is always considered. Inspired by Rosenberg's work [42], we train a teacher- student model to use the hybrid dataset. Our teacher detector is learned from strong annotations only, while the student detector is learned from both ground-truth and/or processed pseudo object labels.

2.4. Active Learning

In general, active learning is a sequential decision making process that iteratively selects the most useful examples an oracle should annotate and add to the labeled training set. It aims at training more accurate models with the minimum data required. This field has been widely studied in the context of image and video classification [41, 2, 27, 25, 61, 16], object detection [50, 47, **?**, 28, 26, 39, 24], action localization [18], human pose estimation [35] and visual question answering [37]. A commonly used approach to selecting new training images is by means of their entropy score [55]. The intuition is that higher entropy examples attribute to more learning information. More recent research directly predicts the improvements of adding a new sample to the training set, and uses this measurement as a selection criteria [27]. As specified in previous work, active learning aims at maximizing performance while minimizing the human cost in labeling the training samples [41, 35, 18].

The BAOD problem can also be formulated as a sequential decision making process. We study a few well-known selection techniques in our new active learning pipeline, which contains two active processes: (1) select the next batch of training images to annotate and (2) decide the type of annotation for each selected image. Thus, we focus on the annotation sequence that can provide the most useful information to the detector. To the best of our knowledge, we are the first to use active learning and hybrid training to study object detection.

3. Budget Aware Object Detection (BAOD)

In this section, we explore several ways to sequentially create an object detection dataset with a fixed budget, which includes both image-level and instance-level annotations. Then, we introduce a hybrid supervised learning procedure to take advantage of this hybrid labeled training set. A complete overview of the whole active learning procedure is shown in Figure 2.

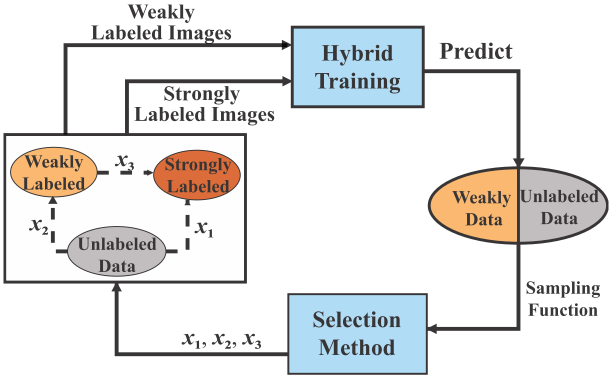

Figure 2: **Overview of active learning pipeline to construct a hybrid labeled dataset.** For any weakly labeled or unlabeled image in the image pool (circular shapes), the selection method (bottom blue rectangle) decides which type of action to apply on the image based on the sample function and image status: weakly label (x_1) or strong label (x_2, x_3). Then such image is appended into the hybrid dataset and we train an object detection model with the hybrid supervision.

3.1. Hybrid Dataset using Active Learning

Our hybrid supervised dataset should consider which images to include in training and how to label them. To tackle this problem, we propose an active hybrid learning framework which acts as an active learning agent to simulate the annotation process. At every active step t, we have a budget constraint d to spend on annotations. Each image in this set belongs to one of three pools: unlabeled U_t, weakly labeled W_t, or strongly labeled S_t. At each active step, the hybrid supervised dataset $W_t \cup S_t$ is used to train a hybrid detector, which will be described in detail in Section 3.2. This detector is used to find a selection function that takes an action on U_t or W_t pools. As shown in Figure 2, we have three possible actions with their associated cost: (1) annotate strongly $\boldsymbol{x}_1$ (*i.e.* send the image from U_t to S_t), (2) annotate weakly $\boldsymbol{x}_2$ (*i.e.* send the image from U_t to W_t), or (3) strongly annotate a weakly annotated image $\boldsymbol{x}_3$ (*i.e.* send the image from W_t to S_t). Once the actions have been made, the image sets U_t, W_t, and S_t are updated. We proceed iteratively until we either run out of images in $P_t = U_t \cup W_t$ or run out of budget d.

To this end, we study three active learning sampling functions (Random, Uncertainty, and Learning Active Learning [27]) within four action selection methods: Random Sampling (RS), Uncertainty Sampling (US), Optimization based on US and Optimization based on Learning Active Learning (LAL).

For RS, we randomly choose an *active batch* of images at each active step to include into W_t or S_t. For US, images are sorted in descending order of uncertainty (measured by entropy) to choose high uncertainty images to train on with full supervision[1], based on an uncertainty score denoted as s_k, and the top images are included into W_t or S_t. We adhere to the following annotation policy for RS and US when a budget constraint is enforced. For both RS and US, we prioritize weak labels first. In other words, so long as the budget constraint is not exceeded, (1) the image batch that is selected in each active step for these two methods will only contain images from U_t that will be weakly labeled or, (2) if U_t becomes empty, images from W_t will be selected and get a strongly label.

