Active Learning Strategies for Weakly-Supervised Object Detection

Huy V. Vo[1,2] Oriane Siméoni[2] Spyros Gidaris[2] Andrei Bursuc[2]
Patrick Pérez[2] Jean Ponce[1,3]

[1]Inria and DI/ENS (ENS-PSL, CNRS, Inria) [2]Valeo.ai
[3]Center for Data Science, New York University

{van-huy.vo, jean.ponce}@inria.fr
{oriane.simeoni,spyros.gidaris,andrei.bursuc,patrick.perez}@valeo.com

Abstract. Object detectors trained with weak annotations are affordable alternatives to fully-supervised counterparts. However, there is still a significant performance gap between them. We propose to narrow this gap by fine-tuning a base pre-trained weakly-supervised detector with a few fully-annotated samples automatically selected from the training set using "box-in-box" (BiB), a novel active learning strategy designed specifically to address the well-documented failure modes of weakly-supervised detectors. Experiments on the VOC07 and COCO benchmarks show that BiB outperforms other active learning techniques and significantly improves the base weakly-supervised detector's performance with only a few fully-annotated images per class. BiB reaches 97% of the performance of fully-supervised Fast RCNN with only 10% of fully-annotated images on VOC07. On COCO, using on average 10 fully-annotated images per class, or equivalently 1% of the training set, BiB also reduces the performance gap (in AP) between the weakly-supervised detector and the fully-supervised Fast RCNN by over 70%, showing a good trade-off between performance and data efficiency. Our code is publicly available at https://github.com/huyvvo/BiB.

Keywords: object detection, weakly-supervised, active learning

1 Introduction

Object detectors are critical components of visual perception systems deployed in real-world settings such as robotics or surveillance. Many methods have been developed to build object detectors with high predictive performance [31,32,33,36,54] and fast inference [52,53]. They typically train a neural network in a fully-supervised manner on large datasets annotated manually with bounding boxes [23,24,47]. In practice, the construction of these datasets is a major bottleneck since it involves large, expensive and time-consuming data acquisition, selection and annotation campaigns. To address this challenge, much effort has been put in devising object detection approaches trained with less (or even no) human annotation. This includes semi-supervised [39,51,63,76], weakly-supervised [7,15,29,38,55,69,80], few-shot [25,41,43,66], active [1,8,14,35,58,59] and unsupervised [13,60,62,67,72,74] learning frameworks for object detection.

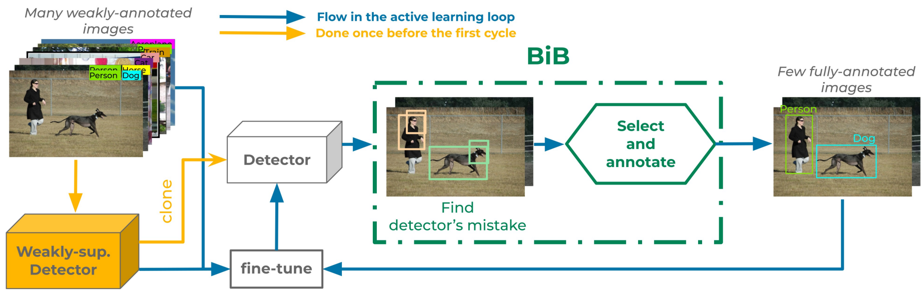


Fig. 1: Overview of our approach. A base object detector is first trained only with image-level tags, then fine-tuned in successive stages using few *well-selected* images that are fully annotated. For their selection, we propose *"box-in-box"* (BiB), an acquisition function designed to discover recurring failure cases of the weakly-supervised detector, e.g., failure to localize whole objects or to separate distinct instances of the same class.

Weakly-supervised object detection (WSOD) typically only uses image-level category labels during training [7,55,69]. This type of annotation is much cheaper than bounding boxes and, in some cases, it can be even obtained automatically, e.g., leveraging tags on online photos, photo captions in media or time-stamped movie scripts. WSOD is thus an affordable alternative to fully-supervised object detection in terms of annotation cost. However, weakly-supervised detectors often struggle to correctly localize the full extent of objects [55,69]. Several recent works [6,49] show that a good trade-off between performance and annotation cost can be achieved by annotating bounding boxes in a small set of randomly selected training images and by training the detector with a mix of weak and full supervision. However, there are better alternatives to random selection. Active learning (AL) methods [14,79] offer means to *select* images that should be the most helpful for the training of an object detector model, given some criterion.

In this work, we propose to combine both worlds, by augmenting the weakly-supervised framework with an active learning scheme. Our active learning strategy specifically targets the known failure modes of weakly-supervised detectors. We show that it can be used to significantly narrow the gap between weakly-supervised detectors and expensive fully-supervised ones with a few fully-annotated images per class. We start with a weakly-annotated dataset, e.g., a set of images and their class labels, with which we train a weakly-supervised detector. We apply our new active learning strategy that we call *box-in-box* (BiB) to iteratively select from the dataset a few images to be fully annotated. New full annotations are added to the training set and used to fine-tune the detector. Given the fine-tuned detector, we select another batch of images to be fully annotated. This process is repeated several times to improve the detector (Figure 1). Previous works have attempted to combine weak supervision with active learning, but they all start with an initial set of hundreds to thousands of fully-annotated images. As shown in Section 4, our approach only requires a small number of fully-annotated images ($50-250$ on VOC07 [24] and $160-800$ on COCO [47]) to significantly improve the performance of weakly-supervised detectors. Our main contributions are: (1) We propose a new approach to improve weakly-supervised object detectors, by using a few fully-annotated images, carefully selected with the help of active learning. Contrary to typical active learning approaches, we initiate the learning process without any fully-annotated data; (2) We introduce BiB, an active selection strategy that is tailored to address the limitations of weakly-supervised detectors; (3) We validate our proposed approach with extensive experiments on VOC07 and COCO datasets. We show that BiB outperforms other active strategies on both datasets, and reduce significantly the performance gap between weakly- and fully-supervised object detectors.

2 Related Work

Weakly-supervised object detection is a data-efficient alternative to fully-supervised object detection which only requires image-level labels (object categories) for training a detector. It is typically formulated as a multiple instance learning problem [19] where images are bags and region proposals [71,83] are instances. The model is trained to classify images using scores aggregated from their regions, through which it also learns to distinguish *object* from *non-object* regions. Since training involves solving a non-convex optimization problem, adapted initialization and regularization techniques [15,17,44,64,65] are necessary for good performance. Bilen *et al.* [7] proposes WSDDN, a CNN-based model for WSOD which is improved in subsequent works [18,40,55,68,69]. Tang *et al.* [69] proposes OICR which refines WSDDN's output with parallel detector heads in a self-training fashion. Trained with only image-level labels, weakly-supervised object detectors are often confused between object parts and objects, or between objects and groups of objects [55]. Although mitigating efforts with better pseudo labels [55,68], better representation [38,55] or better optimization [3,75] achieve certain successes, such confusion issues of weakly-supervised detectors remain due to the lack of a formal definition of objects and their performance is still far behind that of fully-supervised counterparts. In this work, we show that fine-tuning weakly-supervised detectors with strong annotation on *a few carefully selected* images can alleviate these limitations and significantly narrow the gap between weakly- and fully-supervised object detectors.

