Box-Level Active Detection – Supplementary Material

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

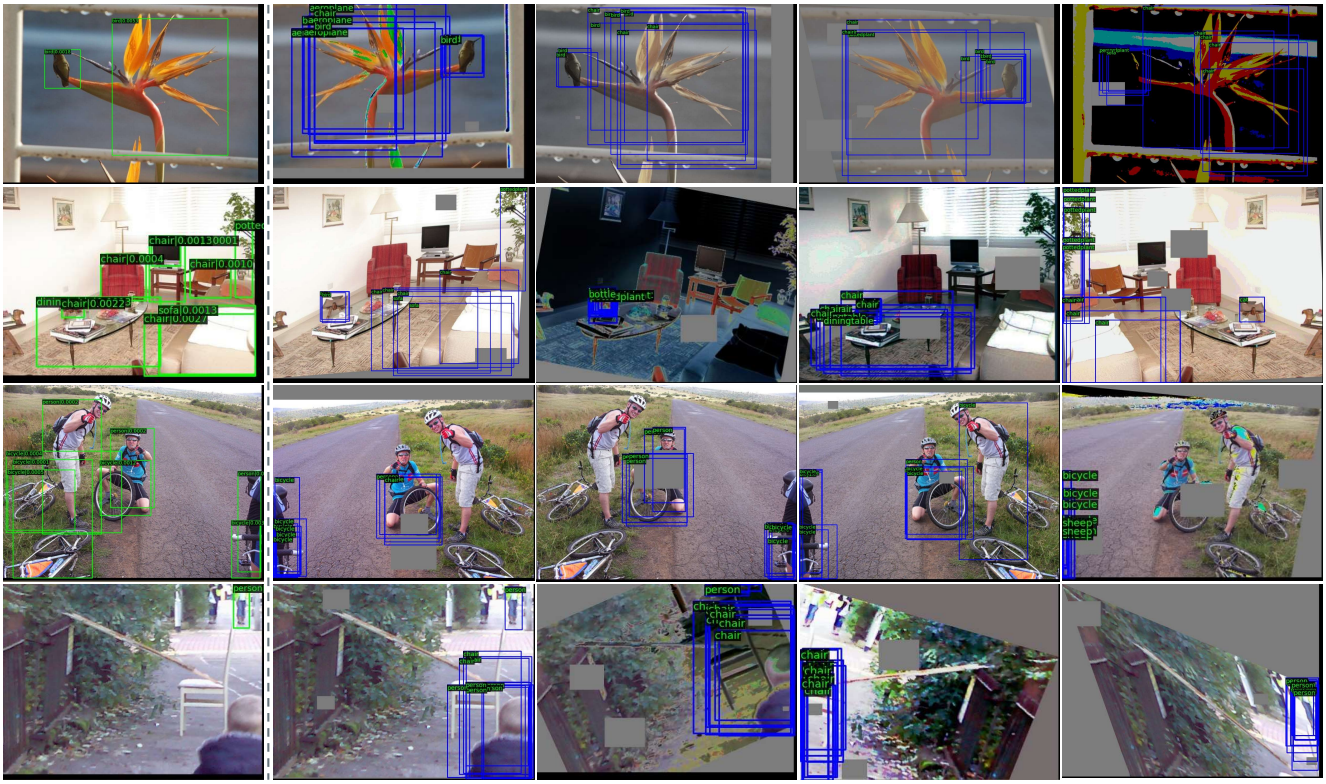

Figure 1. Reference generated by the chairman model (1st column with boxes in green and box-level acquisition scores), and corresponding member predictions (2nd-5th columns with boxes in blue). Experiments are performed at the first active learning cycle under the VOC-sup setting. We only show top-ranking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

Contents

*Corresponding Authors.

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.

A. 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 consistently improves over cycles.

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.

Methods	M	#Trainable Parameters (M)	Inference Time (s)
MeanEntropy-Ensemble	3	123.51	0.0310
MeanEntropy-MCDropout	25	41.17	0.2400
sComPAS (ours)	1	41.17	0.0277
	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.

