# Box-Level Active Detection

Mengyao Lyu[1,2,3]   Jundong Zhou[1,2,3]   Hui Chen[1,2]   Yijie Huang[4]   Dongdong Yu[4]
Yaqian Li[4]   Yandong Guo[4]   Yuchen Guo[1,2]   Liuyu Xiang[5*]   Guiguang Ding[1,2*]
[1]Tsinghua University   [2]BNRist   [3]Hangzhou Zhuoxi Institute of Brain and Intelligence
[4]OPPO Research Institute   [5]Beijing University of Posts and Telecommunications

{mengyao.lyu,jundong.zhou}@outlook.com {huangyijie,yudongdong,liyaqian,guoyandong}@oppo.com
{jichenhui2012,yuchen.w.guo}@gmail.com xiangly@bupt.edu.cn dinggg@tsinghua.edu.cn

## Abstract

*Active learning selects informative samples for annotation within budget, which has proven efficient recently on object detection. However, the widely used active detection benchmarks conduct **image-level evaluation**, which is unrealistic in human workload estimation and biased towards crowded images. Furthermore, existing methods still perform **image-level annotation**, but equally scoring all targets within the same image incurs waste of budget and redundant labels. Having revealed above problems and limitations, we introduce a **box-level active detection** framework that controls a box-based budget per cycle, prioritizes informative targets and avoids redundancy for fair comparison and efficient application.*

*Under the proposed box-level setting, we devise a novel pipeline, namely **Complementary Pseudo Active Strategy (ComPAS)**. It exploits both human annotations and the model intelligence in a complementary fashion: an efficient input-end committee queries labels for informative objects only; meantime well-learned targets are identified by the model and compensated with pseudo-labels. ComPAS consistently outperforms 10 competitors under 4 settings in a unified codebase. With supervision from labeled data only, it achieves 100% supervised performance of VOC0712 with merely 19% box annotations. On the COCO dataset, it yields up to 4.3% mAP improvement over the second-best method. ComPAS also supports training with the unlabeled pool, where it surpasses 90% COCO supervised performance with 85% label reduction. Our source code is publicly available at https://github.com/lyumengyao/blad.*

## 1. Introduction

Reducing the dependency on large-scale and well-annotated datasets for deep neural networks has received
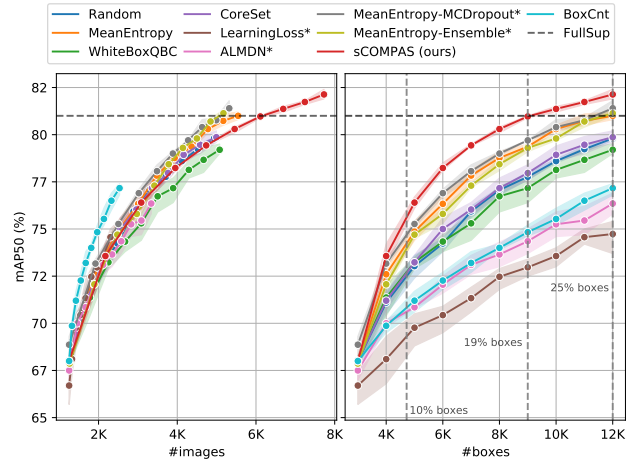


Figure 1. Active detection methods evaluated on VOC0712 under image-level (Left) and box-level (Right) settings. BoxCnt is our hack that simply queries potentially the most crowded images, which demonstrates that image-level evaluation is highly biased. Methods marked with * have specialized detector architectures.

a growing interest in recent years, especially for the detection task, where the box-level annotation is highly demanding. Among data-efficient training schemes, active detection methods [1, 4, 6, 7, 13, 20, 24, 35] iterate over detector training, performance evaluation, informative image acquisition and human annotation. Despite recent progress, previous pool-based active detection methods still consider the subject of interest at the image-level: they conduct **image-level evaluation**, where the budget is controlled by the number of labeled images per cycle; afterwards, they perform exhaustive **image-level annotation**, where all instances of the same image are labeled. Such an image-level framework suffers from unfairness in model performance comparison and leads to a waste of annotation resources.

On the one hand, existing methods under the image-level evaluation assume equal budget for every image. However, in real-world use cases, the workload of annotators is

---

*Corresponding Authors.

measured by bounding boxes [10, 23]. As the image-level budget fails to reflect actual box-based costs, active detection methods are allowed to obscurely gain an advantage by querying box supervision as much as possible until the image-based budget is run out. In fact, according to our experiment shown in Fig. 1L, naively sampling potentially the most crowded images (dubbed as "BoxCnt") can surpass all elaborately designed methods, demonstrating the unfairness of image-level evaluation. On the other hand, during human annotation, simply performing image-level exhaustive annotation is wasteful, since the informativeness of different targets involved in the same image can vary sharply. For example, a salient target of a common category might have been well-learned, whereas a distant or occluded variant could be more informative. As a result, annotating all instances amongst the same image as equals leads to a waste of resources and redundant annotations (See Fig. 4).

After revealing the above problems and limitations, we propose a new **box-level active detection** framework towards fair comparison and non-redundant human annotation. For evaluation, our framework includes a more practical and unbiased criterion that controls the amount of queried boxes per cycle, enabling competing methods to be assessed directly within realistic box-based budgets (as illustrated in Fig. 1R). Considering the annotation, we advocate a box-level protocol that prioritizes top-ranked targets for annotation and discards well-learned counterparts to avoid redundancy. Under the proposed framework, we develop a novel and efficient method namely **Com**plementary **P**seudo **A**ctive **S**trategy (**ComPAS**). It seamlessly integrates human efforts with model intelligence in actively acquiring informative targets via an *input-end committee*, and meantime remedying the annotation of well-learned counterparts using *online pseudo-labels*.

In consideration of the active acquisition, concentrating resources on the most informative targets makes box-level informativeness estimation crucial. Among active learning strategies, multi-model methods, such as Ensemble [2] and MCDropout [9], have demonstrated superiority. Built upon a model-end ensemble, query-by-committee methods select the most controversial candidates based on the voting of model members to minimize the version space [26, 27]. However, directly adapting them to detection not only multiplies the computational cost in the committee construction, but complicates the detection hypothesis ensemble on the box-level. Therefore, to harness the power of diversity without a heavy computational burden, orthogonal to model-end ensembles, we construct an input-end committee during the sampling stage. Variations are drawn from ubiquitous data augmentations and applied to unlabeled candidates, among which each perturbation can be considered as a *cheap but effective committee* member towards version space minimization. When it comes to the box-level

hypothesis ensemble, instead of performing pair-wise label assignment among all members [24], we *reduce the ensemble burden* by analyzing the disagreement between predictions of a reliable reference and other members. Then the disagreement is quantified for both classification and localization to exploit the rich information in annotations.

Later during box-level annotation, the oracle only yields labels for challenging, controversial targets, leaving consistent ones unlabeled. Those unlabeled targets would be considered as the background class during the following training cycles, which severely harms the performance and poses a new challenge. To compensate well-learned targets for missing annotation, we combine sparse ground truths with online pseudo-label generation, where in contrast to active sampling, confident model predictions are accepted as self-supervision signals. The proposed box-level pipeline supports both labeled-only and mixed-supervision learning settings w/ or w/o the unlabeled image pool involved during training, which makes a fairer comparison with fully- and semi-supervised state-of-the-arts (SOTAs).

Our contributions can be summarized as follows:
- We propose a box-level active detection framework, where we control box-based budgets for realistic and fair evaluation, and concentrate annotation resources on the most informative targets to avoid redundancy.
- We develop ComPAS, a novel method that seamlessly integrates model intelligence into human efforts via an input-end committee for challenging target annotation and pseudo-labeling for well-learned counterparts.
- We provide a unified codebase with implementations of active detection baselines and SOTAs, under which the superiority of ComPAS is demonstrated via extensive experiments.

## 2. Related Work

**Active scoring functions.** Pool-based active detection strategies rely on scoring functions to rank sample candidates for annotation, which can be categorized into uncertainty-based [4, 12, 24, 34, 35], diversity-based [1, 25] and hybrid methods [20, 32]. Uncertainty-based methods prioritize unconfident predictions based on posterior probability distributions [2, 9, 35], a specified loss prediction module [34], or Gaussian mixture heads [4]. To avoid sampling bias [5, 21] in batch mode active learning induced by uncertainty, another strategy is to promote diversity for a more representative dataset, which is achieved by core-set selection [1, 25]. However, diversity-based methods incline to sample data points as far as possible to cover the data manifold without considering density. Thus, hybrid methods that make trade-offs between diversity and uncertainty are proposed [20, 32]. Besides adaption from classification-oriented methods, some recent research specially considers
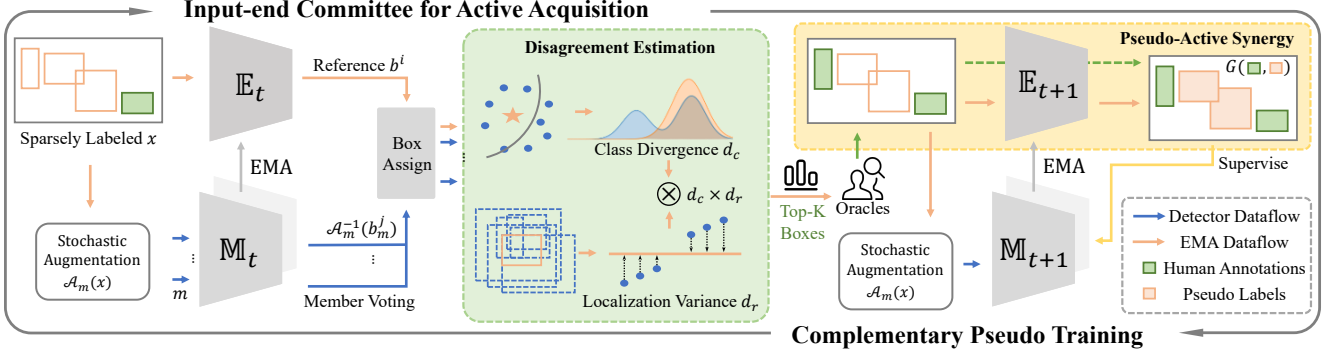
Figure 2. Overview of our ComPAS pipeline for box-level active detection, which iterates between active acquisition via the input-end committee and complementary training based on pseudo-active synergy. Only workflow of sparsely labeled images is shown for generality.

the localization subtask, of which the uncertainty is estimated via the inconsistency between RPN proposals and final predictions [13], or a mixture density model [4]. However, they either impose limitations on the detector architecture, or require certain modifications to it, thus cannot be generalized. In contrast, our localization informativeness is efficiently estimated between stochastic perturbations of candidates without dependency on model architecture.