There are several ways to evaluate s_k for an image [55, 40]. We follow convention and model this score in Eq. 1:

$$s_k = \frac{1}{M} \sum_{i=1}^{M} \sum_{p \in \boldsymbol{p}_i} -p \log(p) \tag{1}$$

This is an entropy measure for each of the M bounding boxes in an image using the classification score predicted by the current detection model. Collecting M predictions from an image I_k, each prediction has a probability score vector $\boldsymbol{p}_i \in [0,1]^c$ for the c object categories.

Below, we explain the remaining active selection methods: Optimization based on US and LAL.

3.1.1 Optimization Based Active Selection using US

At the t-th active step, let $P_t = \{I_k\}_{k=1}^N$ be the N images that can be further annotated, and $\boldsymbol{s} \in \mathbb{R}^N$ be their corresponding uncertainty scores (as defined above). We have three possible action vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3 \in \{0,1\}^N$, as defined earlier. If the k-th element of $\boldsymbol{x}_i$ is 1, we annotate the image I_k using option i. Assume the next active batch of annotations has a linear impact on the model performance increment δ_t.

$$\delta_t = f_1(P_t)^\top \boldsymbol{x}_1 + f_2(P_t)^\top \boldsymbol{x}_2 + f_3(P_t)^\top \boldsymbol{x}_3 \tag{2}$$

To quantitatively show the contribution of new images, we also assume that the uncertainty score is a complete statistic of an unlabeled or weakly labeled set of images. Then, we can simply approximate (f_1, f_2, f_3) as linear functions. We observe that many active learning studies [30, 18] have experimentally shown that incorporating images with a higher uncertainty score in training may further improve the detector, but at the risk of having more difficulty in producing true positive predictions. Therefore, we model (f_1, f_2, f_3) as linear functions that tend to favor actions $\boldsymbol{x_1}, \boldsymbol{x_3}$ over action $\boldsymbol{x_2}$ for images with high enough uncertainty score (*i.e.* higher than the median). As such, and by combining the intuitions above, the expected increment in performance is modeled as:

$$\delta_t \approx \boldsymbol{s}^\top (\boldsymbol{x_1} + \boldsymbol{x_3}) + (\mu \boldsymbol{1} - \boldsymbol{s})^\top \boldsymbol{x_2} \tag{3}$$

For the active selection method based on optimization using US, we seek to maximize δ_t, while staying within the budget constraint. To model this latter constraint, we define a as the cost of strongly annotating an unlabeled image, b as the cost of weakly annotating an unlabeled image, and c as the cost of strongly annotating an already weakly labeled image. Therefore, this selection method seeks to solve the following binary optimization problem for $(\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3})$:

$$\max_{\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3} \in \{0,1\}^N} \boldsymbol{s}^\top (\boldsymbol{x_1} + \boldsymbol{x_3}) + (\mu \boldsymbol{1} - \boldsymbol{s})^\top \boldsymbol{x_2}$$
$$\text{s.t.} \begin{cases} \boldsymbol{x_3} \leq \boldsymbol{\psi} \\ \boldsymbol{x_1} + \boldsymbol{x_2} \leq \boldsymbol{1} - \boldsymbol{\psi} \\ \boldsymbol{1}^\top (a\boldsymbol{x_1} + b\boldsymbol{x_2} + c\boldsymbol{x_3}) \leq d \\ \boldsymbol{1}^\top (a\boldsymbol{x_1} + b\boldsymbol{x_2} + c\boldsymbol{x_3}) \geq d - a \end{cases} \tag{4}$$

Here, we define a vector ψ as the indicator vector for images that have already been weakly annotated, *i.e.* the k-th component of ψ is 1 if $I_k \in W_t$ and 0 otherwise. Using this indicator, the first two constraints in Eq (4) enforce that only one action is performed on each image, *i.e.* among all the k-th components of $(\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3})$, only one can be 1. The third and fourth constraints enforce that the budget be used as much as possible in each active step.

The linear binary problem in Eq (4) is NP-hard in general, so exact solvers (*e.g.* the Branch and Bound Algorithm) tend to have long run-times especially when N is large. As a tradeoff between optimization accuracy and per active step run-time, we employ a conventional linear relaxation of the original problem to form an approximate linear

[1] We also conducted an experiment in which images in US were sorted in ascending order and low uncertainty images were chosen in every step. We discuss this in greater detail in the supplementary material.

