This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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 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.

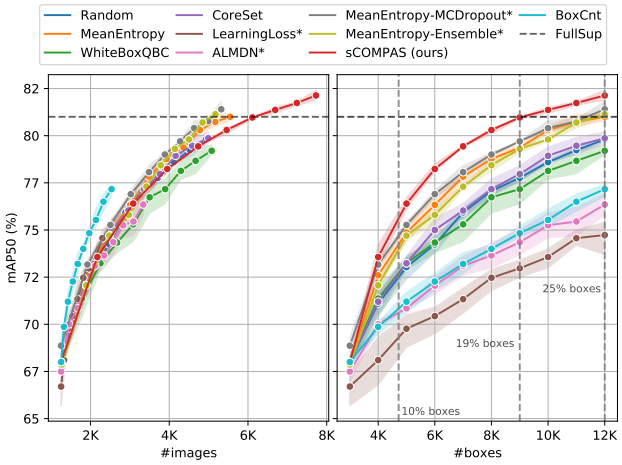

Figure 1. Active detection methods evaluated on VOC0712 under image-level (Left) and box-level (Right) settings. BoxCnt is our hack that simply queries potentially the most crowded images, which demonstrates that image-level evaluation is highly biased. Methods marked with * have specialized detector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

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

*Corresponding Authors.

is 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

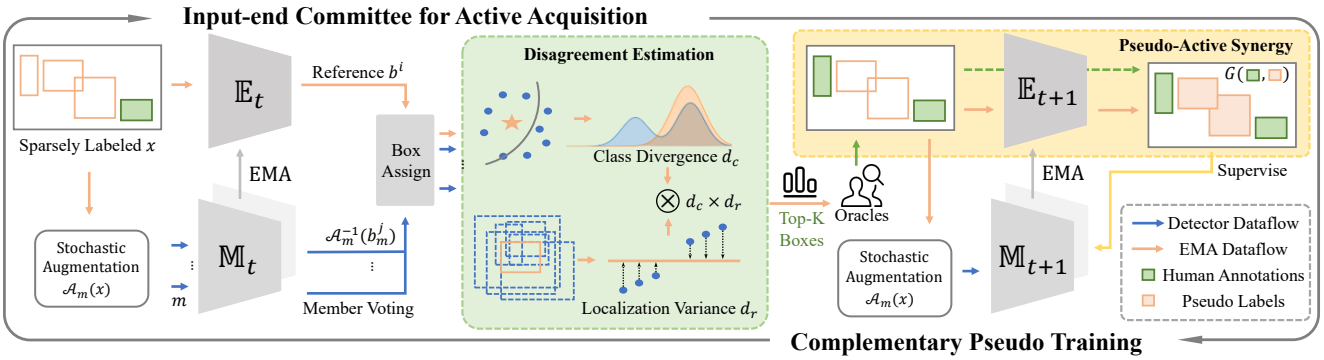

Figure 2. Overview of our ComPAS pipeline for box-level active detection, which iterates between active acquisition via the input-end committee and complementary training based on pseudo-active synergy. Only workflow of sparsely labeled images is shown for generality.

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 iteration 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:

$$d^i = d^i_c \times d^i_r, \tag{5}$$

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 $\{x^i_l\}$ from L_t are fed into the detector and learned in a supervised way:

$$\mathcal{L}_l = \frac{1}{N_l}\sum_i \mathcal{L}_{cls}(x^i_l, y^i_l) + \mathcal{L}_{loc}(x^i_l, t^i_l), \tag{6}$$

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 λ_c to obtain reliable boxes $\{\hat{b}^i_c\}$ 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 λ_r are kept to supervise the regression head, which is denoted as $\{\hat{b}^i_r\}$.

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:

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

where we supplement sparse ground truth labels y_s with pseudo-labels $\hat{y}_{sc}$ whose corresponding boxes $\hat{b}_{sc}$ have less than λ_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:

$$\begin{aligned}\mathcal{L}_s = \frac{1}{N_s}\sum_i^{N_s} &\mathcal{L}_{cls}(\mathcal{A}(x^i_s), G(y^i_s, \hat{y}^i_{sc})) + \\ &\mathcal{L}_{loc}(\mathcal{A}(x^i_s), G(t^i_s, \hat{t}^i_{sr})),\end{aligned} \tag{8}$$

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:

$$\mathcal{L}_u = \frac{1}{N_u}\sum_i^{N_u}(\mathcal{L}_{cls}(\mathcal{A}(x^i_u), \hat{y}^i_{uc}) + \mathcal{L}_{loc}(\mathcal{A}(x^i_u), \hat{b}^i_{ur})), \tag{9}$$