**Multi-model score ensemble.** Based on the above active scoring functions, multi-model methods ensemble different hypotheses obtained via multiple training repetitions [2], stochastic forward passes [9], different model scales [24] or duplicated detection heads [4]. Despite their effectiveness because of increased variety, ensemble methods have not been widely discussed in detection due to the computational burden and the box-level ensemble difficulty. To reduce the computational cost, some of the previous methods attempted at altering model inputs, such as image flipping [7] and noise interfering [13], but those variations are simple and limited. During the result ensemble, unlike classification methods that can directly average over posterior distributions, existing adaptations towards detection mainly avoid the obstacles by image-level scoring followed by model-level aggregation [4]. However, when applied on the box-level, it requires pair-wise label assignment [24], which further incurs computational cost. In contrast, our input-based committee promotes diversity via stronger positional and color perturbations applied on more input members, and disagreement is efficiently analyzed between a reference and members.

**Implementation and evaluation.** While most recent research on active detection is still evaluated on the image-level, our analysis has revealed that it is unrealistic and heavily biased. Furthermore, we suggest box-level annotation, which is attempted but neither well-explored [31] nor applicable to the in-domain task [29]. To this end, we present a strong pipeline that integrates both human annotations and machine predictions on the box-level. We also note that previous active detection

methods are compared without detector uniformity (*e.g.* SSD [17], Faster R-CNN [22], RetinaNet [15]), learning standardization (*e.g.* runtime settings), benchmark consistency (VOC12/0712 [8], COCO2014/2017 [16]), and supervision differentiation (fully- or semi-supervised). To help advance reproducible research, we introduce a shared implementation of methods based on the same detector, train with similar procedures in a unified codebase, evaluate under the box-level criterion and support both labeled-only and mixed-supervision learning.

## 3. ComPAS for Box-level Active Detection

### 3.1. Problem Formulation

The pipeline of the box-level active detection is initialized at the active learning cycle $t = 0$. A small set of images $L_0$ is randomly sampled and fully annotated with bounding boxes, whereas the majority of images are remained unlabeled $U_0$. Based on the current data pools, a generic object detector $\mathbb{M}(\theta_0)$ is obtained, evaluated and used for inferring on the unlabeled pool. Then a scoring function evaluates the informativeness of each unlabeled candidate and queries an oracle for labels. Different from image-level active detection methods that consider an image as the minimum annotation unit, we actively select top-ranked informative bounding box proposals for annotators to identify the objects of interest within the candidate regions. Such protocol actively prompts annotators with the potential boxes for correction, rather than leaving them passively spotting, marking and verifying all instances for every class among an image, which greatly helps narrow down the spatial search space and semantic options.

Since the first active sampling process $t \geq 1$, sparsely labeled images become available as $S_t$. Then the detector $\mathbb{M}(\theta_t)$ is updated accordingly, with labeled images only ($L_t \cup S_t$), or with all images involved ($L_t \cup S_t \cup U_t$) in the mixed-supervision setting. In the subsequent active acquisition cycles, both sparsely labeled and unlabeled images are evaluated, among which top-ranked boxes with low overlap with existing ground truths are prompted for labels. This it-

eration repeats itself until the stopping criterion is reached.

As illustrated in Fig. 2, under the box-level active detection scenario, we propose a Complementary Pseudo Active Strategy (ComPAS), where the synergy between hard ground truth mining during active sampling (Sec. 3.2) and easy pseudo-label generation (Sec. 3.3) is exploited.

## 3.2. Active Acquisition via Input-end Committee

Ensemble-based active learning has proven effective for classification [2, 9] as well as detection under the box-level evaluation (shown in Fig. 1). The ensemble is also dubbed as a *committee* [24, 27] when the disagreement amongst *member* hypotheses is estimated. However, existing ensemble strategies mainly rely on model parameter duplication, referred to as model-end diversity, which induces extra computational cost. Furthermore, efficiently aggregating bounding box results is also non-trivial. Previous methods adapt the procedure via the instance-level integration to obtain image-level scores, followed by model-level averages [2, 4, 9] which cannot be applied to box-level detection. Or otherwise, aggregating multiple sets of box results would incur pair-wise label assignment, due to the fact that we have to traverse every prediction from all other members to construct a committee for each instance [24].

Orthogonal to the model-end diversity in principle, we instead propose to introduce invariant transformations on the input-end. The posterior disagreement is thus estimated amongst multiple stochastic views of the input, which can be considered as committee members. Drawing variations from data augmentation instead of model ensemble greatly alleviates the burden of training. To achieve complexity reduction for result assignment, inspired by the *consensus* formulation [19, 26], we keep an exponential moving average (EMA) $\mathbb{E}(\theta')$ of the detector $\mathbb{M}(\theta)$ as a *chairman* to generate box *references*:

$$\theta'_{tr} = \alpha \theta'_{tr-1} + (1-\alpha)\theta_{tr}, \tag{1}$$

where $tr$ indicates the training step within one cycle. As shown in Fig. 2, the chairman model generates more reliable predictions [14] $\mathbb{E}(x)$ with regard to the input $x$. Meanwhile, the detector bears more diversity and produces competing hypotheses $\{b_m^j\}$ for a batch of $M$ stochastic augmentations $\mathcal{A}_m(x)$. Based on the chairman predictions as a reference, measuring disagreement between it and all other member hypotheses can effectively reduce the assignment complexity. Note that those augmentations are fed into the network as batches and run in parallel in practice, instead of being forwarded in multiple passes. Next, we detail our disagreement quantification for classification and localization.

**Disagreement on classification.** In estimating the potential value of a box to the classification branch, we prioritize controversial regions in the input space. Specifically, given the box candidates $\{b^i\}$ predicted by the chairman,

member boxes $\{b_m^j\}$ are assigned to each reference box in $\{b^i\}$ using, though not limited to, the detector-defined assignment strategy, such as the max-IoU assigner.

Given a matched pair of boxes $\{b^i, b_m^j\}$, we measure the classification disagreement based on the cross entropy between the one-hot chairman prediction $\mathbf{q}^i$ and the posterior predictive member distribution $\mathbf{q}^j$:

$$d_c^{ij} = -\mathbb{E}_{\mathbf{q}^i}[\log \mathbf{q}^j]. \tag{2}$$

And the disagreement about box $b^i$ is aggregated among $M$ committee members:

$$d_c^i = \frac{1}{M} \sum_m^M \left( \frac{1}{k_{mi}} \sum_j^{k_{mi}} d_c^{ij} \right), \tag{3}$$

where $k_{mi}$ denotes the number of positively matched member predictions in the $m$-th stochastic view. A larger value indicates higher disagreement amongst the input-end committee over a box candidate. It shows that the current model cannot consistently make invariant label predictions under varying degrees of image perturbations, and thus it should be queried for human annotations.

**Disagreement on localization.** While it is straightforward to adopt the prediction distribution as the confidence indicator, $d_c$ can only reflect the committee disagreement on classification. Considering the multi-tasking nature of detection, we are motivated to measure the controversy over localization.

Inspired by [13, 33], with multiple stochastic perturbations applied on the input, we estimate the variation of their box regression results. The intuition behind it is that, if the predicted position is seriously interfered due to randomness, the judgment of the current model on the target concept might not be trustworthy, and thus should be aided by human annotations. The reverse applies when the predictions remain stable despite input variations.

Specifically, with the same chairman-member label assigner used for the classification counterpart, a box reference $b^i$ is matched by multiple candidates $\{b_m^j\}$ generated by $M$ members. We apply respective inverse transformations on those boxes, which are aligned as $\{\mathcal{A}_m^{-1}(b_m^j)\}$ and fed into the localization branch of the chairman model $\mathbb{E}^{reg}$. Then the disagreement over the location of $b^i$ is measured based on the chairman re-calibrated boxes:

$$d_r^i = \frac{1}{4} \sum_k^4 \hat{\sigma}_k(\{\mathbb{E}^{reg}(\mathcal{A}_m^{-1}(b_m^j))\}). \tag{4}$$

In doing so, the localization task is decomposed into four regression tasks based on coordinates. $\hat{\sigma}_k$ represents the standard deviation of the $k$-th coordinate, which is normalized by the average of box height and width.

Overall, for the box-level detection task, our scoring function is formulated as follows:

$$d^i = d_c^i \times d_r^i, \tag{5}$$

based on which we rank all reference boxes for unlabeled regions, and provide labels for top-ranking boxes if they meet certain IoU-based criterion with interested targets during the annotation procedure.

Measuring controversy in both classification and localization exploits human annotations at the bounding box level. Built upon the scoring function, our voting committee is constructed with input-end stochasticity to avoid duplicated training, and the reference formulation further reduces assignment complexity in box-level active acquisition. With controversial regions of the input space being efficiently identified and annotated, the generalization error minimization is gradually achieved in subsequent cycles.

### 3.3. Sparse- and Mixed-Supervision Training

Ever since the first active sampling, sparsely-labeled images are incorporated into the queried pool, where annotated targets provide additional information, and meantime unlabeled ones bring the noise. More severely, our setting prioritizes challenging targets, which we empirically found to be small-sized, distant or occluded, whereas salient and dominant objects are more likely to be left unlabeled. As a result, the label absence of confident objects provides incorrect supervision signals, and proposals associated with them are mistakenly classified as hard negatives. If not properly handled, the sparse annotation problem would have a detrimental effect on the detection performance (See Sec. 4.3).

Despite the significant label absence, as described in Sec. 3.2, the silver lining is that human annotations have been provided for targets that the previous detector fails to interpret, leaving the easier ones to be concerned about. We find the pseudo-label generation *complementary* to it, where targets with confident model predictions are kept for self-training, while challenging targets with uncertain predictions are filtered out. With both active sparse training and pseudo-label generation, we can reduce noise incurred by missing labels, as well as alleviate the error accumulation of pseudo signals. To exploit labeled, sparsely labeled and optionally unlabeled images, we adopt the SOTA pseudo-label generation scheme inspired by [18, 30, 33].