Semi-supervised object detection methods exploits a mix of some fully-annotated and many unlabelled-data. Two dominant strategies arise among these methods: consistency-based [39,70] and pseudo-labeling [45,51,63,76,77,84]. The latter can be further extended with strategies inspired by active learning [45,76] for selecting boxes to be annotated by humans.

Combining weakly- and semi-supervised object detection. These approaches seek a better trade-off between performance and annotation cost than individual strategies. All images from the training set have weak labels and a subset is also annotated with bounding boxes. This setup enables the exploration of the utility of additional types of weak labels, e.g., points [10,56] or scribbles [56]. Others leverage fully-annotated images to train detectors that can correct wrong predictions of weakly-supervised detectors [49] or compute more reliable pseudo-boxes [6]. Similarly to [6,49], we train a detector with only a few annotated images, but different from them, we focus on how to best select the images to annotate towards maximizing the performance of the detector.

Active learning for object detection aims at carefully *selecting* images to be fully annotated by humans, in order to minimize human annotation efforts. Most methods exploit *data diversity* [30,58] or *model uncertainty* [8,14] to identify such images. These strategies, originally designed for generic classification tasks [59], have been recently derived and adapted to object detection [14,79], a complex task involving both classification (object class) and regression (bounding box location). Data diversity can be ensured by selecting data samples using image features and applying k-means [82], k-means++ initialization [35] or identifying a core-set – a *representative* subset of a dataset [1,30,58]. Model uncertainty for AL can be computed from image-level scores aggregated from class predictions over boxes [8,35,50], comparing predictions of the same image from different corrupted versions of it [22,42,28] or from different steps of model training [37,57], voting over predictions from an ensemble of networks [5,12,35], Bayesian Neural Networks [27,35] or single forward networks mimicking an ensemble [14,79]. Multiple other strategies have been proposed for selecting informative, difficult or confusing samples to annotate by: learning to discriminate between labeled and unlabeled data [20,21,34,81], learning to predict the detection loss [78], the gradients [4] or the influence of data on gradient [48]. In contrast to classical active learning methods in which the initial model is trained in a fully-supervised fashion using a randomly sampled initial set of images, our initial model is only trained with weakly-annotated data. This is a challenging problem, but often encountered in practice when new collections of data arrive only with weak annotations and significant effort is required to select which images to annotate manually prior to active learning.

Combining weak supervision and active learning. Closer to us, [16,26,50] investigate how weakly-supervised learning and active learning can be conducted together in the context of object detection. Desai et al. [16] propose to use clicks in the center of the object as weak labels which include localization information and are stronger than image-level tags. Pardo et al. [50] also mix strong supervision, tags and pseudo-labels in an active learning scenario. Both [16,50] rely on Faster R-CNN [54] and [26] on a FPN [46] – detectors that are hard to train only with weak labels. All start with 10% of the dataset fully labeled, which is more than the total amount of fully annotated data we consider in this work.

3 Proposed Approach

3.1 Problem Statement

We assume that we are given n images $\mathcal{I} = \{\mathbf{I}_i\}_{i\in\{1\ldots n\}}$ annotated with labels $\mathcal{Q} = \{\mathbf{q}_i\}_{i\in\{1\ldots n\}}$. Here $\mathbf{q}_i \in \{0,1\}^C$ is the class label of image $\mathbf{I}_i$, with C being

Algorithm 1: WSOD with Active Learning.

Input: Set $\mathcal{I}$ of weakly-labelled images, set $\mathcal{Q}$ of weak annotations, number of cycles T, budget per cycle B.

Result: Detector M^T, bounding box annotations $\mathcal{G}^T$.

$M^0 \leftarrow \texttt{train}(\mathcal{I}, \mathcal{Q})$ ▷ weakly-supervised pre-training

for $t = 1$ *to* T **do**

  $A^t \leftarrow \texttt{select}(W^{t-1}, M^{t-1}, \mathcal{I}, \mathcal{Q}, B)$ ▷ select a batch A^t of B images

  $\mathcal{G}^t \leftarrow \mathcal{G}^{t-1} \cup \texttt{label}(\mathcal{I}, A^t)$ ▷ annotate new selection

  $S^t \leftarrow S^{t-1} \cup A^t,\ \ W^t \leftarrow W^{t-1} \setminus A^t$ ▷ update the sets

  $M^t \leftarrow \texttt{fine-tune}(\mathcal{I}, \mathcal{Q}, \mathcal{G}^t, M^0)$ ▷ fine-tune the model

end

the number of classes in the dataset. Let M^0 be a weakly-supervised object detector trained using only $\mathcal{Q}$. The goal of our work is to iteratively select a *very small set of images* to fully annotate with bounding boxes and fine-tune M^0 on the same images with both weak and full annotation so as to maximize its performance. To that end, we propose a novel *active learning* method properly adapted to the aforementioned problem setting.

3.2 Active Learning for Weakly-Supervised Object Detection

As typical in active learning, our approach consists of several cycles in which an acquisition function first uses the available detector to select images that are subsequently annotated by a human with bounding boxes. The model is then updated with this additional data. With the updated detector, a new cycle of acquisition is performed (see Algorithm 1).

Let $W^t \subset \{1, \ldots, n\}$ be the set of indices of images with class labels only, and $S^t \subset \{1, \ldots, n\}$ the set with bounding-box annotations at the t-th active learning cycle. The active learning process starts with $W^0 = \{1, \ldots, n\}$ and $S^0 = \varnothing$. Then, at each cycle $t > 0$, the acquisition function selects from W^{t-1} a set A^t of B images to be annotated with bounding boxes, with B the fixed annotation budget per cycle. By definition, we have that $A^t \subset W^{t-1}$ and $|A^t| = B$. For the selection, the acquisition function exploits the detector M^{t-1} obtained at the end of the previous cycle. After selecting A^t, the sets of fully and weakly-annotated images are updated with $S^t = S^{t-1} \cup A^t$ and $W^t = W^{t-1} \setminus A^t$ respectively. We define as $\mathcal{G}^t = \{\mathbf{G}_i\}_{i \in S^t}$ the bounding-box annotations for images with indices in S^t. Finally, at the end of cycle t, we fine-tune M^0 on the entire dataset, using the bounding box annotations for images with indices in S^t and the original image-level annotations for others.