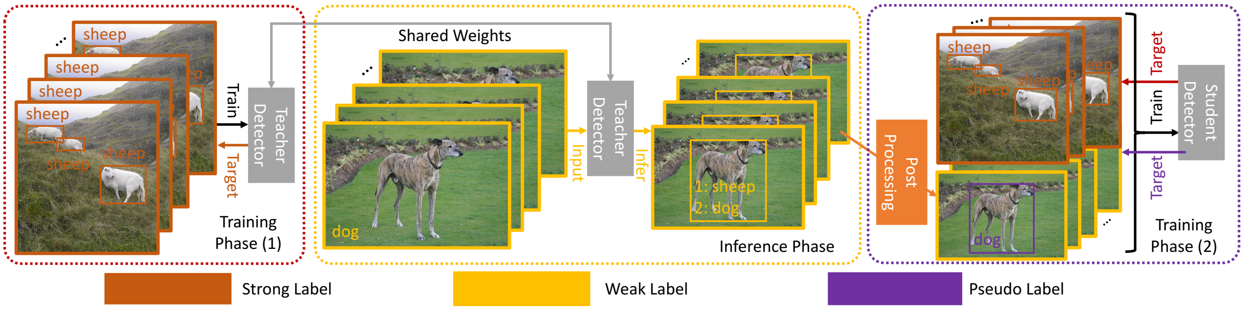

Figure 3: **Illustration of the Hybrid Learning framework.** Given an image collection with hybrid labels, we firstly use a warm-up detection model to generate pseudo instance labels (*e.g.* black solid rectangles in the inference phase.) After cleaning noise, the second detection model learns from both the ground truth and pseudo instance labels.

program (LP), which can be efficiently solved at large-scale with off-the-shelf LP solvers. More details about this problem are shown and proven in the supplementary material.

3.1.2 Optimization Based Active Selection using LAL

The uncertainty score $\boldsymbol{s}$ is not the only selection measure that has been studied in the active learning literature. In the Learning Active Learning (LAL) method proposed by Konyushkova *et al.* [27], the increment in performance function δ_t is learned to be a function of the current model state as well. More concretely, given images $I_k \in P_t$ and the current detection model, we build a feature vector $\boldsymbol{v} = [\boldsymbol{O}_t, \boldsymbol{s}]$ that concatenates both the current model state $\boldsymbol{O}_t$, represented as the average precision curves under five different Intersection over Union (IoU) thresholds, and the uncertainty scores $\boldsymbol{s}$. Following the LAL method in [27], we train a Support Vector Regression (SVR) model to regress the actual increment in mAP performance from $\boldsymbol{v}$ at each active step for both weak and strong annotation actions. Obviously, these SVRs are trained on a separate detection dataset than the one used to evaluate BAOD. At each active step and by denoting the output predictions of these SVRs as $\boldsymbol{h}_w$ and $\boldsymbol{h}_s$, we formulate the same constrained optimization problem in Eq (4) but with the *learned* objective:

$$\delta_t \approx \boldsymbol{h}_w^\top \boldsymbol{x_2} + \boldsymbol{h}_s^\top (\boldsymbol{x_1} + \boldsymbol{x_3}) \qquad (5)$$

3.2. Hybrid Supervision for Object Detection

At each active step, we train a hybrid supervised detector using both strong and weak supervision, *i.e.* using images from W_t and S_t (refer to Figure 3 for an illustration).

Teacher-Student Model. In this framework, we train an initial detector with the initial strongly annotated image set S_0. In each active step afterwards and to overcome undesirable local minima, we learn this detector using pretraining on Imagenet [51]. However, the model can also be fine-tuned from previous active steps to reduce computation time. Training on S_t at each active step can only learn a decent detector, but it can also transfer knowledge to weakly annotated images. This detector works as a teacher that predicts objects in every image in the weakly labeled image set W_t. If these predictions are post-processed properly, they can be viewed as pseudo labels, which we merge with the ground truth instance-level labels in S_t to train a fully supervised student detector. The student detector is also pretrained from Imagenet. Since it has both strong and weak annotations from training samples in $W_t \cup S_t$, we expect the student detector to perform better than its teacher.

Post-processing. The predictions from the teacher model at each step have many redundant or erroneous pseudo labels. We use minimal knowledge to post-process them so as to focus our study on the active selection methods and the benefit of hybrid training. More concretely, given an image I_k with a weak annotation $\boldsymbol{\omega} \in \{0,1\}^c$, where c is the number of categories, we assume that the teacher model gives M positive predictions with localization $\boldsymbol{P} \in \mathbb{R}^{M\times 4}$ (4 components defining a bounding box), classification $\boldsymbol{A} \in \{0,1\}^{M\times c}$, and the confidence score for each of the M positive predictions $\boldsymbol{q} \in [0,1]^M$. $\boldsymbol{P}$, $\boldsymbol{A}$ and $\boldsymbol{q}$ are obtained from the teacher model. Among these M predictions, we seek to mine the pseudo labels in the form of a sparse M-dimensional binary vector $\boldsymbol{y}$ that solves the following constrained optimization:

$$\max_{\boldsymbol{y}\in\{0,1\}^M} \boldsymbol{y}^\top \boldsymbol{q}$$
$$\text{s.t.} \begin{cases} \boldsymbol{y}^\top \boldsymbol{A}(\boldsymbol{1} - \boldsymbol{w}) = 0 \\ \text{IoU}(\boldsymbol{P}_i, \boldsymbol{P}_j) \leq \alpha \quad \forall\, y_i, y_j = 1 \\ 1 \leq \|\boldsymbol{y}\|_0 \leq \beta \end{cases} \qquad (6)$$

The intuition behind solving this problem is that we seek to maximize the confidence score across all pseudo labels indexed by $\boldsymbol{y}$. The first constraint enforces image-level consistency of the pseudo labels with the ground truth weak annotations. The second constraint removes predictions that are highly overlapping (similar to an Non-Maximal-Suppression post-processing step). The third constraint enforces a sparsity condition on $\boldsymbol{y}$, thus limiting the number of possible pseudo labeled bounding boxes (potential objects) to be between 1 and β. In our experiments, we take ($\alpha = 0.3, \beta = 3$). Details on this optimization are in the supplementary material.

4. Experiments and Analysis

4.1. Experimental Setup

Datasets. As in most WSOD methods, we use the PASCAL VOC 2007 (VOC07) or 2012 (VOC12) datasets [12] to perform most of the experiments. Given the active learning pipeline of our method, we emulate the active learning procedure using VOC07-12 annotations to selectively annotate 20 categories in 5011 images in VOC07 or 16551 images in the union of VOC07 and VOC12 (VOC0712). All detection models are evaluated on the VOC07 test dataset. For each annotation type (weak or strong), we assume that each image has a fixed annotating cost/time, which is not necessarily true in practice but it simplifies the analysis. In most of the experiments, we set ($a = 34.5, b = 1.6, c = a - b$), in unit seconds, according to the annotation procedure of [51]. In Section 4.4, we vary these cost value to ($a = 7, b = 1.6, c = a - b$) following the more efficient annotation procedure of [40][2]. The cost of fully annotating VOC07 trainval is denoted as 100% (or total) budget for every experiment.

Evaluation Metrics. To evaluate the active selection methods with the hybrid detector, we compute a budget-performance curve at various budget limits. The budget axis varies the percentage of the total budget, and the model detection performance is taken to be mAP. In doing so, we propose a new budget-aware metric denoted as **Budget-Average mAP**, measured as the normalized area under the budget-mAP curve for a certain budget range. We take three ranges $[10\%, 30\%]$, $[30\%, 50\%]$, and $[50\%, 100\%]$ to evaluate our experiment such that three to five data points are located in each range and the metric is less affected by noise. The first range starts from 10% since we need a small fully labeled warm up set to initialize the fully supervised detector, and start our pipeline. This warm up set is randomly selected and fixed in all experiments. The last budget range is wider because the performance curve saturates at high budgets, and we observe more subtle changes in performance. All budget-performance curves are shown in the supplementary material.

Implementation Details. In every active step, we choose Faster-RCNN as the object detection model (teacher and student), which is trained in the hybrid supervised way described in Section 3.2. Both teacher and student models use VGG16 as the backbone network pre-trained on ImageNet [51]. We follow the default setup in [58] to train Faster-RCNN. During training, the total number of epochs is set to 10, the learning rate is 0.01 for the first 8 epochs and 0.001 for the remaining. The batchsize is set to 16 on four-GPU cluster nodes equipped with Titan Xp. The student model is cloned from the ImageNet pretrained VGG16 in every active step and has the same training schedule as the teacher model[3]. For LAL experiment, we collect 10 categories from the MS-COCO dataset [34] to train the SVRs. Since the LAL method is dataset agnostic [27], we take the SVR training categories to be different from the 20 categories in PASCAL VOC.

4.2. Advantages of Uncertainty Sampling and Hybrid Training

Uncertainty Sampling. In order to explore the influence of the uncertainty score on model performance, we emulate the annotation process of the oracle using VOC07 and incorporate it into the active learning pipeline with *only* strong annotations and use a FSOD method to train our model. We use two selection methods (RS and US) to build the dataset for FSOD. From the first two columns in Table 1 that compare these two sampling methods, we see that collecting images with large uncertainty score is more effective and preferable than random sampling in the three budget ranges.

Hybrid Training. Table 1 also shows the influence of the active selection scheme on model performance but when the hybrid supervised learning process is used to combine weak and strong annotations at each active step. We again use RS and US to select the images in each step for hybrid training. It is clear that both hybrid methods outperform their corresponding FSOD counterpart at every budget range, especially when the budget range is low (under 50%).