**Supervised loss for labeled images.** Fully labeled images $\{x_l^i\}$ from $L_t$ are fed into the detector and learned in a supervised way:

$$\mathcal{L}_l = \frac{1}{N_l} \sum_i \mathcal{L}_{cls}(x_l^i, y_l^i) + \mathcal{L}_{loc}(x_l^i, t_l^i), \tag{6}$$

where $N_l$ is the number of fully labeled images, $y$ represents ground truth class labels and $t$ denotes corresponding box

locations. $\mathcal{L}_{cls}$ and $\mathcal{L}_{loc}$ represent loss functions used by the detector for classification and localization respectively.

As described in Eq. 1, we keep a temporal smoothed version of the detector, which is also denoted as a *teacher* model. Here we refer to it as *chairman* following Sec. 3.2 for consistency.

**Pseudo-label generation.** The data batch is appended with randomly sampled sparse or unlabeled images if available. The weakly augmented input image $x$ is processed by the chairman to generate pseudo-label candidates $\{b^i\}$, while the strongly augmented version $\mathcal{A}(x)$ is fed into the detector to improve data diversity.

In accordance with our acquisition strategy, pseudo-labels for classification and localization are filtered based on different criteria to ensure precision. Specifically, we apply confidence thresholding with a high threshold $\lambda_c$ to obtain reliable boxes $\{\hat{b}_c^i\}$ for classification. With regard to localization, similar as in Eq. 4, we apply positional perturbations $a(b^i)$ on pseudo-labels for the chairman model to refine. Candidates with predictive fluctuations lower than a threshold $\lambda_r$ are kept to supervise the regression head, which is denoted as $\{\hat{b}_r^i\}$.

**Pseudo-active synergy for sparse images.** Although the confidence thresholding is known to accumulate false negative errors due to the low recall of pseudo-labels, it is less likely to happen in our active sparse training setting. Because annotations of the most challenging targets have been provided. For a sparsely labeled image $x_s$, the pseudo-active synergy is exploited as follows:

$$G(y_s, \hat{y}_{sc}) = y_s \cup \{\hat{y}_{sc}^i \mid IoU(\hat{b}_{sc}^i, b_s^j) \le \lambda_g, \forall b_s^j \in b_s\}, \tag{7}$$

where we supplement sparse ground truth labels $y_s$ with pseudo-labels $\hat{y}_{sc}$ whose corresponding boxes $\hat{b}_{sc}$ have less than $\lambda_g$ jaccard overlap with the ground truth ones. And the same de-duplication process applies to the localization branch, which results in $G(t_s, \hat{t}_{sr})$. The supervision quality for sparse images is thus enhanced after the completion:

$$\mathcal{L}_s = \frac{1}{N_s} \sum_i^{N_s} \mathcal{L}_{cls}(\mathcal{A}(x_s^i), G(y_s^i, \hat{y}_{sc}^i)) + \\ \mathcal{L}_{loc}(\mathcal{A}(x_s^i), G(t_s^i, \hat{t}_{sr}^i)), \tag{8}$$

where $N_s$ is the number of sparsely labeled images.

**Mixed-supervision with unlabeled images.** In the pool-based active learning scenario, unlabeled images are also available during training, which can be utilized to boost performance [7, 20, 35]. Without any human annotation available, the loss function is formulated as follows:

$$\mathcal{L}_u = \frac{1}{N_u} \sum_i^{N_u} (\mathcal{L}_{cls}(\mathcal{A}(x_u^i), \hat{y}_{uc}^i) + \mathcal{L}_{loc}(\mathcal{A}(x_u^i), \hat{b}_{ur}^i)), \tag{9}$$
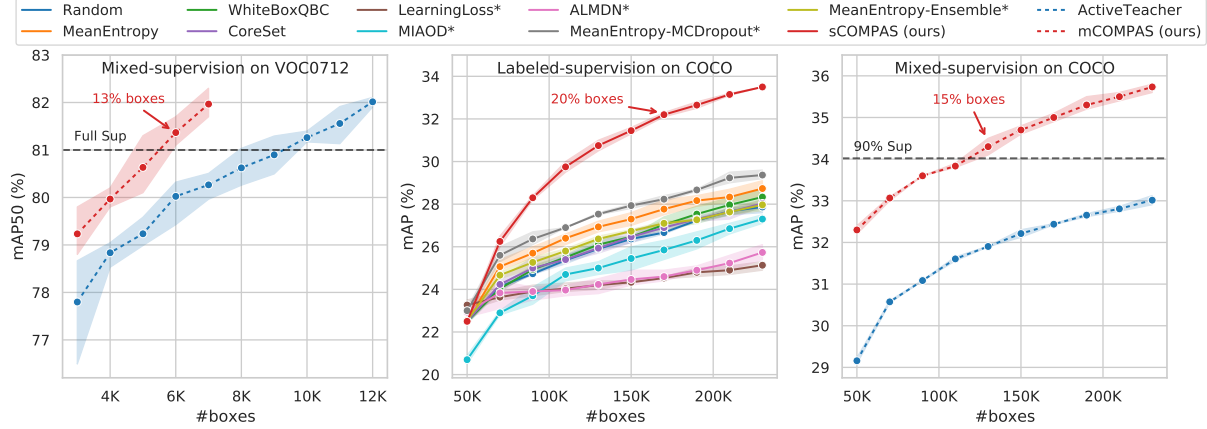
Figure 3. Box-level comparative results on (Left) VOC-semi, (Middle) COCO-sup and (Right) COCO-semi. Solid lines are performed with labeled-supervision, whereas dashed lines indicate training with unlabeled images.

in which $N_u$ denotes the number of unlabeled images, and $\hat{y}_{uc}$ and $\hat{b}_{ur}$ are pseudo-labels and boxes for the two sub-tasks after thresholding respectively.

**Overall training objectives.** In the labeled-only setting, our objective function is formulated as $\mathcal{L}_l + \frac{N_s}{N_l}\mathcal{L}_s$, where we use the sample ratio $\frac{N_s}{N_l}$ to control the contributions of the sparse data flow. Likewise, the objective function for the mixed-supervision setting is $\mathcal{L}_l + \frac{N_s}{N_l}\mathcal{L}_s + \frac{N_u}{N_l}\mathcal{L}_u$.

In grouping hard annotations and easy pseudo-labels together, ComPAS strategy leverages both human brainpower and machine intelligence. It frees object detectors from image-level exhaustive annotations and greatly reduces labor costs.

## 4. Experiments

### 4.1. General Setup

**Datasets.** We study previous and the proposed methods under the box-level evaluation setting on 1) PASCAL VOC0712 [8] dataset, of which the trainval split contains 16,551 images with 40K boxes from 20 classes, and we validate on VOC07 test split; 2) Microsoft COCO [16] dataset, which includes 118K images with about 860K boxes for 80 classes on the train2017 split, and 5K images for validation.

**Baselines and Evaluation.** Depending on the holistic involvement of unlabeled images during training, existing active learning strategies are divided into **labeled-supervised** methods (Random, MeanEntropy, WhiteBoxQBC [24], CoreSet [25], LearningLoss [34], MIAOD[1] [35], ALMDN [4], MCDropout [9], Ensemble [2] and our supervised-ComPAS (denoted as sComPAS)) and **mixed-supervised** ones (ActiveTeacher [20] and our mixed-ComPAS (mComPAS)). Under different supervision and datasets, we refer to our experimental settings as

**VOC-sup, VOC-semi, COCO-sup** and **COCO-semi**.

On VOC0712 dataset, to initialize the labeled pool, we randomly sample images for exhaustive annotation until the budget of 3K boxes is reached, and append 1K boxes per cycle. On COCO, images of 50K boxes are randomly sampled and annotated at first, and 20K boxes are labeled per cycle based on respective query strategies. We conduct all experiments in the main paper for 10 cycles unless the fully supervised (FS) performance has been reached. We report mean average precision @0.5 (mAP50) for VOC0712 and @0.5:0.95 (mAP) for COCO. The mean and standard deviation of results for three independent runs are reported.

**Implementation.** Our detector implementation and training configurations are based on Faster R-CNN [22] with ResNet-50 [11] backbone under the mmdetection [3] codebase. For a fair comparison, we re-implement MeanEntropy, WhiteBoxQBC [24], CoreSet [25], LearningLoss [34], MIAOD [35], ALMDN [4], MCDropout [9] and Ensemble [2] based on their respective public code (if available) or paper descriptions[2]. Details of their implementations can be found in the supplementary material.

In each independent run, the exact same data split is used for all methods, among which methods with specialized architectures have different initial performances (marked with *). During the training of each cycle, we train 12500 or 88000 iterations with a batch-size of 16 for VOC0712 or COCO datasets respectively to be consistent with the fully supervised setting. Unless otherwise stated, SGD optimizer is adopted with learning rates set as 0.01 or 0.02 for VOC0712 or COCO, which is decayed by 10 at 8/12× and 11/12× total iterations. On VOC0712, we train from scratch in each cycle, whereas for COCO we fine-tune from the previous checkpoint for 0.3× iterations with 0.1× learning rate. In terms of the semi-supervision training of com-

---

[1]MIAOD [35] samples an unlabeled pool of the same size as the labeled pool, and thus is excluded from holistic mixed-supervision comparison.

[2]We verify that the performance of our re-implementations with sufficient regularization and augmentation can surpass their reported results.
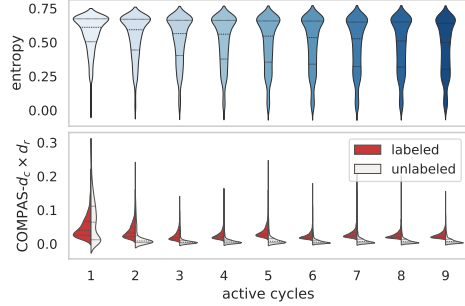
Figure 4. Violin plots showing the box-level score distributions of newly acquired images in each active learning cycle calculated by (Top) image-level MeanEntropy and (Bottom) box-level ComPAS.
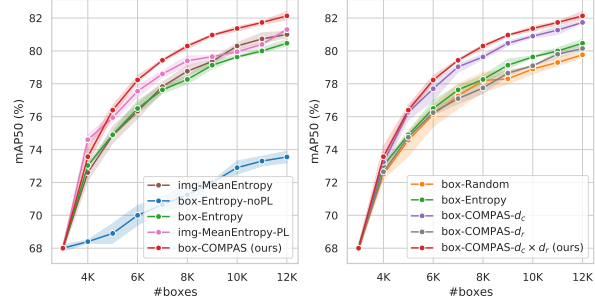


Figure 5. Analysis on (Left) the extension from image-level annotation to the box-level and (Right) alternatives to the box-level active acquisition strategy on VOC0712.

petitors and our mixed-supervision setting, we double the training iterations following the common practice, and leave the rest unchanged. For the proposed method, thresholds $\lambda_c, \lambda_r, \lambda_g$ are set as $0.9, 0.02$ and $0.4$ respectively, which are not specially tuned. We adopt $M$=10 committee members. Diverse augmentations (*e.g.* flip, color distortion) are applied for all methods to make full use of available data.