3.3 BiB: An Active Learning Strategy

With a very small annotation budget, we aim at selecting the "best" training examples to "fix" the mistakes of the base weakly-supervised detector. We propose

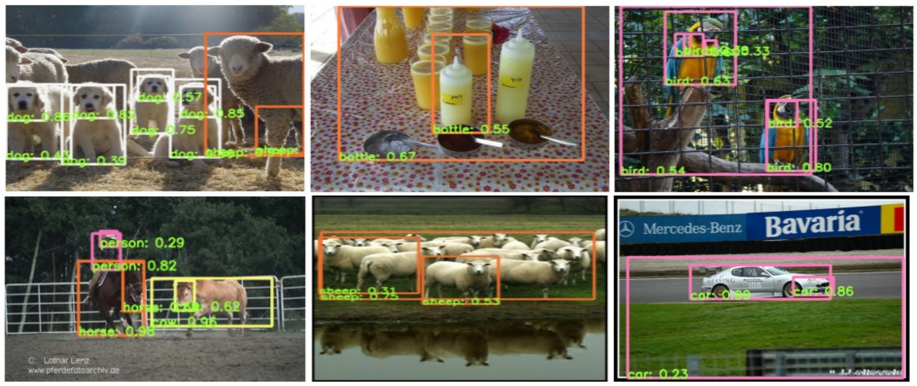


Fig. 2: Example of box-in-box (BiB) pairs among the predictions of the weakly-supervised object detector. The existence of such pairs is an indicator of the detector's failure on those images. Best viewed in color.

BiB, an acquisition strategy tailored for this purpose. It first *discovers (likely) detection mistakes* of the weakly-supervised detector, and then selects *diverse detection mistakes*. Our selection strategy is summarized in Algorithm 2.

Discovering BiB patterns. Weakly-supervised detectors often fail to accurately localize the full extent of the objects in an image, and tend to focus instead on the most *discriminative parts* of an object or to group together multiple object instances [55]. Several examples of these errors are shown in Figure 2. In the first column, a predicted box focuses on the most discriminative part of an object while a bigger one encompasses a much larger portion of the same object. Another recurring mistake is when two or more distinct objects are grouped together in a box, but some correct individual predictions are also provided for the same class (second column). The two kinds of mistakes can also be found in the same image (third column). We name "box-in-box" (BiB) such detection patterns where two boxes are predicted for a same object class, a small one being "contained" (within some tolerance, see below) in a larger one. We take BiB pairs as an indicator of model's confusion on images.

More formally, let $\mathbf{D}_i$ be the set of boxes detected in image $\mathbf{I}_i$ and let d_A and d_B be two of them. We consider that (d_A, d_B) is a BiB pair, which we denote with is-bib$(d_A, d_B) =$ `True`, when: (i) d_A and d_B are predicted for the same class, (ii) d_B is at least μ times larger than d_A (i.e., $\frac{\text{Area}(d_B)}{\text{Area}(d_A)} \geq \mu$), and (iii) the intersection of d_B and d_A over the area of d_A is at least δ (i.e., $\frac{\text{Intersection}(d_A,d_B)}{\text{Area}(d_A)} \geq \delta$). Hence, the set $P_i = \{p_{i,j}\}_{j=1}^{|P_i|}$ of BiB pairs is found in image $\mathbf{I}_i$ by the procedure

$$\text{find-bib}(\mathbf{D}_i) = \{(d_A, d_B) \in \mathbf{D}_i \times \mathbf{D}_i |\, \text{is-bib}(d_A, d_B)\}. \tag{1}$$

We observe that in such a BiB pair, it is likely that at least one of the boxes is a detection mistake. We thus propose to select images to be fully annotated among those containing BiB pairs.

Algorithm 2: BiB acquisition strategy.

Result: Set A^t of selected images.
Input: Budget B, model M^{t-1}, image set $\mathcal{I}$, index set W^{t-1} of weakly-annotated images, set $\hat{\mathcal{P}}$ of already selected BiB pairs (if empty, see text for initialization)

for $i \in W^{t-1}$ **do**
  $\mathbf{D}_i \leftarrow \texttt{Detect}(\mathbf{I}_i|M^{t-1})$ ▷ Predict boxes
  $P_i \leftarrow \{p_{i,j}\}_{j=1}^{|P_i|} = \text{find-bib}(\mathbf{D}_i)$ ▷ Discover BiB patterns
end
Select diverse detection mistakes
$A^t \leftarrow \varnothing$
while $|A^t| < B$ **do**
  for $p \in \cup_{i \in W^{t-1} \setminus A^t} P_i$ **do**
    $w_p \leftarrow \min_{\tilde{p} \in \hat{\mathcal{P}}} \|F(p) - F(\tilde{p})\|$ ▷ Comp. dist. to selected pairs
  end
  $p^* \sim \text{Prob}(\{w_p\}_p)$ ▷ Randomly select a pair
  $i^* \leftarrow \texttt{get-imid}(p)$ ▷ Get index of the image containing p
  $\hat{\mathcal{P}} \leftarrow \hat{\mathcal{P}} \cup P_{i^*}$, $A^t \leftarrow A^t \cup \{i^*\}$ ▷ Updates
end

Selecting diverse detection mistakes. Given the set of all BiB pairs over $\mathcal{I}$, the acquisition function considers the *diversity* of the pairs in order to select images. In particular, we follow *k-means++ initialization* [2] – initially developed to provide a good initialization to k-means clustering by iteratively selecting as centroids data points that lie further away from the current set of selected ones. This initialization has previously been applied onto image features in the context of active learning for object detection [35] or on model's gradients for active learning applied to image classification [4]. Here we focus and apply the algorithm to pairs of detected boxes.

We denote with $\hat{\mathcal{P}}$ the set of BiB pairs from the already selected images. For each pair p not in $\hat{\mathcal{P}}$, we compute the minimum distance w_p to the pairs in $\hat{\mathcal{P}}$: $w_p \leftarrow \min_{\tilde{p} \in \hat{\mathcal{P}}} \|F(p) - F(\tilde{p})\|$, where $F(p)$ is the feature vector of p, i.e., the concatenation of the region features corresponding to the two boxes of p each extracted using the model M^{t-1}. We then randomly pick a new pair p^*, using a weighted probability distribution where a pair p is chosen with probability proportional to w_p. We finally select the image $\mathbf{I}_{i^*}$ that contains p^*, add its index i^* to A^t and its BiB pairs to $\hat{\mathcal{P}}$. Note that at the beginning of the selection process in each cycle, $\hat{\mathcal{P}}$ contains the pairs of images selected in the previous cycles and is empty when the first cycle begins. In the latter case, we start by selecting the image $\mathbf{I}_{i^*}$ that has the greatest number of pairs $|P_{i^*}|$[1] and add the pairs in P_{i^*} to $\hat{\mathcal{P}}$ before starting the selection process above.

With this design, BiB selects a diverse set of images that are representative of the dataset while addressing the known mistakes of the weakly-supervised detec-

[1] In case of a draw, an image is randomly selected.

tor. We show some examples selected by BiB and demonstrate its effectiveness in boosting the performance of the weakly-supervised detector in Section 4.