Table 1: **Budget-Average mAP using fully and hybrid training pipelines with random and uncertainty selection.** Uncertainty sampling is always better than random sampling selection, and hybrid training is always better than FSOD.

Selection Method	**FSOD**		**Hybrid**	
Sampling Function	**RS**	**US**	**RS**	**US**
Low Budget Range	52.5	53.1	**56.4**	55.1
Mid Budget Range	62.5	63.6	64.5	**65.8**
High Budget Range	67.9	68.7	68.7	**69.3**

4.3. Optimization-Based Active Selection

Here, we evaluate the optimization-based selection methods (using US and LAL) described in Section 3.1, when they are combined with hybrid supervised training. Table 2 compares these methods by measuring their average mAP at each budget range and compares them to the RS and US selection methods for references. These results indicate that the optimization methods are the best ways to combine and use the hybrid annotations at every budget, where the optimization-based US method is slightly better

[2] We left more discussion on the annotating time assumption in the supplementary material.

[3] The source code of the framework will be made publicly available.

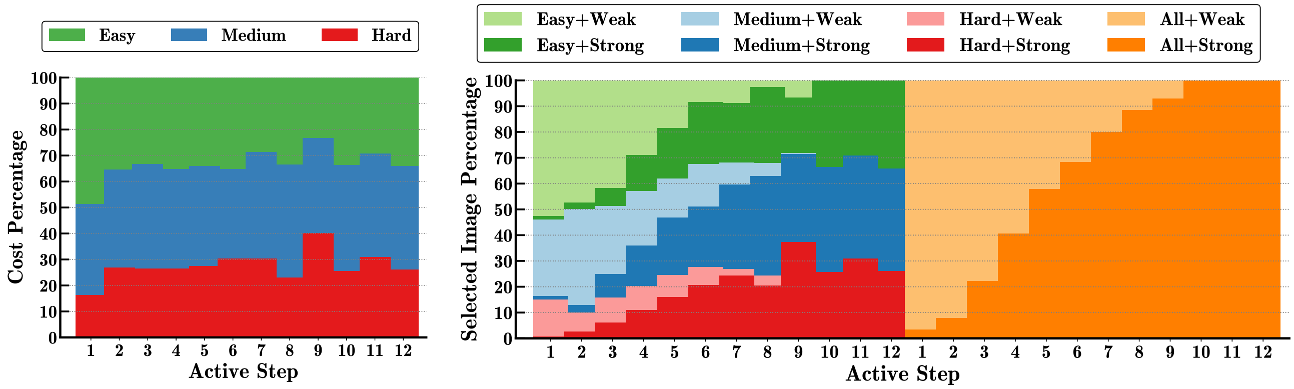

Figure 4: **Comparison of the cost and image number on every active selection step.** *Left*: Budget usage distribution to learn different difficulty categories. More budget is used to annotate Easy images (green area) at beginning. The cost spent on Hard images (red area) grows up when the active model is mature. *Right*: Selected images distribution in different difficulty categories and different annotation type. The selection agent gives more weak annotations (light color) at the first steps. Given more budget, the proportion of strong annotations (dark color) increase. We run out of unlabeled images after 9-th step. The mapping is motivated and shown in the supplementary material.

than its LAL counterpart. As such, we denote the former method as the BAOD approach, which was mentioned and highlighted in Figure 1. Interestingly, we observe that the RS method requires 62% of the total VOC07 budget (all images are strongly annotated) to achieve 95% of the detection performance at that budget (*i.e.* 67.4% mAP). In comparison, the BAOD method requires only 48.5% of the total budget to reach the same performance. This performance gap attests to the effectiveness of this method. We include more results in the supplementary material.

Table 2: **Budget-Average mAP using simple hybrid training and optimization methods.** US based optimization is slightly better than LAL one. The optimization methods perform better than the simple hybrid random selection and uncertainty selection methods in the three budget ranges.

Selection Method	**Hybrid**		**Optimization**	
Sampling Function	**RS**	**US**	**LAL**	**US** (BAOD)
Low Budget Range	56.4	55.1	56.3	**57.1**
Mid Budget Range	64.5	65.8	65.9	**66.0**
High Budget Range	68.7	69.3	69.3	**69.5**

4.4. Effect of Per-Image Annotation Cost

The cost of strong annotations can vary due the annotation strategy that is used. For instance, Papadopoulos *et al.* [40] created a method that reduces the time needed to draw bounding boxes to 7 seconds per image on average. Here, we use this cost ($a = 7$) to study the behavior of our BAOD method given a smaller gap between strong and weak annotation. We report these results in Table 3, which depicts the best performing FSOD method and the RS hybrid baseline for reference. We observe the same relative behavior between the methods as in the case when a was several times higher. However, the performance gap between these methods decreases in this case because the number of images that can be weakly annotated for the cost of one strong annotation is much smaller.