All experiments on VOC0712 were conducted on NVIDIA RTX 3090, and those of COCO were performed on Tesla V100.

## 4.2. Main Results

**Image-level vs. box-level evaluation**. The comparison between the image-level and box-level evaluation settings under VOC-sup is shown in Fig. 1. Although most of the SOTAs and our hacking method BoxCnt work well under the image-level evaluation, their scoring functions obscurely prioritize crowded images, or their highly ranked targets are severely interfered by invaluable counterparts from the same images. Thus, when evaluated under the box-level criterion, resources wasted on the latter ones emerge, and some previous conclusions are no longer tenable.

**Performance comparison**. Results under VOC-sup and COCO-sup are presented in Fig. 1R and Fig. 3M respectively. As can be seen, within the same box-based budgets, the proposed method outperforms baselines and SOTAs at each active learning cycle by a large margin. Under the labeled-supervision setting for VOC, we obtain 100% supervised performance with only 9K ground truth boxes, which efficiently saves approximately 81% label expenditure. The superiority of our method is also clearly demonstrated on COCO, where sComPAS can exploit rich knowledge from both human annotations and model intelligence. It consistently beats the second-best model-end ensemble-based method by a large margin, and outperforms it by 4.3% mAP in the last cycle in a robust and efficient manner.

In leveraging the unlabeled pool, as shown in Fig. 3LR, we first notice that the proposed active learning strategy retains its overall supremacy: surpassing the 100% supervised

performance requires less than 13% boxes for mComPAS on VOC0712, and on the COCO dataset it only requires 15% boxes to achieve 90% fully supervised capability. In comparison, ActiveTeacher [20] adopts an advanced semi-supervised model [28] for pseudo-label generation, but its acquisition function is solely based on predictive class distributions, which cannot exploit the pseudo-active synergy or well capture sample informativeness.

Under four benchmark settings, the consistent improvements of ComPAS over active learning cycles demonstrate the effectiveness of the input-end committee in identifying informative targets to benefit the detector. Built upon it, our superior results over competitors further show that the proposed pipeline can maximize return over investment by the pseudo-active synergy on the box-level.

## 4.3. Quantitative Analysis

**Redundant annotations in the image-level annotation**. We take MeanEntropy sampling, the best image-level single-model method, to demonstrate the redundancy problem. In Fig. 4, MeanEntropy shows a long-tail phenomenon in the score distributions of newly acquired images, which gets even more acute in later cycles. It indicates that the scoring functions of image-level methods are interfered by less informative targets. Passively annotating them along with highly-ranked ones results in redundancy. In contrast, our box-level method actively annotates valuable targets and leaves the rest unlabeled, maximizing the return over investment. As the iteration proceeds, the divergence between distributions of labeled and unlabeled boxes consistently increases, demonstrating the improvement of model capability and acquisition reliability.

Next, to show the extension of annotation protocol from image-level to the box-level, in Fig. 5L, we take MeanEntropy sampling as a baseline, and apply the box-level annotation protocol, ComPAS model design choices and our scoring function step-by-step under the VOC-sup setting.

**Impact of box-level sparse annotation**. Image-level exhaustive annotation (*img-MeanEntropy*), despite the la-
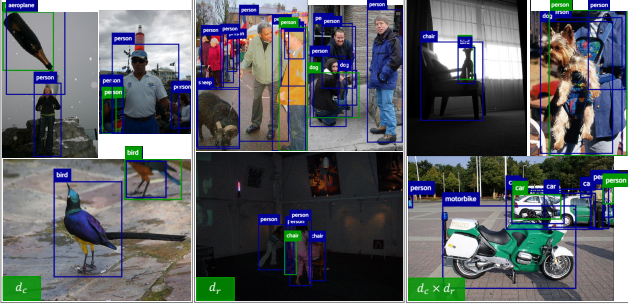
Figure 6. Qualitative results of targets top-ranked by our scoring functions (in green) and complementary pseudo-labels (in blue).

bel redundancy, guarantees the stable training of detection models. When the annotation is disentangled into the box-level, without specific handling (*box-Entropy-noPL*), the missing label problem has a detrimental effect on the model due to the incorrect supervision signal. To alleviate the problem, we introduce pseudo-label generation for sparse images (*box-Entropy*), where sparse labels for challenging targets are supplemented with confident model predictions via an IoU-based grouping strategy, which rectifies supervision signals and significantly boosts performance. However, we notice that it is outperformed by the image-level counterpart in later cycles, which indicates that the box-level annotation poses a greater challenge to the budget allocation, under which entropy-based sampling is not an optimal informativeness estimation solution. Thus, we propose *box-ComPAS*, which is analyzed later in this section.

**ComPAS model design**. The pseudo-label generation scheme designed for sparse annotations of box-level methods can also be used to boost the performance for fully supervised training. To present the effect of it, we simply apply it on all labeled images for image-level MeanEntropy (*img-MeanEntropy*). We first observe that pseudo-labeling is especially effective in the low data regime, but the performance increment is limited in later cycles as the knowledge grows. We also note that our method retains superiority although only sparse images are fed for pseudo-labeling, which demonstrates that the effectiveness of ComPAS is attributed to informative box selection, while pseudo-labeling is mainly used to compensate for acquired knowledge.

**Box-level scoring function**. Under the box-level annotation protocol, we experiment with alternatives to our scoring function during the active acquisition stage, which includes Random, Entropy, our classification disagreement estimation $d_c$ in Eq. 3 alone, our localization disagreement estimation $d_r$ in Eq. 4 alone, and the proposed classification-localization hybrid metric $d_c \times d_r$ presented in Eq. 5. All alternatives are performed with the input-end committee ensemble same as ours. As the results in Fig. 5R suggest, baseline methods, such as entropy-based sampling that performs well for image-level annotation, are

not optimal box-level informativeness indicators. In contrast, $d_c$ estimates the cross entropy between the consensus and member prediction distributions, which well captures the classification informativeness. But built upon $d_c$ and $d_r$, our hybrid metric further incorporates disagreement estimation about the localization subtask, benefiting both detection heads from human annotations. It shows that our acquisition function reflects the challenge of boxes being correctly and robustly detected given the current level of knowledge, so that highly ranked boxes can play a complementary role with pseudo-labels in the subsequent training cycles.

### 4.4. Qualitative Analysis.

The complementarity between actively queried targets (in green) and pseudo-labels (in blue) are visualized in Fig. 6. We present top-ranked boxes scored by our classification metric $d_c$, localization metric $d_r$ as well as the hybrid metric $d_c \times d_r$ respectively, and give the chairman-generated pseudo-labels from the same learning cycle. We empirically find that actively queried targets are more likely to be small, occluded or deviant, where the model fails to guarantee invariant predictions under strong input variations. In contrast, targets left by our scoring function tend to be salient and ubiquitous, whose online pseudo-labels usually get better and better in the next cycles and play a complementary role. More visualizations are shown in the supplementary.

## 5. Conclusion

In this paper, we reveal the pitfalls of image-level evaluation for active detection and propose a realistic and fair box-level evaluation criterion. We then advocate efficient box-level annotation, under which we formulate a novel active detection pipeline, namely Complementary Pseudo Active Strategy (ComPAS) to exploit both human annotations and machine intelligence. It evaluates box informativeness based on the disagreement amongst a near-free input-end committee for both classification and localization to effectively query challenging targets. Meantime, the detector model addresses the sparse training problem by pseudo-label generation for well-learned targets. Under both labeled-only and mixed-supervision settings on VOC0712 and COCO datasets, ComPAS outperforms competitors by a large margin in a unified codebase.

## 6. Acknowledgment

# References

[1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *European Conference on Computer Vision*, pages 137–153. Springer, 2020. 1, 2

[2] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan 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. 2, 3, 4, 6

[3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[4] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, and Jose M. Alvarez. Active learning for deep object detection via probabilistic modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10264–10273, October 2021. 1, 2, 3, 4, 6

[5] Sanjoy Dasgupta. Two faces of active learning. *Theoretical computer science*, 412(19):1767–1781, 2011. 2

[6] Sai Vikas Desai, Akshay L Chandra, Wei Guo, Seishi Ninomiya, and Vineeth N Balasubramanian. An adaptive supervision framework for active learning in object detection. In *BMVC*, 2019. 1

[7] Ismail Elezi, Zhiding Yu, Anima Anandkumar, Laura Leal-Taixe, and Jose M Alvarez. Not All Labels Are Equal: Rationalizing the Labeling Costs for Training Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14492–14501, 2022. 1, 3, 5

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 3, 6

[9] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017. 2, 3, 4, 6

[10] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[12] Tao He, Xiaoming Jin, Guiguang Ding, Lan Yi, and Chenggang Yan. Towards better uncertainty sampling: Active learning with multiple views for deep convolutional neural network. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1360–1365, 2019. 2

[13] Chieh-Chi Kao, Teng-Yok Lee, Pradeep Sen, and Ming-Yu Liu. Localization-aware active learning for object detection. In *Asian Conference on Computer Vision*, pages 506–522. Springer, 2018. 1, 3, 4

[14] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017. 4

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3, 6

[17] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 3

[18] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 5

[19] Andrew McCallum, Kamal Nigam, et al. Employing em and pool-based active learning for text classification. In *ICML*, volume 98, pages 350–358. Madison, 1998. 4

[20] Peng Mi, Jianghang Lin, Yiyi Zhou, Yunhang Shen, Gen Luo, Xiaoshuai Sun, Liujuan Cao, Rongrong Fu, Qiang Xu, and Rongrong Ji. Active Teacher for Semi-Supervised Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14482–14491, 2022. 1, 2, 5, 6, 7

[21] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *arXiv preprint arXiv:2009.00236*, 2020. 2

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3, 6

[23] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Alexander G Schwing, and Jan Kautz. Ufo2: A unified framework towards omni-supervised object detection. In *European Conference on Computer Vision*, pages 288–313. Springer, 2020. 2

[24] Soumya Roy, Asim Unmesh, and Vinay P Namboodiri. Deep active learning for object detection. In *BMVC*, page 91, 2018. 1, 2, 3, 4, 6