3.4 Training Detectors with both Weak and Strong Supervision

We provide below details about the step of model fine-tuning performed at each cycle. For clarity, we drop the image index i and the cycle index t in this section.

Training with weak annotations. We adopt the state-of-the-art weakly-supervised method MIST [55] as our base detector. MIST follows [69] which adapts the detection paradigm of Fast R-CNN [32] to weak annotations. It leverages pre-computed region proposals extracted from unsupervised proposal algorithms, such as Selective Search [71] and EdgeBoxes [83]. In particular, given image $\mathbf{I}$ which has only weak labels q (class labels) and its set of region proposals $\mathcal{R}$, simply called regions, the detection network extracts the image features with a CNN backbone and compute for each region a feature vector using a region-wise pooling [32]. Then, the network head(s) on top of the CNN backbone process the extracted region features in order to predict for each of them the object class and modified box coordinates. To build a detector that can be effectively trained using only image-wise labels, MIST has two learning stages, *coarse detection with multiple instance learning* and *detection refinement with pseudo-boxes*, each implemented with different heads but trained simultaneously in an online fashion [69].

The *Multiple Instance Learner* (MIL) head is trained to minimize the multi-label classification loss $\mathcal{L}^{\text{MIL}}$ using weak labels through which it produces classification scores for all regions in $\mathcal{R}$. MIST selects from them the regions with the highest scores (with non-maximum suppression) as coarse predictions, which we denote with $\mathbf{D}^{(0)}$. Then, such predictions are iteratively refined using K consecutive *refinement heads*. Each refinement head $k \in \{1 \dots K\}$ predicts for all regions in $\mathcal{R}$ their classification scores for the $C+1$ classes (C object classes plus 1 background class) and box coordinates per object class. The refinement head k is trained by minimizing:

$$\mathcal{L}_w^{(k)}(\mathbf{I}, \mathcal{R}, \mathbf{D}^{(k-1)}) = \mathcal{L}_{\text{cls}}^{(k)}(\mathbf{I}, \mathcal{R}, \mathbf{D}^{(k-1)}) + \mathcal{L}_{\text{reg}}^{(k)}(\mathbf{I}, \mathcal{R}, \mathbf{D}^{(k-1)}), \tag{2}$$

which combines an adapted instance classification loss, $\mathcal{L}_{\text{cls}}^{(k)}$, and the box regression loss $\mathcal{L}_{\text{reg}}^{(k)}$ of Fast R-CNN [32], using as targets the pseudo-boxes $\mathbf{D}^{(k-1)}$ generated by MIST from the region scores of the previous head. The final loss for image $\mathbf{I}$ is:

$$\mathcal{L}_w = \mathcal{L}^{\text{MIL}}(\mathbf{I}, \mathcal{R}, \mathbf{q}) + \sum_{k=1}^{K} \mathcal{L}_w^{(k)}(\mathbf{I}, \mathcal{R}, \mathbf{D}^{(k-1)}). \tag{3}$$

For more details about MIST, please refer to the appendix and [55].

Adding strong annotations. In our proposed approach, we obtain ground-truth bounding boxes for *very few* images in the set $\mathcal{I}$. In order to integrate such strong annotations to the weakly-supervised framework, we simply replace the pseudo-annotations in Equation 2 with box annotations $\mathbf{G}$, now supposed available for image $\mathbf{I}$. The resulting loss for the refinement head k reads $\mathcal{L}_s^{(k)}(\mathbf{I}, \mathcal{R}, \mathbf{G}) = \mathcal{L}_{\text{cls}}^{(k)}(\mathbf{I}, \mathcal{R}, \mathbf{G}) + \mathcal{L}_{\text{reg}}^{(k)}(\mathbf{I}, \mathcal{R}, \mathbf{G})$, and the final loss on image $\mathbf{I}$ in this case is $\mathcal{L}_s = \mathcal{L}^{\text{MIL}}(\mathbf{I}, \mathcal{R}, \mathbf{q}) + \sum_{k=1}^{K} \mathcal{L}_s^{(k)}(\mathbf{I}, \mathcal{R}, \mathbf{G})$. As such, during the fine-tuning of the detector M^0, we use $\mathcal{L}_w$ to train on images for which only class labels are available and $\mathcal{L}_s$ when images are provided with bounding-boxes.

Difficulty-aware proposal sampling. In this framework, we use thousands of pre-computed proposals in $\mathcal{R}$ for each image. This is necessary when no box annotations are provided. However, when ground-truth boxes are available, better training can be achieved by sampling a smaller number of proposals [32,56]. In particular, we select a subset of 512 proposals with 25% of *positive* boxes, i.e., those whose IoU with one of the ground-truth boxes exceeds 0.5, and 75% of *negative* boxes, i.e., those whose IoU ≤ 0.3 with all ground-truth boxes. However, we have noticed that negatives are over-sampled from the background or often appear uninformative. We propose to improve negative proposal sampling by using the network predictions to select those classified as objects. We perform a first forward pass and average classification scores obtained over the K refinement heads; we then apply row-wise softmax and select proposals with the highest class scores, excluding background. We show in our experiments that this sampling method allows better training and yields better performance.

4 Experimental Results

In this section, we first introduce the general setup of our experiments. We then present an ablation study of different components of BiB before comparing BiB to different existing active learning strategies. Finally, we demonstrate the effectiveness of our proposed pipeline compared to the state of the art.

4.1 Experimental Setup

Datasets and evaluation. We evaluate our method on two well-known object detection datasets, Pascal VOC2007 [24] (noted VOC07) and COCO2014 [47] (COCO). Following previous works [6,55], we use the *trainval* split of VOC07 for training and the *test* split for evaluation, respectively containing 5011 and 4952 images. On COCO, we train detectors with the *train* split (82783 images) and evaluate on the *validation* split (40504 images) following [6]. We use the average precision metrics AP50 and AP, computed respectively with an IoU threshold of 0.5 and with thresholds in [0.5 : 0.95]. We report results corresponding to N-shot experiments – where $N \times C$ images are selected – and N% experiments, where about N percents of the training set are selected to be fully-annotated.

Architecture. Though BiB can be applied on any weakly-supervised detector, we use MIST [55] as our base weakly-supervised detector for it has public code and has been shown to be a strong baseline. We modify MIST to account for images containing bounding box annotations during training as detailed in Section 3.4. The Concrete Drop Block (CDB) [55] technique is used in MIST in experiments on VOC07 but dropped in COCO experiments to save computational cost. We use our difficulty-aware proposal sampling in all experiments unless stated otherwise. We train with a batch size of 8 during training and a learning rate of 4e−4 for MIST and 4e−6 for CDB when the latter is used. During training, images are drawn from the sets of images with weak and strong annotation uniformly at random such that the numbers of weakly- and fully-annotated images considered are asymptotically equal.