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 respectively. 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 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.

B. 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.

C. 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 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].

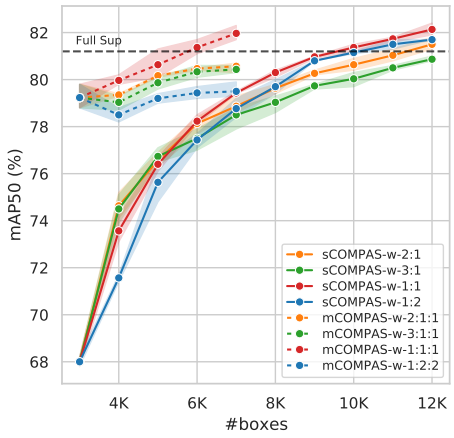

Figure 3. Analysis of the loss weight ratio among different levels of supervision (full-labeled: partial-labeled(: unlabeled)) under the VOC-sup and VOC-semi settings.

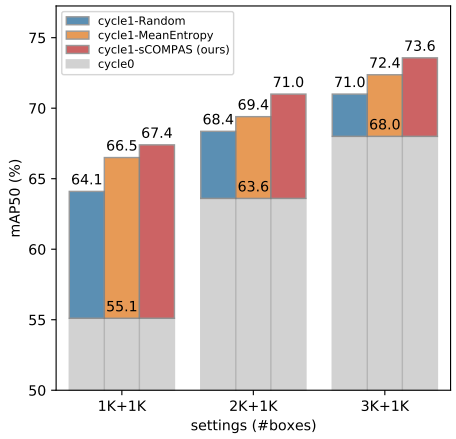

Figure 4. Weaker initialization (cycle0) and the started state (cycle1) under lower-data regimes under the VOC-sup setting.

D. 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.

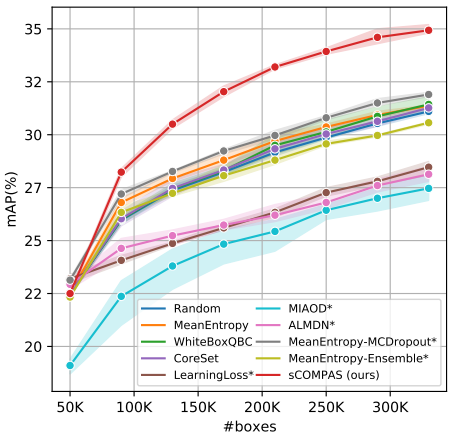

Figure 5. Box-level comparative results on COCO-sup with a 40K per-cycle budget.