[25] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. 2, 6

[26] Burr Settles. Active learning: Synthesis lectures on artificial intelligence and machine learning. *Long Island, NY: Morgan & Clay Pool*, 2012. 2, 4

[27] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992. 2, 4

[28] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 7

[29] Ying-Peng Tang, Xiu-Shen Wei, Borui Zhao, and Sheng-Jun Huang. Qbox: Partial transfer learning with active querying

for object detection. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13, 2021. 3

[30] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 5

[31] Sai Vikas Desai and Vineeth N Balasubramanian. Towards fine-grained sampling for active learning in object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4010–4014, 2020. 3

[32] Jiaxi Wu, Jiaxin Chen, and Di Huang. Entropy-Based Active Learning for Object Detection With Progressive Diversity Constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9397–9406, 2022. 2

[33] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 4, 5

[34] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 93–102, 2019. 2, 6

[35] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5330–5339, 2021. 1, 2, 5, 6

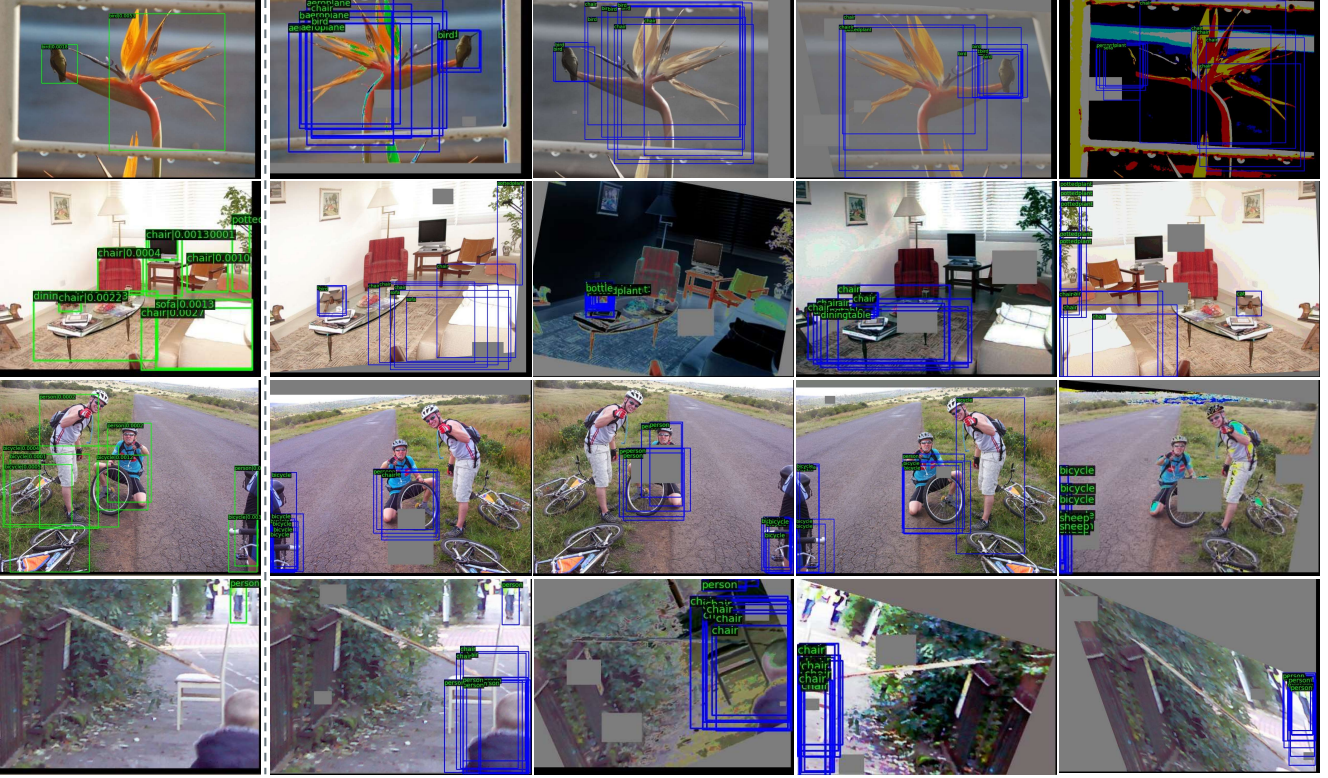# Box-Level Active Detection – Supplementary Material



Figure 1. Reference generated by the chairman model (1st column with boxes in green and box-level acquisition scores), and corresponding member predictions (2nd-5th columns with boxes in blue). Experiments are performed at the first active learning cycle under the VOC-sup setting. We only show top-ranking member predictions for ease of visualization.

## Contents

## 1. Efficacy of the Acquisition Reference

In this section, we validate the use of acquisition reference in terms of both accuracy and efficiency. The comparison is conducted under the VOC-sup setting on the same server with 4 NVIDIA RTX 3090. We set the committee size as 10 in accordance with our main results.

**The existence of the reference**. Without the reference, the *Members2Members* variant constructs a committee for each instance by traversing every prediction from all other members, as done in WhiteBoxQBC [19]. In contrast, with a reference incorporated, the assignment procedure can be reduced to comparing between it and other member hypotheses. As shown in Tab. 1, if facilitated by the reference, such as in our *ChairmanRef2Members*, the inference for acquisition takes about 0.1667 seconds per image in practice. It is approximately $4\times$ faster than the Members2Members ablation, which indicates that the reference is essential when scaling up to larger datasets. Besides, since the augmentations used to construct the committee is diverse and relatively strong, basing the informativeness estimation solely on member outputs is susceptible to noise and randomness. As the detection accuracy shows, with a robust and reliable reference, the quality of active sampling

Figure 2. Reference generated by the chairman model (1st column with boxes in green and box-level acquisition scores), and corresponding member predictions (2nd-5th columns with boxes in blue). Experiments are performed at the first active learning cycle under the COCO-sup setting. We only show top-ranking member predictions for ease of visualization.

| sComPAS Ablations | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Inference Time (s) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Members2Members | 68.00±0.20 | 73.15±0.41 | 75.85±1.26 | 78.00±0.20 | 78.60±0.20 | 79.50±0.34 | 80.00±0.06 | 80.55±0.27 | 81.10±0.20 | 81.20±0.20 | 0.6739 |
| DetectorRef2Members | 68.00±0.20 | 73.40±0.34 | 76.35±0.55 | 77.70±0.34 | 78.75±0.41 | 79.50±0.20 | 79.75±0.13 | 79.95±0.27 | 80.10±0.14 | 80.70±0.28 | 0.1656 |
| ChairmanRef2Members (ours) | 68.00±0.20 | **73.57±0.57** | **76.40±0.30** | **78.23±0.23** | **79.43±0.06** | **80.30±0.17** | **80.97±0.06** | **81.37±0.15** | **81.73±0.12** | **82.13±0.25** | 0.1667 |

Table 1. Ablation study on the existence and the source of the acquisition reference under the VOC-sup setting. The accuracy (%) and time consumption are averaged over 3 runs. Inference Time in seconds denotes the average forward time per image during the acquisition stage.

consistently improves over cycles.

**The source of the reference**. To obtain reliable references as query candidates, the original image is fed into the chairman model, whose predictions are adopted as reference in our proposed ComPAS design. For comparison with the chairman, we experiment with the *Detector-Ref2Members* variant, where the reference is obtained from the detector itself instead of its temporal ensemble. As the results in Tab. 1 show, with almost equal computational cost, the quality of chairman-generated references beat the non-EMA alternative by a large margin.

**Visual inspection of our reference**. In Fig. 1 and Fig. 2, we visualize top-ranked reference hypotheses according to our disagreement scores, as well as member predictions assigned to them. The experiment is performed at the 1st cycle under the VOC-sup and COCO-sup settings respec-

tively. For ease of visualization, we trim member predictions based on their scores, *i.e.*, contributions to the acquisition score of the reference box, so that only top-ranking box predictions are shown. We observe that the chairman model can recognize well-learned targets and locate some salient objects as reference in a stable manner, whereas those perturbed members can bring considerable variations and randomness. Member predictions challenge the judgments of the chairman or supplement with potential candidates, so that controversial regions of the input space can be found. For example, in Fig. 1, the bird in the first row obtains a consensus, whereas the branch next to it is mistakenly recognized by the chairman as a bird with spread wings, which is disputed by the committee. Similar observations can also be made from the camera held by the girl shown in Fig. 2. Besides, the proposed metric also prioritizes targets that are

12

| Cycle | M | | |
|---|---|---|---|
| | 1 | 4 | 10 |
| 0 | 68.00±0.20 | 68.00±0.20 | 68.00±0.20 |
| 1 | 72.97±0.35 | **73.83±0.47** | 73.57±0.57 |
| 2 | 76.37±0.21 | **76.83±0.32** | 76.40±0.30 |
| 3 | 78.10±0.46 | 78.17±0.38 | **78.23±0.23** |
| 4 | 78.93±0.31 | 79.20±0.10 | **79.43±0.06** |
| 5 | 79.67±0.31 | 79.90±0.35 | **80.30±0.17** |
| 6 | 80.33±0.12 | 80.50±0.35 | **80.97±0.06** |
| 7 | 80.87±0.47 | 80.83±0.25 | **81.37±0.15** |
| 8 | 81.37±0.12 | 81.33±0.23 | **81.73±0.12** |
| 9 | 81.70±0.26 | 81.83±0.15 | **82.13±0.25** |

Table 2. Sensitivity to the input-end committee size under the VOC-sup setting.

challenging to localize. Take the 3rd row in Fig. 2 for example: the committee has a disagreement over the train body, which leads to a higher acquisition score. The qualitative results further indicate that the proposed disagreement quantification under strong variations well identifies the input space where the current model neglects. Once actively annotated, they can effectively provide informativeness and guarantee consistent improvements in later model updates.

## 2. Sensitivity to the Input-end Committee Size

As there is no consensus on the appropriate committee size to use [21], we experiment under the VOC-sup setting with a varying number of members $M$. As shown in Tab. 2, although one member prediction can work well under the proposed pipeline, providing more data variations on the input-end helps the model identify more informative and representative samples and provides robustness, which guarantees further improvements and stability in a cheap but effective way.

## 3. Comparison with Ensemble-based Methods

We further compare the proposed ComPAS with well-performed ensemble-based methods, including Ensemble [2] and MCDropout [7]. Following [4], those multi-model methods are implemented based on MeanEntropy, which is also the best single-model baseline in our experiments. The detection accuracy has been presented in Fig. 1, Fig. 3 and Tab. 1 of the main paper, and the numerical results of figures are reported in Tab. 4 and Tab. 5 respectively. In Tab. 3, we detail the size of the ensemble/committee, the required training time and the inference time per image for informativeness estimation. The experiments are conducted under the VOC-sup setting on the same server with 4 NVIDIA RTX 3090.