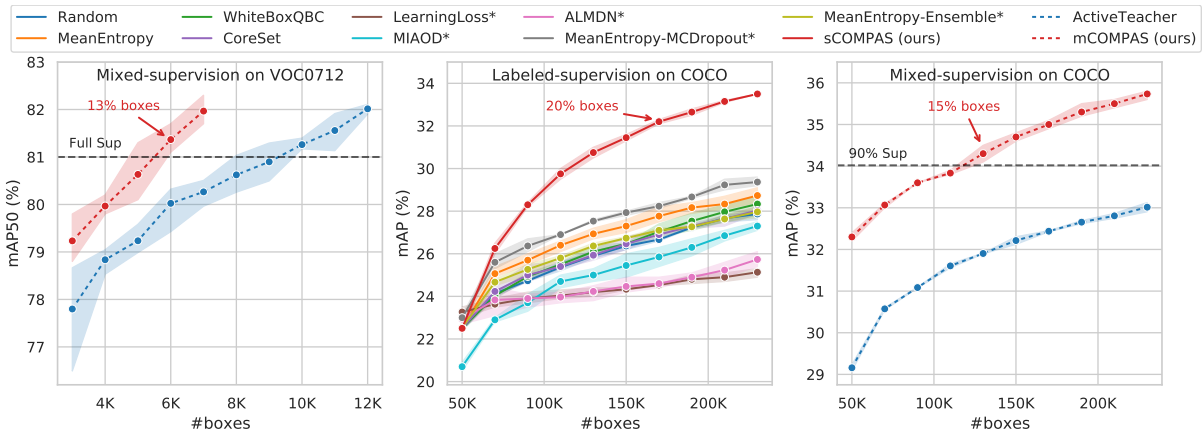

Figure 3. Box-level comparative results on (Left) VOC-semi, (Middle) COCO-sup and (Right) COCO-semi. Solid lines are performed with labeled-supervision, whereas dashed lines indicate training with unlabeled images.

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 subtasks 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.

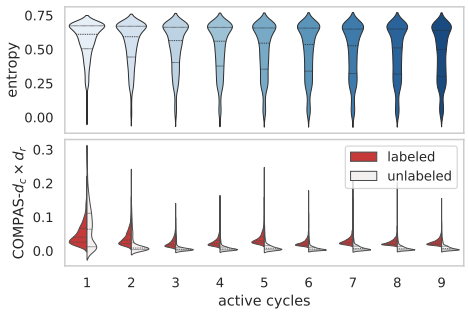

Figure 4. Violin plots showing the box-level score distributions of newly acquired images in each active learning cycle calculated by (Top) image-level MeanEntropy and (Bottom) box-level ComPAS.

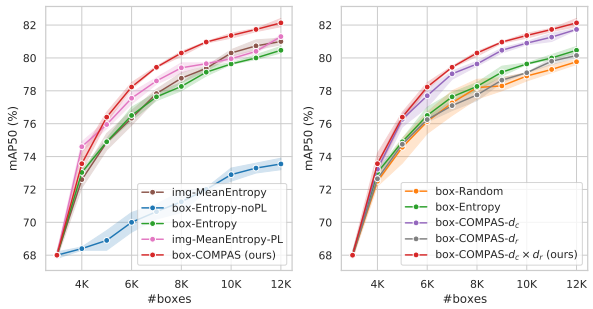

Figure 5. Analysis on (Left) the extension from image-level annotation to the box-level and (Right) alternatives to the box-level active acquisition strategy under the VOC-sup setting.

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-

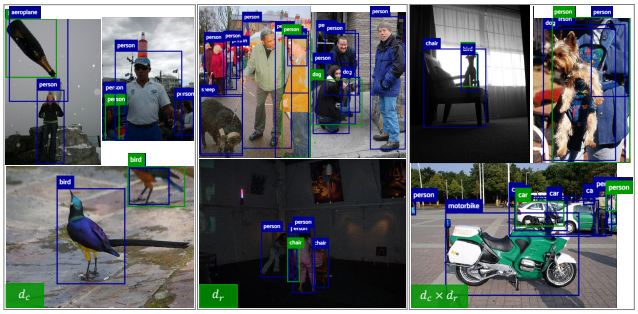

Figure 6. Qualitative results of targets top-ranked by our scoring functions (in green) and complementary pseudo-labels (in blue).

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

This work was partly supported by National Key R&D Program of China (No. 2022ZD0119400), National Natural Science Foundation of China (Nos. 61925107, 62271281, U1936202), Zhejiang Provincial Natural Science Foundation of China under Grant (No. LDT23F01013F01), China Postdoctoral Science Foundation (BX2021161) and Tsinghua-OPPO JCFDT.

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