Active learning setup. We emulate an active learning setup by ignoring available bounding box annotations of images considered weakly annotated in our experiments. On both dataset, we run MIST [55] three times to account for the training's instability and obtain three base weakly-supervised detectors. We fine-tune each base weakly-supervised detector twice on VOC07 and once on COCO, giving respectively 6 and 3 repetitions. We always report averaged results, and in some cases also their standard deviation. Detailed results for all experiments are provided in supplemental material. The number of fine-tuning iterations is scaled linearly with the number of strong images in the experiment. Concretely, the base weakly-supervised detector is fine-tuned over 300 iterations for every 50 strong images in VOC07 and 1200 iterations for every 160 images on COCO.

Active learning baselines. We compare BiB to existing active learning strategies. We first compare our method to random selections, either uniform sampling (*u-random*) or balanced per class sampling (*b-random*). We compare to uncertainty-based selection and aggregate box entropy scores per image using sum or max pooling, noted *entropy-sum* and *entropy-max* respectively. Finally, we leverage weak detection losses to select high impact images (*loss*). We report here results obtained with the detection loss of the last refinement branch in MIST, which we find outperforms others losses; a detailed comparison can be found in supplemental material along with a complete description of other methods. We also use the greedy version of the *core-set* selection method [58]; and a weighted version that weights distances in core-set with uncertainty scores (entropy-max) [35], named *coreset-ent*. For our BiB, we set $\mu = 3$ and $\delta = 0.8$, and provide a study on their influence in the supplemental material.

4.2 Ablation Studies

We perform in Table 1 an ablation study to understand the relative importance of the difficulty-aware proposal sampling (*DifS*), the selection based on k-means++ initialization and the use of box-in-box pairs in our method. The second row corresponds to *u-random*. We apply the diversity selection (e.g., following k-means++ initialization) on image-level features, predictions, and BiB pairs. The experiments are conducted on VOC07 and for each variant of our method, we perform five active learning cycles with a budget of 50 images per cycle. It appears that *DifS* significantly improves results over both random and BiB, confirming that targeting the model's most confusing regions is helpful. K-means++ initialization does not help when applied on image-level features but

Table 1: Ablation study. Results in AP50 on VOC07 with 5 cycles and a budget $B = 50$. We provide averages and standard deviation results over 6 repetitions. *DifS* stands for the difficulty-aware region sampling module. Images are selected by applying k-means++ init. (*K selection*) on image-level features (*im.*), confident predictions' features (*reg.*) or BiB pairs.

DifS	K selection im.	reg.	BiB	Number of images annotated 50	100	150	200	250
				56.3 ± 0.4	58.0 ± 0.5	58.9 ± 0.4	60.0 ± 0.3	60.5 ± 0.4
✓				56.5 ± 0.4	58.4 ± 0.4	59.3 ± 0.7	60.2 ± 0.4	61.1 ± 0.5
✓	✓			57.1 ± 0.4	58.3 ± 0.5	59.3 ± 0.6	59.8 ± 0.4	60.3 ± 0.4
✓		✓		**58.4 ± 0.4**	60.2 ± 0.4	61.5 ± 0.6	62.6 ± 0.4	**63.4 ± 0.3**
			✓	57.9 ± 0.7	60.1 ± 0.4	61.2 ± 0.5	62.1 ± 0.5	62.6 ± 0.4
✓			✓	**58.5 ± 0.8**	**60.8 ± 0.5**	**61.9 ± 0.4**	**62.9 ± 0.5**	**63.5 ± 0.4**

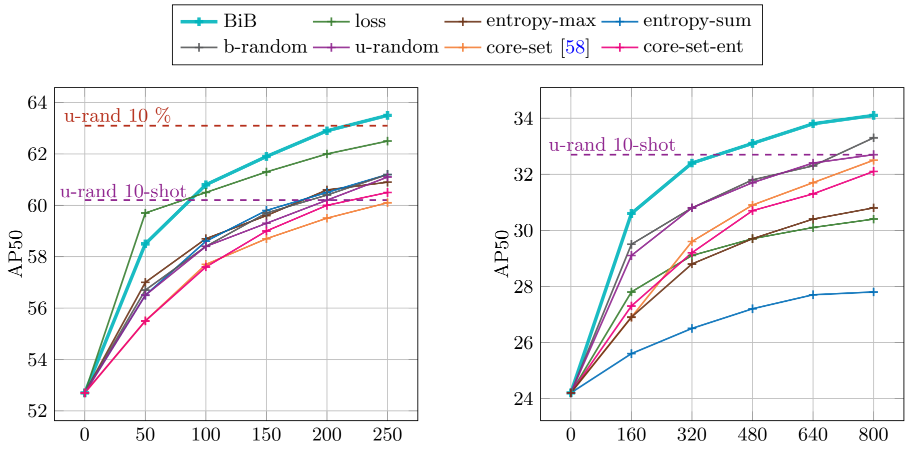
yields significant performance boosts over random when combined with region-level features. Finally, the use of BiB pairs shows consistent improvements over *region*, confirming our choices in BiB's design.

4.3 Comparison of Active Strategies

In order to compare BiB to baselines, we conduct 5 active learning cycles with a budget of $B = 50$ images (1% of the training set) per cycle on VOC07 and of $B = 160$ images (0.2% of the training set, 2 fully annotated images per class on average) on COCO. We present results in Figure 3. The detailed numbers are provided in the supplemental material. It can be seen that the ranking of the examined baseline methods w.r.t. their detection performance is different on the two datasets. This is explained by the fact that the two datasets have different data statistics. COCO dataset contains many cluttered images, with an average of 7.4 objects in an image, and VOC07 depicts simpler scenes, with an average of only 2.4 objects. However, BiB consistently improves over other baselines.

Results on VOC07 (Figure 3a) show that BiB and *loss* significantly outperform every method in all cycles. BiB also surpasses *loss* except in the first cycle. Entropy and variants of *random* perform comparably and slightly better than variants of core-set. Balancing the classes consistently improves the performance of random strategy, albeit with a small margin. Interestingly, BiB reaches the performance of *random* at 10% setting ($\approx$ 500 images) with only $\approx$ 200 fully-annotated images. Similarly, it needs fewer than 100 fully annotated images to attain *random*'s performance in the 10-shot ($\approx$ 200 images) setting.