Table 3: **Budget-Average mAP using a lower cost for strong annotations.** If we assume the weak and strong annotation costs are more close (7 seconds and 1.5 seconds), US based optimization (BAOD) still performs better than the simple hybrid random selection and uncertainty selection methods in the three budget ranges.

Selection Method	**FSOD**	**Hybrid**	**Optimization**
Sampling Function	**RS**	**US**	**US** (BAOD)
Low Budget Range	53.1	52.3	**54.2**
Mid Budget Range	62.8	63.1	**63.6**
High Budget Range	68.3	68.6	**69.3**

4.5. Improving Detection using Fixed Budget

This experiment simulates a real-world application of our method. We combine VOC07 and VOC12 data to simulate a larger pool of unlabeled images (called VOC0712) to choose from. As reported in Table 4, when the budget is set at 87.2% of the total budget (annotation budget for VOC07), we learn an object detector whose performance is the same as an FSOD detector trained on the entire strongly labeled VOC07 training set. This is a budget saving of 12.8%. Now, if we choose to use this total budget on VOC0712, our method achieves a 73.0% mAP, which is 2% improvement over the aforementioned FSOD detector on the same VOC07 test set. If all the 16551 images in the union of VOC07 and VOC12 set are strongly labeled, we use an extra 230% of budget and only improve 3.4% mAP.

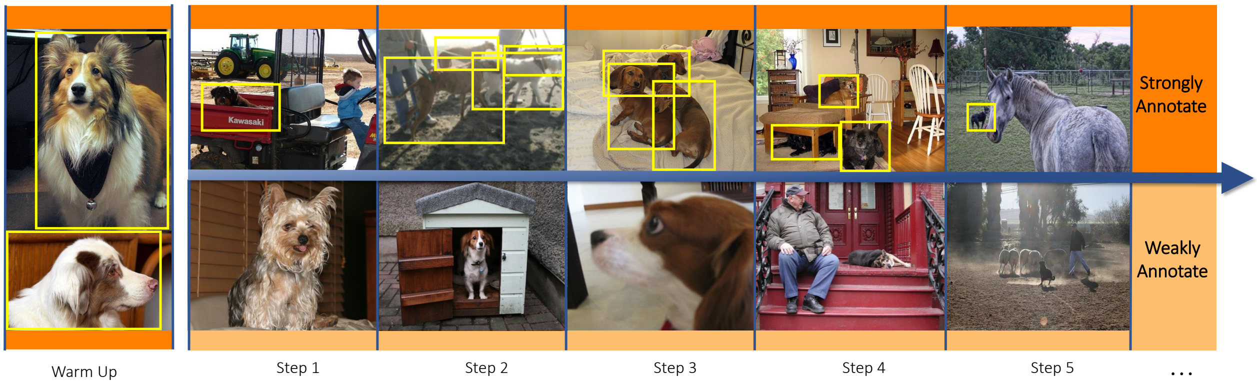

Figure 5: **Visualization of the selected images in each step.** *Left*: Two examples in the warm-up set which is fully annotated by 10% budget. *Up-Right*: Strongly annotated images per step. They are hard examples including occlusion, multiple instance or tiny scale. *Bottom-Right*: Weakly annotated images per step. They are simple in the beginning but the difficulty increases when the detector is mature.

Table 4: **Simulation of a larger unlabeled image pool.** With 87.2% budget, BAOD achieves the same performance as a detector trained on fully annotated VOC07. If the budget equals to the total budget of VOC07, BAOD achieves 2% mAP improvement over FSOD with the same budget. Further annotating all the images can only improve 3.4% mAP.

Train Set Pool	VOC07	VOC0712		
Max Budget*	100%	**87.2%**	100%	330%
Final mAP	0.710	0.710	0.730	**0.764**

* The values are normalized by VOC07 trainval set cost.

4.6. Easy Images and Weak Annotation First

We analyze the cost and number of images selected at every active selection step to investigate which type of training examples are more helpful in the sequential training process. Based on the final mAP of each category in VOC07 [13], we divide the twenty categories into three groups: Easy, Medium, and Hard.The left plot of Figure 4 illustrates that BAOD spends more budget to annotate **Easy** categories (shown in the green area) in the first several steps, while the cost of **Hard** categories (red area) increases when the detector becomes more accurate. These results align with concepts from curriculum learning, in which a larger number of *easy* samples can be trained on first to bootstrap the model and then *hard* samples are introduced progressively.