As can be seen, even when there is only one member, *i.e.* $M = 1$, our method retains its overall supremacy in both effectiveness and efficiency. Built upon it, we provide flexibility in the committee size to suit the needs of different end applications. With more members incorporated, better detection performance can be further pursued via more input

| Methods | M | #Trainable Parameters (M) | Inference Time (s) |
|---|---|---|---|
| MeanEntropy-Ensemble | 3 | 123.51 | 0.0310 |
| MeanEntropy-MCDropout | 25 | 41.17 | 0.2400 |
| | 1 | 41.17 | 0.0277 |
| sComPAS (ours) | 4 | 41.17 | 0.0741 |
| | 10 | 41.17 | 0.1667 |
| | 15 | 41.17 | 0.2582 |

Table 3. Comparison with the ensemble-based methods under the VOC-sup setting. $M$ represents the size of the ensemble/committee following the implementation of [2, 4]. The number of trainable parameters are reported in millions. Inference Time in seconds denotes the average forward time per image during the acquisition stage.
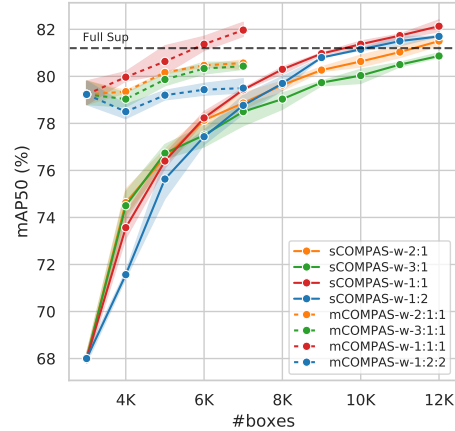


Figure 3. Analysis of the loss weight ratio among different levels of supervision (full-labeled: partial-labeled(: unlabeled)) under the VOC-sup and VOC-semi settings.

variations. Since the committee construction only happens during the acquisition stage, and image perturbations can be processed in one feed-forward pass in practice, the extra costs incurred are marginal in contrast to ensemble [2] and MCDropout [7].

## 4. Analysis of Overall Loss Weights

Our overall objective functions for both training settings give the same relevance to different levels of supervision. To analyze the importance of the fully labeled subset versus others, we specifically finetune the weight of them under both VOC-sup and VOC-semi settings.

Results in Fig. 3 show that fully labeled images are more informative when they dominate the data pool, but increasing their relevance cannot guarantee consistent improvement as the distribution changes along learning cycles. In contrast, re-weighting losses by the sample ratio keeps the method simple but effective.
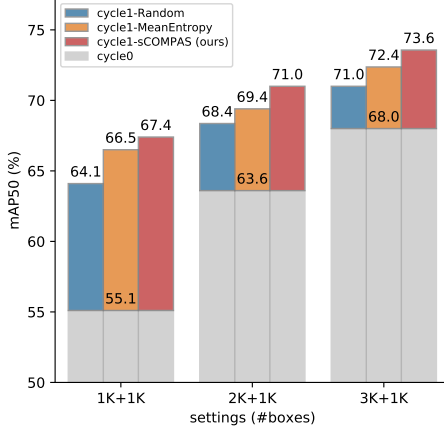
Figure 4. Weaker initialization (cycle0) and the started state (cycle1) under lower-data regimes under the VOC-sup setting.



Figure 5. Box-level comparative results on COCO-sup with a 40K per-cycle budget.

## 5. Performance under Different Active Learning Settings

As active learning is known to be sensitive to settings [13], in this section, we validate our observations under different scenarios.

**Performance with weaker starting points**. In Fig. 4, we start the detector under lower-data regimes without overfitting. Results of the initialization with 3K, 2K and 1K boxes and of the corresponding started states demonstrate that our active sampling strategy is robust to weaker initialization and outperforms competitors.

**Performance with a larger acquisition batch size**. We conduct experiments on the COCO dataset under the COCO-sup setting with a larger acquisition batch size. The iterations are initialized with 50K boxes, the same as the experiments shown in Fig. 3M of the main paper. In each active learning cycle, we append 40K boxes based on respective query and annotation strategies. Results of three independent runs are plotted in Fig. 5 and numerically detailed in Tab. 6 respectively, which show that the proposed method is superior regardless of the acquisition batch size.

## 6. Pseudo-Active Synergy Visualization over Active Learning Cycles

In Fig. 6, We present the iteration of pseudo-labels predicted by the chairman model accompanied by active human annotations across the learning cycles. Images are highlighted in red frames if the active annotation happens in those steps. We find that the proposed acquisition function prioritizes challenging targets, such as small, occluded (*e.g.* cars in the 1st and 4th images) or deviant (*e.g.* the cow in the last images mistakenly recognized as a horse) ones. Meantime, most unlabeled targets can be covered by pseudo-label generation, which gets better in both classification and localization across active learning cycles. Through iterative
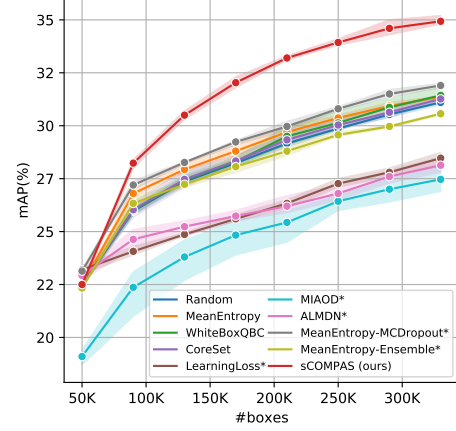
knowledge gain and self-supervision, the synergy between them is effectively exploited.

## 7. Implementation Details

In addition to the general settings of our unified codebase introduced in the main paper, we detail the re-implementation of compared methods in this section.

**FullSup**. For reference, we also report the fully supervised (FS) performance, denoted by FullSup, given the same model, augmentations and runtime settings as in AL experiments. Thus $r\%$ Sup refers to $r\%$ of the FS performance instead of a data split.

**MeanEntropy**. After NMS (Non-maximal suppression), we calculate the entropy of a box candidate based on the predictive probability of its most confident class. Then the uncertainty scores are averaged over the image.

**WhiteBoxQBC** [19]. Following the Algorithm. 1 provided in the paper, we take the NMS outputs of the multiple scales of Faster R-CNN to construct the 'committee', among which all classes and pair-wise bounding box predictions are traversed to estimate the image uncertainty based on predictive *margin*.

**CoreSet** [20]. We apply global average pooling on the multi-level features extracted by the Feature Pyramid Network(FPN) of Faster R-CNN, which are then concatenated as the latent space representation of an image. We implement the k-Center-Greedy algorithm to select unlabeled images during the acquisition stage.

**LearningLoss** [28]. The multi-level features extracted by the Feature Pyramid Network (FPN) of Faster R-CNN are fed into the loss prediction module. The gradient from the loss prediction module is stopped at $0.8\times$ total iterations following the implementation of the paper. We finetune the weight of the learned loss and use 0.3 in consideration of convergence and better performance.

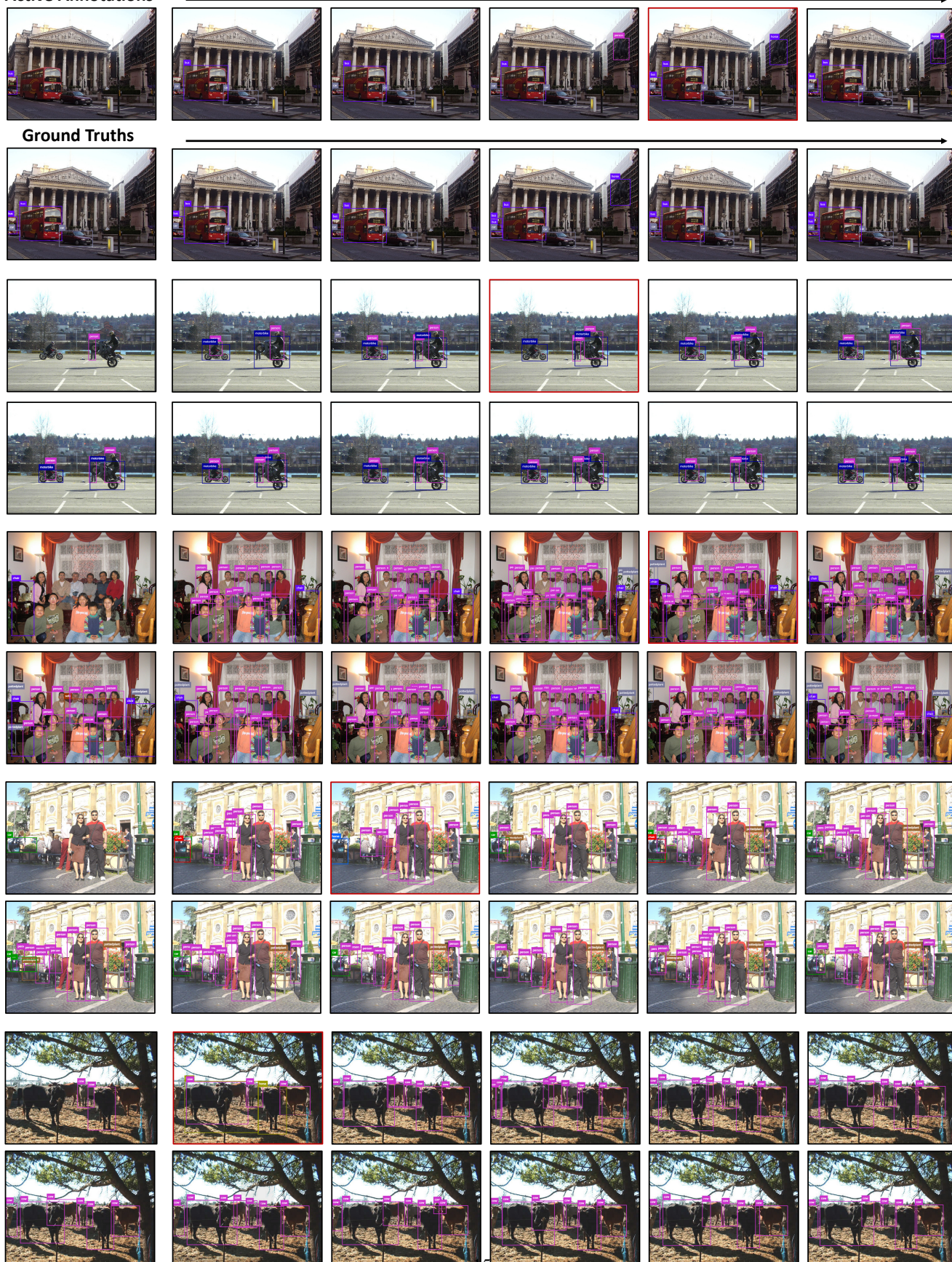**ALMDN** [4] Following the code published by the au-