On COCO, BiB again shows consistent improvement over competitors. However, surprisingly, *loss* fares much worse than BiB and even *random*. To understand these results, we present a representative subset of selected images in Figure 4. It appears that images selected by the *loss* strategy tend to depict complex scenes. Many of them are indoors scenes with lots of objects (people, foods, furniture, ...). The supervision brought by these images is both redundant (two many images for certain classes) and insufficient (no or too few images for



(a) Results on VOC07 (b) Results on COCO

Fig. 3: Detection performances of different active learning strategies in our framework on VOC07 [24] (a) and COCO datasets [47] (b). We perform 5 annotation cycles for each strategy with the budget of $B = 50$ on VOC07 and $B = 160$ on COCO. This corresponds to annotating 1% and 0.2% of the training set per cycle respectively for VOC07 and COCO. Dashed lines in purple and red highlight results obtained with *10-shot* and *10%* images selected with *u-random*. Best viewed in color.

others). This result agrees with those obtained in [14,48] on COCO with the predicted loss method [78]. On the other hand, variants of entropy strategy tend to select very difficult images that are outliers and not representative of the training dataset. They do not perform well on COCO, especially *entropy-sum* which obtains significantly worse results than other strategies. This observation is similar to that of [79]. Diversity-based methods fare better than uncertainty-based methods, with *core-set* and *core-set-ent* performing much better than *entropy* variants. Among the latter two methods, *core-set* performs unsurprisingly better than *core-set-ent*, given *entropy*'s bad performance. BiB outperforms all other methods. It obtains significantly better results than *random*, which other methods fail to do. In addition, BiB attains the same performance as *u-random* (see dashed line) with only half as many annotated images, reducing the performance gap (in AP50) between the base weak detector and the fully-supervised Fast RCNN by nearly 70% with only ten fully annotated images per class on average. It can be seen in Figure 4 that BiB selects a diverse set of images that reflect the model's confusion on object extent.

4.4 Comparison to the State of the Art

We compare the 10-shot performance of our proposed method to the state of the art in Table 2. For BiB, we report the performance of previous experiments

Fig. 4: Images selected by BiB, *entropy-max* and *loss* strategies on COCO dataset.

Table 2: Performance of BiB compared to the state of the art on VOC07 ($B = 50$) and COCO ($B = 160$) datasets. The *10-shot* setting corresponds to 4 and 5 AL cycles resp. on VOC07 and COCO. All of the compared methods use VGG16 [61] as the backbone.

Setting	Method	VOC07 AP50	COCO AP50	COCO AP
		100%		
Fully supervised	Fast RCNN [32]	66.9	38.6	18.9
	Faster RCNN [54]	69.9	41.5	21.2
		0%		
WSOD	WSDDN [7]	34.8	-	-
	OICR [69]	41.2	-	-
	C-MIDN [29]	52.6	21.4	9.6
	WSOD2 [80]	53.6	22.7	10.8
	MIST-CDB [55]	54.9	24.3	11.4
	CASD [38]	56.8	26.4	12.8
		10-shot		
Weak & few strong	BCNet [49]	57.1	-	-
	OAM [6]	59.7	31.2	14.9
	Ours (u-rand)	60.2	32.7	16.4
	Ours (BiB)	62.9	34.1	17.2

(Figure 3) at cycle 4 on VOC07 and cycle 5 on COCO. All compared methods use a Fast R-CNN or Faster R-CNN architecture with a VGG16 [61] backbone. Most related to us, OAM [6] and BCNet [49] also seek to improve the performance of weakly-supervised detectors with a few fully-annotated images. We can see that BiB significantly outperforms them in this setting. In particular, on COCO, we observe from Table 2 and Figure 3 that BiB obtains comparable results to 10-shot OAM with only 2 shots (160 images) and significantly better results with 4 shots. Similarly, on VOC07, BiB surpasses the performance of OAM with only a half of the number of fully-annotated images used by the latter. We additionally consider the 10% setting and compare BiB to other baselines on the VOC07 dataset (see Table 3). In this setting, a random selection following our method ('Ours (u-rand)') gives an AP50 of 63.1, outperformed by BiB ('Ours (BiB)') which achieves an AP50 of 65.1. In comparison, our main competitors perform worse: OAM (63.3), BCNet (61.8), EHSOD [26] (55.3) and BAOD [50] (50.9).

Compared to WSOD methods, we obtain significantly better results with a small amount of full annotations. BiB enables a greater boost over weakly-supervised detectors than *random* and narrows significantly the performance gap between weakly-supervised detectors and fully-supervised detectors. It reduces the gap between the state of the art weakly-supervised detector CASD [38] and Fast RCNN [32] by 5.5 times with 10% of the training images fully annotated on VOC07 and by 3.5 times with only 10 fully annotated images on average per class

Table 3: Per-class AP50 results on VOC07. BiB yields significant boosts in hard classes such as *bottle*, *chair*, *table* and *potted plant*. Results of MIST are the average of three runs using the authors' public code and differ from the numbers in the original paper.

method	sup.	aero	bike	bird	boat	bottl	bus	car	cat	chair	cow	table	dog	horse	moto	pers	plant	sheep	sofa	train	tv	mean
MIST*	✗	69.0	75.6	57.4	22.5	24.8	71.5	76.1	55.9	27.6	70.3	43.9	37.5	50.8	**75.9**	18.5	23.9	60.8	54.7	69.3	68.1	52.7
BAOD [50]	10%	51.6	50.7	52.6	41.7	36.0	52.9	63.7	69.7	34.4	65.4	22.1	66.1	63.9	53.5	59.8	24.5	60.2	43.3	59.7	46.0	50.9
BCNet [49]	10%	64.7	73.1	55.2	37.0	39.1	**73.3**	74.0	75.4	35.9	69.8	56.3	**74.7**	77.6	71.6	66.9	25.4	61.0	61.4	**73.8**	69.3	61.8
OAM [6]	10%	65.6	73.1	59.0	**49.4**	42.5	72.5	78.3	**76.4**	35.4	72.3	57.6	73.6	**80.0**	72.5	**71.1**	28.3	64.6	55.3	71.4	66.2	63.3
Ours (u-r.)	10%	**70.5**	77.2	62.3	38.5	38.5	72.3	**79.4**	73.6	38.6	73.8	55.7	66.5	71.4	75.3	65.5	33.8	65.4	62.7	72.3	69.7	63.1
Ours (BiB)	10%	68.9	**78.1**	**62.7**	41.4	**47.8**	72.4	**79.2**	70.3	**44.9**	**74.7**	**66.2**	62.2	72.1	**75.6**	69.8	**43.1**	**66.2**	**65.0**	71.4	**70.7**	**65.1**

on COCO. This is arguably a better trade-off between detection performance and data efficiency than both weakly- and fully-supervised detectors.

Per-class study. Additionally, we present in Table 3 the per-class results for different methods on VOC07. It can be seen that variants of our approach (u-random and BiB) consistently boost the performance on all classes over MIST [55] (except on *aeroplane* and *motorbike* where they perform slightly worse than MIST). Notably, BiB yields larger boosts on *hard* classes such as *table* (+23 points w.r.t. our baseline MIST), *chair* (+17.3), *bottle* (+23) and *potted plant* (+19.2). On those classes, a random selection with our approach is worse than BiB by more than 7 points. Overall, BiB obtains the best results on most classes.