The right plot of Figure 4 measures the percentage of selected images per active step that belong to various subsets of the data (combinations of annotation type chosen and difficulty). Interestingly, the BAOD model tends to select more weak annotations at the beginning from all three groups of objects, since this kind of label is cheap and informative. Given a larger budget, the agent increases the proportion of strong annotations to further improve model performance. Note that there are no new weakly labeled images after active step 9 because all the images are already annotated.

4.7. Qualitative Results of the Active Selection

Figure 5 shows some selected strongly and weakly labeled images based on the BAOD experiment in section 4.5. Each row of the images is from the *dog* category in the VOC07 or VOC12 trainval set, and each column of images is selected at the same active step.

We show that in the first five active steps, strongly annotated images contain dog instances that are difficult to detect due to occlusion, multiple close instances or small-scale. In contrast, the selected weakly annotated images during the same steps are relatively easier to locate. As the model gets mature the difficulty for both levels of annotations increases. For example, the image chosen in step 4 contains three small dogs, and the dogs appeared in step 5's images are small and black which makes them barely visible.

5. Conclusion

In summary, we introduce a novel budget-aware perspective to study the unexplored dimensions of the object detection problem. With a fixed budget, we compare both optimization and learn based sample methods to build diverse hybrid supervised object detection datasets which consist of both image level supervision and instance level supervision. The evaluation of detectors that learned from these budget fixed datasets shows that the handcrafted optimization method on uncertainty score outperforms other general active learning methods including random sampling, active learning, and reinforcement learning (shown in supplementary material). With the optimal set-up, our proposed budget-aware approach can achieve the performance of a strongly supervised detector on PASCAL-VOC 2007 while saving 12.8% of its original annotation budget. Furthermore, when 100% of the budget is used, our approach surpasses this performance by 2 percentage points of mAP.

References

[1] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem. Finding tiny faces in the wild with generative adversarial network. *CVPR*, 2018. 1

[2] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9368–9377, 2018. 3

[3] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with convex clustering. In *CVPR*, pages 1081–1089, 2015. 2

[4] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, pages 2846–2854, 2016. 2

[5] P. Chattopadhyay, R. Vedantam, R. R. Selvaraju, D. Batra, and D. Parikh. Counting everyday objects in everyday scenes. In *CVPR*, pages 4428–4437, 2017. 1

[6] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1409–1416, 2013. 3

[7] G. Chéron, J.-B. Alayrac, I. Laptev, and C. Schmid. A flexible model for training action localization with varying levels of supervision. *arXiv preprint arXiv:1806.11328*, 2018. 3

[8] R. G. Cinbis, J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *TPAMI*, 39(1):189–203, 2017. 2

[9] J. Collomosse, T. Bui, M. J. Wilber, C. Fang, and H. Jin. Sketching with style: Visual search with sketches and aesthetic context. In *ICCV*, pages 2679–2687, 2017. 1

[10] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: object detection via region-based fully convolutional networks. *CoRR*, abs/1605.06409, 2016. 2

[11] X. Du, M. El-Khamy, J. Lee, and L. Davis. Fused dnn: A deep neural network fusion approach to fast and robust pedestrian detection. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 953–961. IEEE, 2017. 1

[12] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 2, 6

[13] R. Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 2, 8

[14] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013. 2

[15] R. Gokberk Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *CVPR*, pages 2409–2416, 2014. 2

[16] M. Hasan and A. K. Roy-Chowdhury. Context aware active learning of activity recognition models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4543–4551, 2015. 3

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. *arXiv preprint arXiv:1703.06870*, 2017. 2

[18] F. C. Heilbron, J.-Y. Lee, H. Jin, and B. Ghanem. What do i annotate next? an empirical study of active learning for action localization. In *ECCV*, 2018. 3, 4

[19] J. Hoffman, D. Pathak, T. Darrell, and K. Saenko. Detector discovery in the wild: Joint multiple instance and representation learning. In *CVPR*, pages 2883–2891, 2015. 2

[20] P. Hu and D. Ramanan. Finding tiny faces. In *CVPR*, pages 1522–1530. IEEE, 2017. 1

[21] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu. Deep self-taught learning for weakly supervised object localization. *arXiv preprint arXiv:1704.05188*, 2017. 2

[22] D. Kang and A. Chan. Crowd counting by adaptively fusing predictions from an image pyramid. *arXiv preprint arXiv:1805.06115*, 2018. 1

[23] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *ECCV*, pages 350–365. Springer, 2016. 2

[24] C.-C. Kao, T.-Y. Lee, P. Sen, and M.-Y. Liu. Localization-aware active learning for object detection. *arXiv preprint arXiv:1801.05124*, 2018. 3

[25] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with gaussian processes for object categorization. In *ICCV*, pages 1–8. IEEE, 2007. 3