E. 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.

F. 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 lo-

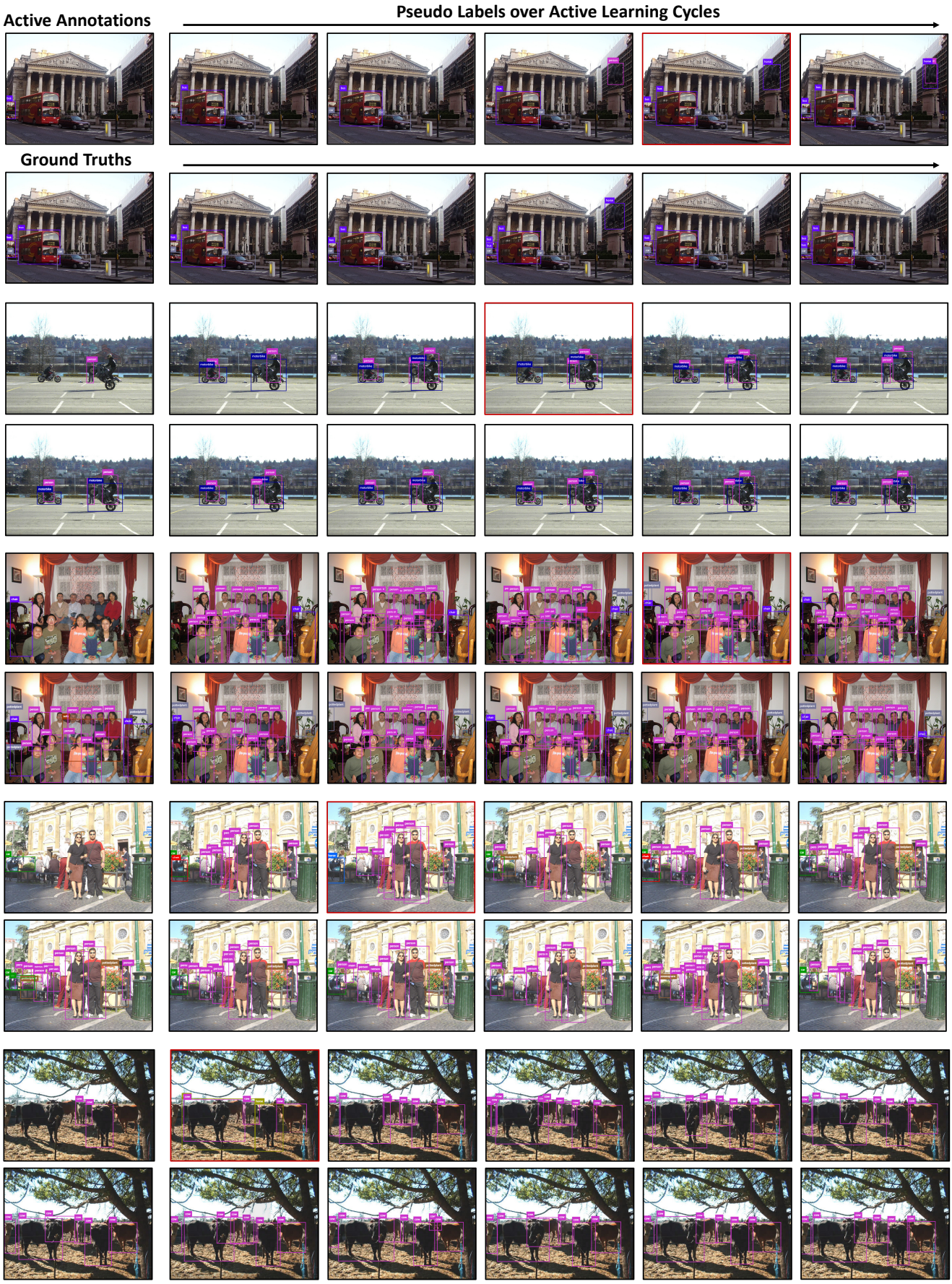

Figure 6. Iteration of pseudo labels across active learning cycles accompanied by actively annotated targets. For each set of images, we highlight the image in red frame if the active annotation happens in that step.

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.

calization across active learning cycles. Through iterative knowledge gain and self-supervision, the synergy between them is effectively exploited.

G. 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 authors, 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.

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.

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.

H. 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 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 annotated 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*. scrib-

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.

bles 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.

I. 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. 7

[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, 6, 7

[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. 7, 8

[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. 3, 6, 7

[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. 7

[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, 8

[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. 3, 6

[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. 7

[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. 7

[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. 7

[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. 7

[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. 7

[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, 2022. 4

[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. 7

[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. 7

[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. 8

[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. 7

[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. 8

[19] Soumya Roy, Asim Unmesh, and Vinay P Namboodiri. Deep active learning for object detection. In *BMVC*, page 91, 2018. 2, 6, 7

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

[21] Burr Settles. Active learning: Synthesis lectures on artificial intelligence and machine learning. *Long Island, NY: Morgan & Clay Pool*, 2012. 3, 7

[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. 7

[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. 7

[24] Huy V. Vo, Oriane Siméoni, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, and Jean Ponce. Active learning strategies for weakly-supervised object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, October 2022. 8

[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. 7

[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. 7

[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. 7

[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. 6

[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. 6, 7

[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. 7

[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. 7, 8