5 Conclusion and Future Work

We propose a new approach to boost the performance of weakly-supervised detectors using a few fully annotated images selected following an active learning process. We introduce BiB, a new selection method specifically designed to tackle failure modes of weakly-supervised detectors and show a significant improvements over random sampling. Moreover, BiB is effective on both VOC07 and COCO datasets, narrowing significantly the performance between weakly- and fully-supervised object detectors, and outperforming all methods mixing many weak and a few strong annotations in the low annotation regime.

In this work, we combine weakly-supervised and active learning for reducing human annotation effort for object detectors. There are other types of methods that require no annotation at all, such as unsupervised object discovery [60,73] and self-supervised pre-training [9,11], that could help improving different component of our pipeline, e.g., region proposals or the detection architecture. Future work will be dedicated to improving our approach by following those directions.

Acknowledgements. This work was supported in part by the Inria/NYU collaboration, the Louis Vuitton/ENS chair on artificial intelligence and the French government under management of Agence Nationale de la Recherche as part of the "Investissements davenir" program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute). It was performed using HPC resources from GENCIIDRIS (Grant 2021-AD011013055). Huy V. Vo was supported in part by a Valeo/Prairie CIFRE PhD Fellowship.

References

1. Agarwal, S., Arora, H., Anand, S., Arora, C.: Contextual diversity for active learning. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 137–153. Springer (2020) 2, 4
2. Arthur, D., Vassilvitskii, S.: K-means++: The advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). p. 10271035 (2007) 7
3. Arun, A., Jawahar, C., Kumar, M.P.: Dissimilarity coefficient based weakly supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3
4. Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A.: Deep batch active learning by diverse, uncertain gradient lower bounds. In: Proceedings of the International Conference on Learning Representations (ICLR) (2020) 4, 7
5. Beluch, W.H., Genewein, T., Nürnberger, A., Köhler, J.M.: The power of ensembles for active learning in image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 4
6. Biffi, C., McDonagh, S.G., Torr, P.H.S., Leonardis, A., Parisot, S.: Many-shot from low-shot: Learning to annotate using mixed supervision for object detection. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 2, 4, 9, 13, 14
7. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 1, 2, 3, 13
8. Brust, C.A., Kading, C., Denzler, J.: Active learning for deep object detection. In: Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP) (2019) 2, 4
9. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the International Conference on Computer Vision (ICCV) (2021) 14
10. Chen, L., Yang, T., Zhang, X., Zhang, W., Sun, J.: Points as queries: Weakly semi-supervised object detection by points. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 8819–8828 (2021) 4
11. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020) 14
12. Chitta, K., Alvarez, J.M., Lesnikowski, A.: Large-scale visual active learning with deep probabilistic ensembles. arXiv preprint arXiv:1811.03575 (2019) 4
13. Cho, M., Kwak, S., Schmid, C., Ponce, J.: Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 2
14. Choi, J., Elezi, I., Lee, H.J., Farabet, C., Alvarez, J.M.: Active learning for deep object detection via probabilistic modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 2, 4, 12
15. Cinbis, R., Verbeek, J., Schmid, C.: Weakly supervised object localization with multi-fold multiple instance learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2017) 1, 3
16. Desai, S.V., Lagandula, A.C., Guo, W., Ninomiya, S., Balasubramanian, V.N.: An adaptive supervision framework for active learning in object detection. In: Proceedings of the British Machine Vision Conference (BMVC) (2019) 4

17. Deselaers, T., Alexe, B., Ferrari, V.: Localizing objects while learning their appearance. In: Proceedings of the European Conference on Computer Vision (ECCV). p. 452466 (2010) 3
18. Diba, A., Sharma, V., Pazandeh, A., Pirsiavash, H., Van Gool, L.: Weakly supervised cascaded convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 3
19. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence **89**(12), 3171 (1997) 3
20. Ebrahimi, S., Gan, W., Salahi, K., Darrell, T.: Minimax active learning. ArXiv **abs/2012.10467** (2020) 4
21. Ebrahimi, S., Sinha, S., Darrell, T.: Variational adversarial active learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 4
22. Elezi, I., Yu, Z., Anandkumar, A., Leal-Taixe, L., Alvarez, J.M.: Not all labels are equal: Rationalizing the labeling costs for training object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 4
23. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results 1
24. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL visual object classes challenge 2007 (VOC2007) results (2007) 1, 3, 9, 12
25. Fan, Q., Zhuo, W., Tang, C.K., Tai, Y.W.: Few-shot object detection with attention-rpn and multi-relation detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4013–4022 (2020) 2
26. Fang, L., Xu, H., Liu, Z., Parisot, S., Li, Z.: Ehsod: Cam-guided end-to-end hybrid-supervised object detection with cascade refinement. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). pp. 10778–10785 (2020) 4, 13
27. Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. arXiv preprint arXiv:1703.02910 (2017) 4
28. Gao, M., Zhang, Z., Yu, G., Ark, S., Davis, L., Pfister, T.: Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 4
29. Gao, Y., Liu, B., Guo, N., Ye, X., Wan, F., You, H., Fan, D.: C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In: Proceedings of the International Conference on Computer Vision (ICCV) (2019) 1, 13
30. Geifman, Y., El-Yaniv, R.: Deep active learning over the long tail. ArXiv **abs/1711.00941** (2017) 4
31. Gidaris, S., Komodakis, N.: Object detection via a multi-region and semantic segmentation-aware cnn model. In: Proceedings of the International Conference on Computer Vision (ICCV) (2015) 1
32. Girshick, R.: Fast R-CNN. In: Proceedings of the International Conference on Computer Vision (ICCV) (2015) 1, 8, 9, 13, 14
33. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR) (2014) 1
34. Gissin, D., Shalev-Shwartz, S.: Discriminative active learning. ArXiv **abs/1907.06347** (2019) 4
35. Haussmann, E., Fenzi, M., Chitta, K., Ivanecky, J., Xu, H., Roy, D., Mittel, A., Koumchatzky, N., Farabet, C., Alvarez, J.M.: Scalable active learning for object

detection. In: Proceedings of the IEEE Intelligent Vehicles Symposium (IV) (2020) 2, 4, 7, 10

36. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. In: Proceedings of the International Conference on Computer Vision (ICCV) (2017) 1
37. Huang, S., Wang, T., Xiong, H., Huan, J., Dou, D.: Semi-supervised active learning with temporal output discrepancy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 4
38. Huang, Z., Zou, Y., Kumar, B., Huang, D.: Comprehensive attention self-distillation for weakly-supervised object detection. In: Advances in Neural Information Processing Systems (NeurIPS) (2020) 1, 3, 13
39. Jeong, J., Lee, S., Kim, J., Kwak, N.: Consistency-based semi-supervised learning for object detection. In: Advances in Neural Information Processing Systems (NeurIPS) (2019) 1, 3
40. Jie, Z., Wei, Y., Jin, X., Feng, J., Liu, W.: Deep self-taught learning for weakly supervised object localization. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 3
41. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8420–8429 (2019) 2
42. Kao, C.C., Lee, T.Y., Sen, P., Liu, M.Y.: Localization-aware active learning for object detection. In: Proceedings of the Asian Conference on Computer Vision (ACCV) (2018) 4
43. Karlinsky, L., Shtok, J., Harary, S., Schwartz, E., Aides, A., Feris, R., Giryes, R., Bronstein, A.M.: Repmet: Representative-based metric learning for classification and few-shot object detection. In: roposal learning for semi. pp. 5197–5206 (2019) 2
44. Kumar, M., Packer, B., Koller, D.: Self-paced learning for latent variable models. In: Advances in Neural Information Processing Systems (NIPS) (2010) 3
45. Li, Y., Huang, D., Qin, D., Wang, L., Gong, B.: Improving object detection with selective self-supervised self-training. Proceedings of the European Conference on Computer Vision (ECCV) (2020) 3
46. Lin, T.Y., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 936–944 (2017) 4
47. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollr, P., Zitnick, L.: Microsoft COCO: common objects in context. In: Proceedings of the European Conference on Computer Vision (ECCV) (2014) 1, 3, 9, 12
48. Liu, Z., Ding, H., Zhong, H., Li, W., Dai, J., He, C.: Influence selection for active learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9274–9283 (2021) 4, 12
49. Pan, T., Wang, B., Ding, G., Han, J., Yong, J.: Low shot box correction for weakly supervised object detection. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). p. 890896 (2019) 2, 4, 13, 14
50. Pardo, A., Xu, M., Thabet, A.K., Arbeláez, P., Ghanem, B.: Baod: Budget-aware object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 1247–1256 (2021) 4, 13, 14
51. Radosavovic, I., Dollár, P., Girshick, R.B., Gkioxari, G., He, K.: Data distillation: Towards omni-supervised learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4119–4128 (2018) 1, 3

52. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 1
53. Redmon, J., Farhadi, A.: Yolo9000: Better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 1
54. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS) (2015) 1, 4, 13
55. Ren, Z., Yu, Z., Yang, X., Liu, M.Y., Lee, Y.J., Schwing, A.G., Kautz, J.: Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 1, 2, 3, 6, 8, 9, 10, 13, 14
56. Ren, Z., Yu, Z., Yang, X., Liu, M.Y., Schwing, A.G., Kautz, J.: UFO^2: A unified framework towards omni-supervised object detection. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 4, 9
57. Roy, S., Unmesh, A., Namboodiri, V.P.: Deep active learning for object detection. Proceedings of the British Machine Vision Conference (BMVC) (2018) 4
58. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: Proceedings of the International Conference on Learning Representations (ICLR) (2018) 2, 4, 10, 12
59. Settles, B.: Active Learning Literature Survey. Technical Report, University of Wisconsin-Madison Department of Computer Sciences (2009), `https://minds.wisconsin.edu/handle/1793/60660` 2, 4
60. Siméoni, O., Puy, G., Vo, H.V., Roburin, S., Gidaris, S., Bursuc, A., Pérez, P., Marlet, R., Ponce, J.: Localizing objects with self-supervised transformers and no labels. In: Proceedings of the British Machine Vision Conference (BMVC) (2021) 2, 14
61. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations (ICLR) (2015) 13
62. Sivic, J., Russell, B., Efros, A., Zisserman, A., Freeman, W.: Discovering objects and their location in images. In: Proceedings of the International Conference on Computer Vision (ICCV) (2005) 2
63. Sohn, K., Zhang, Z., Li, C.L., Zhang, H., Lee, C.Y., Pfister, T.: A simple semi-supervised learning framework for object detection. In: arXiv:2005.04757 (2020) 1, 3
64. Song, H.O., Girshick, R., Jegelka, S., Mairal, J., Harchaoui, Z., Darrell, T.: On learning to localize objects with minimal supervision (2014) 3
65. Song, H.O., Lee, Y.J., Jegelka, S., Darrell, T.: Weakly-supervised discovery of visual pattern configurations. In: Advances in Neural Information Processing Systems (NIPS) (2014) 3
66. Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C.: Fsce: Few-shot object detection via contrastive proposal encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7352–7362 (2021) 2
67. Tang, J., Lewis, P.H.: Non-negative matrix factorisation for object class discovery and image auto-annotation. In: Proceedings of the International Conference on Content-Based Image and Video Retrieval (CIVR) (2008) 2
68. Tang, P., Wang, X., Bai, S., Shen, W., Bai, X., Liu, W., Yuille, A.: Pcl: Proposal cluster learning for weakly supervised object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **42**(1), 176–191 (2020) 3

69. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 1, 2, 3, 8, 13
70. Tang, P., Ramaiah, C., Xu, R., Xiong, C.: Proposal learning for semi-supervised object detection. Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV) pp. 2290–2300 (2021) 3
71. Uijlings, J., van de Sande, K., Gevers, T., Smeulders, A.: Selective search for object recognition. International Journal on Computer Vision (2013) 3, 8
72. Vo, H.V., Bach, F., Cho, M., Han, K., LeCun, Y., Pérez, P., Ponce, J.: Unsupervised image matching and object discovery as optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 2
73. Vo, H.V., Pérez, P., Ponce, J.: Toward unsupervised, multi-object discovery in large-scale image collections. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 14
74. Vo, H.V., Sizikova, E., Schmid, C., Pérez, P., Ponce, J.: Large-scale unsupervised object discovery. In: Advances in Neural Information Processing Systems 34 (NeurIPS) (2021) 2
75. Wan, F., Liu, C., Ke, W., Ji, X., Jiao, J., Ye, Q.: C-mil: Continuation multiple instance learning for weakly supervised object detection. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3
76. Wang, K., Yan, X., Zhang, D., Zhang, L., Lin, L.: Towards human-machine cooperation: Self-supervised sample mining for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018) 1, 3
77. Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., Liu, Z.: End-to-end semi-supervised object detection with soft teacher. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 3
78. Yoo, D., Kweon, I.S.: Learning loss for active learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 4, 12
79. Yuan, T., Wan, F., Fu, M., Liu, J., Xu, S., Ji, X., Ye, Q.: Multiple instance active learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 2, 4, 12
80. Zeng, Z., Liu, B., Fu, J., Chao, H., Zhang, L.: Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019) 1, 13
81. Zhang, B., Li, L., Yang, S., Wang, S., Zha, Z., Huang, Q.: State-relabeling adversarial active learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 8753–8762 (2020) 4
82. Zhdanov, F.: Diverse mini-batch active learning. ArXiv **abs/1901.05954** (2019) 4
83. Zitnick, L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: Proceedings of the European Conference on Computer Vision (ECCV) (2014) 3, 8
84. Zoph, B., Ghiasi, G., Lin, T.Y., Cui, Y., Liu, H., Cubuk, E.D., Le, Q.V.: Rethinking pre-training and self-training. Advances in Neural Information Processing Systems (NeurIPS) (2020) 3