[26] K. Konyushkova, R. Sznitman, and P. Fua. Introducing geometry in active learning for image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2974–2982, 2015. 3

[27] K. Konyushkova, R. Sznitman, and P. Fua. Learning active learning from data. In *Advances in Neural Information Processing Systems*, pages 4225–4235, 2017. 3, 4, 5, 6

[28] K. Konyushkova, J. Uijlings, C. H. Lampert, and V. Ferrari. Learning intelligent dialogs for bounding box annotation. In *CVPR*, 2018. 1, 2, 3

[29] S. Laine and T. Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 3

[30] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine Learning Proceedings 1994*, pages 148–156. Elsevier, 1994. 4

[31] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang. Weakly supervised object localization with progressive domain adaptation. In *CVPR*, pages 3512–3520, 2016. 2

[32] L.-J. Li and L. Fei-Fei. Optimol: automatic online picture collection via incremental model learning. *International journal of computer vision*, 88(2):147–168, 2010. 3

[33] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Featuref pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016. 2

[34] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 1, 2, 6

[35] B. Liu and V. Ferrari. Active learning for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4363–4372, 2017. 3

[36] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016. 2

[37] I. Misra, R. Girshick, R. Fergus, M. Hebert, A. Gupta, and L. van der Maaten. Learning by asking questions. *arXiv preprint arXiv:1712.01238*, 2017. 3

[38] C. Mu, J. Zhao, G. Yang, J. Zhang, and Z. Yan. Towards practical visual search engine within elasticsearch. *arXiv preprint arXiv:1806.08896*, 2018. 1

[39] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari. We don't need no bounding-boxes: Training object class detectors using only human verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 854–863, 2016. 1, 2, 3

[40] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari. Extreme clicking for efficient object annotation. In *ICCV*, pages 4940–4949. IEEE, 2017. 1, 2, 4, 6, 7

[41] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang. Two-dimensional active learning for image classification. In *CVPR*, pages 1–8. IEEE, 2008. 3

[42] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He. Data distillation: Towards omni-supervised learning. *arXiv preprint arXiv:1712.04440*, 2017. 3

[43] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018. 1

[44] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 2

[45] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016. 2

[46] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 2

[47] P. K. Rhee, E. Erdenee, S. D. Kyun, M. U. Ahmed, and S. Jin. Active and semi-supervised learning for object detection with imperfect data. *Cognitive Systems Research*, 45:109–123, 2017. 3

[48] M. Rochan and Y. Wang. Weakly supervised localization of novel objects using appearance transfer. In *CVPR*, pages 4315–4324, 2015. 2

[49] C. Rosenberg, M. Hebert, and H. Schneiderman. Semi-supervised self-training of object detection models. In *WACV/MOTION*, pages 29–36, 2005. 3

[50] S. Roy, A. Unmesh, and V. P. Namboodiri. Deep active learning for object detection. In *BMVC*, page 91, 2018. 3

[51] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 2, 5, 6

[52] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell. Weakly-supervised discovery of visual pattern configurations. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1637–1645, 2014. 2

[53] H. Su, J. Deng, and L. Fei-Fei. Crowdsourcing annotations for visual object detection. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, volume 1, 2012. 2

[54] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *ECCV*, pages 431–445. Springer, 2014. 2

[55] D. Wang and Y. Shang. A new active labeling method for deep learning. *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 112–119, 2014. 3, 4

[56] S. Wang, J. Cheng, H. Liu, and M. Tang. Pcn: Part and context information for pedestrian detection with cnns. *arXiv preprint arXiv:1804.04483*, 2018. 1

[57] M. Xu, Y. Bai, and B. Ghanem. Missing labels in object detection. In *CVPR Workshops*, 2019. 3

[58] J. Yang, J. Lu, D. Batra, and D. Parikh. A faster pytorch implementation of faster r-cnn. *https://github.com/jwyang/faster-rcnn.pytorch*, 2017. 6

[59] Y. Zhang, Y. Bai, M. Ding, Y. Li, and B. Ghanem. W2f: A weakly-supervised to fully-supervised framework for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 928–936, 2018. 3

[60] Y. Zhang, M. Ding, Y. Bai, M. Xu, and B. Ghanem. Beyond weakly supervised: Pseudo ground truths mining for missing bounding-boxes object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):983–997, 2019. 3

[61] J.-J. Zhu and J. Bento. Generative adversarial active learning. *arXiv preprint arXiv:1702.07956*, 2017. 3

[62] B. Zhuang, L. Liu, Y. Li, C. Shen, and I. Reid. Attend in groups: a weakly-supervised deep learning framework for learning from web data. *arXiv preprint arXiv:1611.09960*, 2016. 2