Figure 6. Iteration of pseudo labels across active learning cycles accompanied by actively annotated targets. For each set of images, we highlight the image in red frame if the active annotation happens in that step.

| Methods | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| | VOC-sup | | | | | | | | | |
| Random | 68.00±0.20 | 71.10±0.56 | 73.03±0.60 | 74.23±0.38 | 75.93±0.31 | 77.07±0.23 | 77.77±0.21 | 78.60±0.10 | 79.23±0.15 | 79.80±0.10 |
| MeanEntropy | 68.00±0.20 | 72.60±0.44 | 74.87±0.50 | 76.33±0.45 | 77.83±0.31 | 78.77±0.31 | 79.37±0.15 | 80.30±0.20 | 80.73±0.32 | 81.00±0.20 |
| WhiteBoxQBC | 68.00±0.20 | 71.37±0.85 | 73.23±0.31 | 74.33±0.47 | 75.30±0.78 | 76.73±0.74 | 77.17±0.68 | 78.13±0.45 | 78.67±0.65 | 79.20±0.26 |
| CoreSet | 68.00±0.20 | 71.20±0.50 | 73.23±0.32 | 75.00±0.53 | 76.03±0.35 | 77.17±0.47 | 77.97±0.40 | 78.93±0.42 | 79.47±0.45 | 79.87±0.31 |
| LearningLoss* | 66.70±0.89 | 68.10±1.25 | 69.77±0.87 | 70.43±0.81 | 71.33±0.87 | 72.47±0.59 | 72.97±0.45 | 73.57±0.51 | 74.57±0.40 | 74.73±0.91 |
| ALMDN* | 67.50±0.71 | 70.00±0.28 | 70.85±0.07 | 72.05±0.64 | 73.10±0.71 | 73.65±0.92 | 74.35±0.92 | 75.25±0.92 | 75.45±1.06 | 76.35±0.78 |
| MeanEntropy-MCDropout* | **68.87±0.59** | 73.17±0.25 | 75.27±0.15 | 76.90±0.20 | 78.07±0.35 | 79.00±0.10 | 79.70±0.00 | 80.40±0.35 | 80.77±0.15 | 81.40±0.35 |
| MeanEntropy-Ensemble* | 67.87±0.23 | 72.07±0.25 | 74.70±0.20 | 75.80±0.44 | 77.30±0.10 | 78.43±0.31 | 79.30±0.17 | 79.80±0.17 | 80.70±0.26 | 81.13±0.15 |
| BoxCnt | 68.00±0.20 | 69.87±0.32 | 71.20±0.56 | 72.27±0.45 | 73.20±0.56 | 74.00±0.17 | 74.83±0.32 | 75.53±0.49 | 76.50±0.46 | 77.17±0.32 |
| sCOMPAS (ours) | 68.00±0.20 | **73.57±0.57** | **76.40±0.30** | **78.23±0.23** | **79.43±0.06** | **80.30±0.17** | **80.97±0.06** | **81.37±0.15** | **81.73±0.12** | **82.13±0.25** |
| | VOC-semi | | | | | | | | | |
| ActiveTeacher | 77.80±1.15 | 78.84±0.27 | 79.24±0.31 | 80.02±0.52 | 80.27±0.28 | 80.62±0.39 | 80.90±0.40 | 81.26±0.12 | 81.56±0.40 | 82.01±0.10 |
| mCOMPAS (ours) | **79.23±0.51** | **79.97±0.21** | **80.63±0.61** | **81.37±0.31** | **81.97±0.31** | - | - | - | - | - |

Table 4. MAP50 and standard deviation (%) under VOC-sup and VOC-semi settings, with 3K boxes for initialization and 1K boxes for each active learning cycle. Results reported are averaged over 3 independent runs. We stop the learning of mCOMPAS after it far exceeds the fully supervised performance (81% mAP50). The best result of each cycle is highlighted in bold.

| Methods | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| | COCO-sup | | | | | | | | | |
| Random | 22.50±0.10 | 24.13±0.12 | 24.73±0.06 | 25.37±0.21 | 25.90±0.17 | 26.37±0.15 | 26.67±0.06 | 27.23±0.06 | 27.67±0.15 | 27.87±0.23 |
| MeanEntropy | 22.50±0.10 | 25.07±0.29 | 25.70±0.26 | 26.40±0.26 | 26.93±0.25 | 27.30±0.30 | 27.77±0.38 | 28.17±0.25 | 28.33±0.32 | 28.73±0.35 |
| WhiteBoxQBC | 22.50±0.10 | 24.03±0.15 | 24.93±0.15 | 25.50±0.10 | 26.10±0.26 | 26.47±0.06 | 27.03±0.32 | 27.53±0.40 | 27.97±0.42 | 28.33±0.47 |
| CoreSet | 22.50±0.10 | 24.23±0.25 | 25.00±0.44 | 25.40±0.17 | 25.93±0.25 | 26.47±0.31 | 26.90±0.17 | 27.30±0.10 | 27.63±0.29 | 28.03±0.15 |
| LearningLoss* | **23.27±0.25** | 23.63±0.23 | 23.90±0.26 | 24.03±0.23 | 24.20±0.20 | 24.33±0.06 | 24.53±0.06 | 24.80±0.20 | 24.90±0.20 | 25.13±0.21 |
| MIAOD* | 20.70±0.28 | 22.90±0.14 | 23.70±0.57 | 24.70±0.42 | 25.00±0.42 | 25.45±0.78 | 25.85±0.64 | 26.30±0.57 | 26.85±0.35 | 27.30±0.28 |
| ALMDN* | 22.93±0.21 | 23.83±0.67 | 23.90±0.35 | 23.97±0.25 | 24.23±0.38 | 24.47±0.38 | 24.60±0.26 | 24.90±0.17 | 25.23±0.35 | 25.73±0.35 |
| MeanEntropy-MCDropout* | 23.00±0.17 | 25.60±0.36 | 26.37±0.31 | 26.90±0.00 | 27.53±0.06 | 27.93±0.12 | 28.23±0.15 | 28.67±0.06 | 29.23±0.25 | 29.37±0.21 |
| MeanEntropy-Ensemble* | 22.53±0.12 | 24.67±0.15 | 25.27±0.25 | 25.80±0.10 | 26.37±0.12 | 26.73±0.06 | 27.10±0.10 | 27.27±0.06 | 27.63±0.15 | 27.97±0.25 |
| sCOMPAS (ours) | 22.50±0.10 | **26.25±0.35** | **28.30±0.14** | **29.75±0.35** | **30.75±0.35** | **31.45±0.21** | **32.20±0.14** | **32.65±0.21** | **33.15±0.07** | **33.50±0.00** |
| | COCO-semi | | | | | | | | | |
| ActiveTeacher | 29.16±0.10 | 30.57±0.04 | 31.09±0.01 | 31.61±0.07 | 31.90±0.01 | 32.21±0.13 | 32.44±0.02 | 32.66±0.05 | 32.81±0.08 | 33.01±0.16 |
| mCOMPAS (ours) | **32.30±0.10** | **33.07±0.06** | **33.60±0.00** | **33.83±0.06** | **34.30±0.20** | **34.70±0.10** | **35.00±0.10** | **35.30±0.17** | **35.50±0.10** | **35.73±0.12** |

Table 5. MAP and standard deviation (%) under COCO-sup and COCO-semi settings, with 50K boxes for initialization and 20K boxes for each active learning cycle. Results reported are averaged over 3 independent runs. The best result of each cycle is highlighted in bold.

thors, we turn the detection heads of Faster R-CNN into four Gaussian mixture models (GMMs), and take the maximum over epistemic and aleatoric uncertainty for classification and localization as the score of an unlabeled image.

**MIAOD** [29] Following the code published by the authors, the second detection head and an additional multi-instance learning (MIL) head are built upon Faster R-CNN, and the disagreement between the two classification branches is used as an uncertainty indicator. The weight of MIL is finetuned as 0.1 in consideration of convergence and better performance. The results of MIAOD obtained from our re-implementation significantly surpass those reported by the authors on the COCO dataset. However, under the VOC setting, consistent improvement cannot be assured over the active learning cycles after parameter search, among which the best performance we can get is no more than 72% mAP50. Experiments with the code provided by the authors under the same budget also confirm our observation, where we only get around 70% mAP50 in the last cycle. This might suggest that MIAOD is less applicable to the low-data regime. Thus, the results of MIAOD on the VOC dataset are not reported in the main paper.

**MCDropout** [7]. Following the practice in [4], the adaption of MCDropout to the detection task is achieved by image-level estimation followed by averaging the results of 25 forward passes. Dropout layers with $p = 0.1$ are inserted at the last two stages after each bottleneck module of the ResNet backbone. The image-level informativeness can be estimated by different acquisition functions. Here we also follow the implementation in [4] to use entropy, which performs the best under our experimental settings.

**Ensemble** [2]. Similarly, we establish an ensemble of three independent detectors following [4]. The informativeness estimation and result ensemble are the same as the MC-Dropout implementation.

**BoxCnt** BoxCnt is devised by us to attack the image-level estimation for active detection. This hack is achieved by naively prioritizing unlabeled images with the most number of box predictions after NMS.

## 8. Related Work Beyond Active Learning

The main paper discusses widely used and most recent acquisition functions, ensemble models and evaluation methods for active learning. In addition to it, we give a

| Methods | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| | COCO-sup | | | | | | | |
| Random | 22.50±0.10 | 25.97±0.15 | 27.37±0.21 | 28.23±0.15 | 29.17±0.21 | 29.87±0.06 | 30.53±0.15 | 31.10±0.10 |
| MeanEntropy | 22.50±0.10 | 26.80±0.20 | 27.93±0.23 | 28.80±0.26 | 29.70±0.30 | 30.37±0.15 | 30.93±0.15 | 31.37±0.06 |
| WhiteBoxQBC | 22.50±0.10 | 26.00±0.20 | 27.43±0.23 | 28.30±0.20 | 29.50±0.46 | 30.13±0.49 | 30.87±0.38 | 31.43±0.32 |
| CoreSet | 22.50±0.10 | 26.03±0.35 | 27.47±0.40 | 28.33±0.23 | 29.33±0.25 | 30.03±0.23 | 30.63±0.21 | 31.27±0.12 |
| LearningLoss* | **23.20±0.20** | 24.07±0.21 | 24.87±0.12 | 25.60±0.17 | 26.33±0.21 | 27.27±0.21 | 27.80±0.20 | 28.47±0.40 |
| MIAOD* | 19.10±0.36 | 22.37±1.18 | 23.80±0.98 | 24.83±0.83 | 25.43±0.81 | 26.43±0.38 | 27.00±0.56 | 27.47±0.60 |
| ALMDN* | 22.93±0.21 | 24.63±0.45 | 25.23±0.38 | 25.73±0.25 | 26.20±0.44 | 26.80±0.36 | 27.60±0.30 | 28.13±0.35 |
| MeanEntropy-MCDropout* | 23.13±0.06 | 27.20±0.20 | 28.27±0.06 | 29.23±0.12 | 29.97±0.21 | 30.80±0.10 | 31.50±0.20 | 31.90±0.10 |
| MeanEntropy-Ensemble* | 22.33±0.06 | 26.33±0.23 | 27.23±0.15 | 28.07±0.46 | 28.80±0.26 | 29.57±0.12 | 29.97±0.06 | 30.57±0.06 |
| sCOMPAS (ours) | 22.50±0.10 | **28.23±0.21** | **30.50±0.20** | **32.03±0.25** | **33.20±0.10** | **33.93±0.15** | **34.60±0.36** | **34.93±0.23** |

Table 6. MAP and standard deviation (%) under COCO-sup and COCO-semi settings, with 50K boxes for initialization and 20K boxes for each active learning cycle. Results reported are averaged over 3 independent runs. The best result of each cycle is highlighted in bold.

more extensive survey of literature related to the proposed framework and method in this section.

**Consistency regularization**. Active learning estimates predictive consistency as an uncertainty indicator to sample unlabeled candidates, based on which images with the most inconsistent hypotheses are queried for human annotation. To this end, previous methods measure the *disagreement* between multiple models or heads [2, 4, 19, 22, 29], and the *robustness* of the output after noise perturbation [10] or horizontal flip [5] of the image.

While the consistency estimation of active learning mainly happens in the acquisition stage, quite similarly, semi-supervised learning (SSL) [9, 15, 23, 27] encourages consistency in the outputs of realistic image perturbations during training. The shared motivation behind them is to find a smooth manifold for the dataset so that the version space of the model is minimized [1, 15, 21]. Their goals are further aligned in [8] for active classification, where under the semi-supervised learning framework, Gao *et al*. sample images with inconsistent predictions that the model has difficulty self-learning via consistency regularization. Most recent active detection methods also draw on advanced semi-supervised learning (SSL) techniques. For example, Mi *et al*. [12] achieve active sampling under the framework of the unbiased teacher [11], a state-of-the-art semi-supervised detection method to exploit all available data. But their informativeness estimation is based on entropy and class diversity, which is not aligned with the training objective. Elezi *et al*. [5] measures the Kullback-Leibler divergence between the predictive distributions of flipped images. They also provide offline pseudo-labels for unlabeled images with confident predictions, so that human-labeled, pseudo-labeled and unlabeled images are all involved in consistency-based model optimization.

In comparison to previous studies, the proposed ComPAS method provides sufficient perturbations as the testing ground for disagreement estimation, aligns the consistency-oriented goal in model training and active acquisition, and supports both labeled-only and mixed-supervision learning.

**Sparse annotation for object detection**. Sparsely an-

notated object detection (SAOD) [14, 17, 25, 26, 30], deals with the interference of unlabeled labels, which will be considered as hard negatives during the training of detectors. The methods in this field mainly fall into two classes: loss re-weighting and pseudo-labeling. Soft sampling [26] and Background Re-calibration [30] trust the judgements of the detector and accordingly adjust the weights of negative proposals. More recent methods draw on a Siamese network [25] or consistency regularization [17] to generate pseudo-labels for unlabeled regions.

The proposed box-level active detection framework inevitably incurs the similar challenge. While SAOD methods are validated on randomly down-sampled benchmark datasets, our setting distinguishes itself in prioritizing the annotation of challenging targets while requiring remedies for the easier ones. Without specific handling, the performance of detection would be severely interfered. But meantime we can benefit from the prior knowledge of the unlabeled targets: pseudo-labeling that accepts confident model predictions can exactly supplement easier targets with self-supervision.

**Mixed types of supervision for object detection**. Besides the box-level supervision we focus on in this paper, some recent work also includes additional image-level labels [3, 6, 31] and more types of supervision (*e.g.* scribbles in [18]) to facilitate the detection performance. For example, [3, 6, 31] utilize weakly labeled datasets and a proportion of fully labeled images. In comparison, our box-only setting is simpler and more effective. Take the VOC dataset for example, as presented in Tab. 4, sCOMPAS obtains 73.57% mAP50 with merely 8.5% (4K) *boxes*, and then achieves 76.4% mAP50 with 10.6% (5K) *boxes*. Given access to all *unlabeled images*, mCOMPAS reaches 79.97%-80.63% mAP50. It clearly surpasses the state-of-the-art 69.4% mAP50 in [3, 6, 31], where *all image-level labels* and 10% *fully labeled images* are required.

Mixed types of supervision can also be sampled via active learning. Based on a pre-trained weakly supervised detector, BiB [24] proposes to actively provide full box annotations for images. BAOD [16] progressively adds

image-level supervision or full box annotations in each active learning cycle. In contrast to their mixed types of supervision and exhaustive annotation protocol, the proposed pipeline is simple and effective in de-redundancy, and is superior in detection performance.

## 9. Numerical Results

Last, we present the exact numerical results used to plot Fig. 1 and Fig. 3 of the main paper and the results of Fig. 5 in this supplementary file. Tab. 4 reports mAP50 and standard deviation (%) under the VOC-sup and VOC-semi settings. Tab. 5 presents results under COCO-sup and COCO-semi settings with the 20K box-level annotation batch size, and Tab. 6 presents the results with a larger batch size of 40K boxes.

## References

[1] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(11), 2006.

[2] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan 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] Carlo Biffi, Steven McDonagh, Philip Torr, Aleš Leonardis, and Sarah Parisot. Many-shot from low-shot: Learning to annotate using mixed supervision for object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII*, pages 35–50. Springer, 2020.

[4] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, and Jose M. Alvarez. Active learning for deep object detection via probabilistic modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10264–10273, October 2021.

[5] Ismail Elezi, Zhiding Yu, Anima Anandkumar, Laura Leal-Taixe, and Jose M Alvarez. Not All Labels Are Equal: Rationalizing the Labeling Costs for Training Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14492–14501, 2022.

[6] Linpu Fang, Hang Xu, Zhili Liu, Sarah Parisot, and Zhenguo Li. Ehsod: Cam-guided end-to-end hybrid-supervised object detection with cascade refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10778–10785, 2020.

[7] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017.

[8] Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision*, pages 510–526. Springer, 2020.

[9] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Advances in neural information processing systems*, 32:10759–10768, 2019.

[10] Chieh-Chi Kao, Teng-Yok Lee, Pradeep Sen, and Ming-Yu Liu. Localization-aware active learning for object detection. In *Asian Conference on Computer Vision*, pages 506–522. Springer, 2018.

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

[12] Peng Mi, Jianghang Lin, Yiyi Zhou, Yunhang Shen, Gen Luo, Xiaoshuai Sun, Liujuan Cao, Rongrong Fu, Qiang Xu, and Rongrong Ji. Active Teacher for Semi-Supervised Object Detection. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 14482–14491, 2022.

[13] Prateek Munjal, Nasir Hayat, Munawar Hayat, Jamshid Sourati, and Shadab Khan. Towards Robust and Reproducible Active Learning Using Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 223–232.

[14] Yusuke Niitani, Takuya Akiba, Tommi Kerola, Toru Ogawa, Shotaro Sano, and Shuji Suzuki. Sampling techniques for large-scale object detection from sparsely annotated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6510–6518, 2019.

[15] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31, 2018.

[16] Alejandro Pardo, Mengmeng Xu, Ali Thabet, Pablo Arbelaez, and Bernard Ghanem. Baod: budget-aware object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1247–1256, 2021.

[17] Sai Saketh Rambhatla, Saksham Suri, Rama Chellappa, and Abhinav Shrivastava. Sparsely annotated object detection: A region-based semi-supervised approach. *arXiv preprint arXiv:2201.04620*, 2022.

[18] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Alexander G Schwing, and Jan Kautz. Ufo2: A unified framework towards omni-supervised object detection. In *European Conference on Computer Vision*, pages 288–313. Springer, 2020.

[19] Soumya Roy, Asim Unmesh, and Vinay P Namboodiri. Deep active learning for object detection. In *BMVC*, page 91, 2018.

[20] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

[21] Burr Settles. Active learning: Synthesis lectures on artificial intelligence and machine learning. *Long Island, NY: Morgan & Clay Pool*, 2012.

[22] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992.

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

[24] Huy V. Vo, Oriane Siméoni, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, and Jean Ponce. Active learning strategies for weakly-supervised object detection. October 2022.

[25] Tiancai Wang, Tong Yang, Jiale Cao, and Xiangyu Zhang. Co-mining: Self-supervised learning for sparsely annotated object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2800–2808, 2021.

[26] Zhe Wu, Navaneeth Bodla, Bharat Singh, Mahyar Najibi, Rama Chellappa, and Larry S Davis. Soft sampling for robust object detection. In *British Machine Vision Conference (BMVC)*, 2018.

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

[28] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 93–102, 2019.

[29] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5330–5339, 2021.

[30] Han Zhang, Fangyi Chen, Zhiqiang Shen, Qiqi Hao, Chenchen Zhu, and Marios Savvides. Solving missing-annotation object detection with background recalibration loss. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1888–1892. IEEE, 2020.

[31] Yi Zhong, Chengyao Wang, Shiyong Li, Zhu Zhou, Yaowei Wang, and Wei-Shi Zheng. Mixed supervision for instance learning in object detection with few-shot annotation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 648–658, 2022.